Pampouchidou *et al. EURASIP Journal on Image and Video Processing* (2017) 2017:64
DOI 10.1186/s13640-017-0212-3

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Quantitative comparison of motion history image variants for video-based depression assessment

Anastasia Pampouchidou[1*], Matthew Pediaditis[2], Anna Maridaki[3], Muhammad Awais[4], Calliope-Marina Vazakopoulou[3], Stelios Sfakianakis[2], Manolis Tsiknakis[2,3], Panagiotis Simos[5], Kostas Marias[2,3], Fan Yang[1] and Fabrice Meriaudeau[1,4]

Abstract

Depression is the most prevalent mood disorder and a leading cause of disability worldwide. Automated video-based analyses may afford objective measures to support clinical judgments. In the present paper, categorical depression assessment is addressed by proposing a novel variant of the Motion History Image (MHI) which considers Gabor-inhibited filtered data instead of the original image. Classification results obtained with this method on the AVEC'14 dataset are compared to those derived using (a) an earlier MHI variant, the Landmark Motion History Image (LMHI), and (b) the original MHI. The different motion representations were tested in several combinations of appearance-based descriptors, as well as with the use of convolutional neural networks. The F1 score of 87.4% achieved in the proposed work outperformed previously reported approaches.

Keywords: Depression assessment, Affective computing, Machine learning, Image processing, Facial image analysis, Motion history image, Gabor inhibition, Facial landmarks

1 Introduction

Major depressive disorder (MDD) is the most prevalent mood disorder, currently reported as the prime cause of disability worldwide [1]. With rising frequency, MDD is also a major factor associated with suicidal behavior. The gold standard for MDD diagnosis is a clinical interview conducted by a specially trained and experienced mental health professional evaluating the presence of widely accepted criteria, such as those specified in the Diagnostic and Statistical Manual of Mental Disorders [2, 3]. In clinical practice, diagnosis can be supported by self-report scales, such as the Beck Depression Inventory (BDI) [4]. Self-ratings of depressive symptomatology are affected by several biasing factors (e.g., subjective and social acceptance). Therefore, additional, objective measures able to support MDD diagnosis could be proven advantageous in both clinical and research settings [5].

A variety of non-verbal manifestations of MDD have been established in the clinical literature [6, 7]. Some of the visual signs of MDD include head movements, facial expression variability, smiles, and frowns [5]. The present work evaluates the accuracy of automatic, video-based detection of such visual signs. More specifically, the proposed work is based on the hypothesis that individuals suffering from depression display less motion and lower motion variability in their facial expressions than non-depressed persons. The methods presented are derived from the Motion History Image (MHI), an algorithm more commonly employed for action recognition [8]. The performance of three MHI variants is contrasted, combined with appearance-based descriptors and deep learning methods. Algorithm performance is tested on the benchmark dataset provided by the 2014 Audio/Visual Emotion Challenge (AVEC'14). The main contribution of the proposed work lies in the improvement of classification performance for the given dataset, as well as introducing a novel variant of MHI, the Gabor Motion History Image (GMHI).

*Correspondence: anastasia.pampouchidou@gmail.com
[1] Le2i Laboratory, University of Burgundy, Le Creusot, France
Full list of author information is available at the end of the article

The current manuscript is organized in six sections. Section 2 presents existing algorithmic implementations and applications on the same dataset. The current methodology is outlined in Section 3, and experimental results are reviewed in Section 4. Discussion and interpretation of results is undertaken in Section 5, with concluding remarks and recommendations for future work included in Section 6.

2 Related work

Related work is summarized in this section in terms of (a) the employed dataset and (b) algorithms involved in the proposed method. In the first subsection, previous work using AVEC datasets, for categorical assessment of depression based on visual cues, is briefly reviewed. The second subsection presents a review of the algorithms and methods which were combined to realize the proposed approach, namely MHI, Gabor filtering, and deep learning.

2.1 Audio/visual emotion challenge

Although decision support systems based on visual cues have not been incorporated into standard clinical practice, several approaches for developing such systems can be found in the literature. The majority of published approaches materialized within the AVEC "Depression Recognition Sub-challenge" (DSC) [9–11].

2.1.1 AVEC: challenge and dataset

AVEC'13 and AVEC'14 addressed the problem of estimating the severity of depressive symptomatology as a continuous variable as indexed by (self-reported) BDI scores. The dataset consisted of video feeds of undiagnosed volunteers performing the following tasks: vowel pronunciation, solving a task out loud, counting from 1 to 10, reading novel excerpts, singing, and describing a specific scene displayed in pictorial form. During AVEC'13, the complete recording was used for testing the performance of the different approaches, while in AVEC'14, two tasks were selected: the Northwind (reading excerpt) and the Freeform (answering to questions) [10]. AVEC'16 [11] utilized a different dataset (DAIC-WOZ [12]) and provided only the features extracted with the OpenFace software [13], instead of the original video recordings.

Despite the AVEC'13 and AVEC'14 datasets being focused on continuous depression assessment, different approaches have utilized it to address classification of portrayed persons into high- and low-depression severity groups according to the following standard BDI score cut-offs [4]:

- 0–13: Minimal depression
- 14–18: Mild depression
- 19–28: Moderate depression
- 30–63: Severe depression

2.1.2 Categorical depression assessment with AVEC datasets

For the purposes of their participation in AVEC'14, Senoussaoui et al. [14] attempted classification of video feeds into "absence" (as indicated by BDI points $\leq$13) versus "presence" (as indicated by BDI points $>$13) of significant depressive symptomatology. They used the Local Gabor Binary Patterns in Three Orthogonal Planes, provided by the challenge organizers [10], achieving 82% accuracy. Similar results were obtained in the cross-cultural study of Alghowinem et al. [15] who focused on geometrical features derived from eye activity to achieve 81.3% classification accuracy utilizing one subset of the AVEC'13 dataset, among other datasets. Pampouchidou et al. [16] reported 74.5% accuracy with the Local Curvelet Binary Patterns in Pairwise Orthogonal Planes, while Pampouchidou et al. [17] employed geometrical features to achieve an F1 score of 58.6% in the single-modality (i.e., visual) approach and 72.8% when taking into account both audio and visual features.

2.2 Motion history image

MHI is a robust, yet relatively straightforward, algorithm developed to represent the motion that occurs in the course of a complete video recording with a single image [18]. The algorithm produces a grayscale image, in which the white pixels correspond to the most recent movements and the darkest gray correspond to the earliest motion elements. Black pixels indicate absence of movement. It is a popular algorithm for motion analysis [18] and has been extensively used in the field of human action recognition [8].

An early approach of MHI to facial image analysis was that of Valstar et al. [19], who employed MHI in facial action recognition from videos. Meng et al. [20] published a continuous depression assessment approach in their participation to the DSC of AVEC'13; they proposed an extension of MHI, the Motion History Histograms (MHH), which considers patterns of movement. In the DSC of AVEC'14, Pérez Espinoza et al. [21] employed MHI, and for the same challenge, Jan et al. [22] proposed the 1-D MHH, an extension of MHI, which is computed on the feature vector sequence instead of the intensity image. As part of their DSC-AVEC'16 participation, Pampouchidou et al. [23] introduced Landmark Motion History Images (LMHI), which instead of considering intensities from image sequences, considers sequences of facial landmarks.

2.3 Gabor filtering and inhibition

Gabor filters have been frequently used in both facial expression analysis and emotion recognition [24, 25]. In relevant approaches, the feature vector is extracted from the convolution of the original image with a 2D

Gabor wavelet function at different orientations and wavelengths. This describes the spatial frequency structure around each pixel. In [26], the Gabor energy was used for facial emotion recognition, which gives a smoother response to an edge or a line of appropriate width with a local maximum exactly at the edge or in the center of the line. The authors also applied background texture suppression on the response of the filter, by removing an image filtered by the difference of Gaussians (DoG) from the original response for each orientation. This approach, also known as anisotropic inhibition [27], removes noise and provides a sharper representation of facial features.

2.4 Deep learning

Deep learning, which has become increasingly popular during recent years, is a self-learning tool designed to identify patterns in several sets of data samples, extracted from multiple processing layers. Each layer is composed of representation learning methods and is processed in a higher and more abstract level [28]. Convolutional Neural Network (CNN) is a particular deep feedforward network with higher generalization efficiency than other fully connected networks. There are typically two types of layers: the convolutional (conv.) layer and the pooling layer. In the conv. layer, all units are arranged in a feature map and connected to the weights (also known as filter banks), while the weighted sum is inserted to the Rectified Linear Unit (ReLU). The CNN architecture used in the present work was employed in the participation that won the 2012 ImageNet competition (ILSVRC) [29, 30].

Deep learning approaches have been widely tested in the field of facial expression recognition [31]. In the field of depression assessment from visual cues, however, we could find only two published reports. Dibeklioğlu et al. [32] employed Stacked Denoising Autoencoders in a multimodal context to perform video classification according to three levels of depressive symptomatology on the Pittsburgh dataset. Moreover, Zhu et al. [33] employed Deep Convolutional Neural Networks to achieve the highest performance among the unimodal (visual) approaches addressing the aim of AVEC'13 and AVEC'14 competitions.

3 Methodology

The analysis pipeline employed in the current approach is presented in Fig. 2. The first step entails preprocessing, followed by motion representation and feature extraction from the motion images. Dimensionality reduction is performed next, to provide the classifier with the appropriate feature descriptors.

3.1 Preprocessing

Meaningful processing of video frames requires, first, extraction of the region of interest (face). Detection of 2D facial landmarks (c.f. Fig. 1) and extraction of aligned facial images, of size 112×112 pixels, were accomplished using OpenFace, an open source application [13]. A binary "success" score is provided for each frame, with "0" and "1" indicating unsuccessful and successful detection, respectively. In the present work, only successfully detected frames were retained for further processing.

3.2 Motion representation

It is well supported in clinical literature that most of the non-verbal signs of depression are dynamic by nature [6, 7]. Therefore, the use of video-based methods (dynamic), as opposed to frame-based (static), is preferable. In the proposed work, three different motion history images were implemented: (a) the Motion History Image (MHI) as derived from the basic algorithm, (b) the Landmark Motion History Image (LMHI) which relies on facial landmarks, and (c) the Gabor Motion History Image (GMHI). More details regarding the specific motion representation algorithms are presented below with implementation examples illustrated in Fig. 4.

3.2.1 Motion History Image

The MHI is a grayscale image, where white pixels correspond to the most recent movement in the video, intermediate grayscale values to corresponding less recent movements, and black pixels to the absence of movement. The MHI algorithm, with slight variations as explained next, is applied on the aligned face image sequences derived from the preprocessed data using OpenFace as described in Section 3.1.

The MHI H, with a resolution equal to the one of the aligned faces, is computed based on an update function $\Psi(x,y)$ as follows:

$$H_i(x,y) = \begin{cases} 0 & i = 1 \\ i \cdot s & \Psi_i(x,y) = 1 \\ H_{(i-1)}(x,y) & \text{otherwise} \end{cases} \tag{1}$$

where $s = 255/N$, N is the total number of video frames, (x,y) is the position of the corresponding pixel, and i is the frame number. $\Psi_i(x,y)$ represents the presence of movement, derived from the comparison of consecutive frames, using a threshold ξ:

$$\Psi_i(x,y) = \begin{cases} 1 & D_i(x,y) \geq \xi \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $D_i(x,y)$ is defined as a difference distance:

$$D_i(x,y) = \left| I_i(x,y) - I_{(i-1)}(x,y) \right| \tag{3}$$

$I_i(x,y)$ is the pixel intensity value in (x,y) at the ith frame. The final MHI is the $H_N(x,y)$.

3.2.2 Landmark Motion History Image

The LMHI originally proposed in Pampouchidou et al. [23] considers the landmarks derived from OpenFace. The

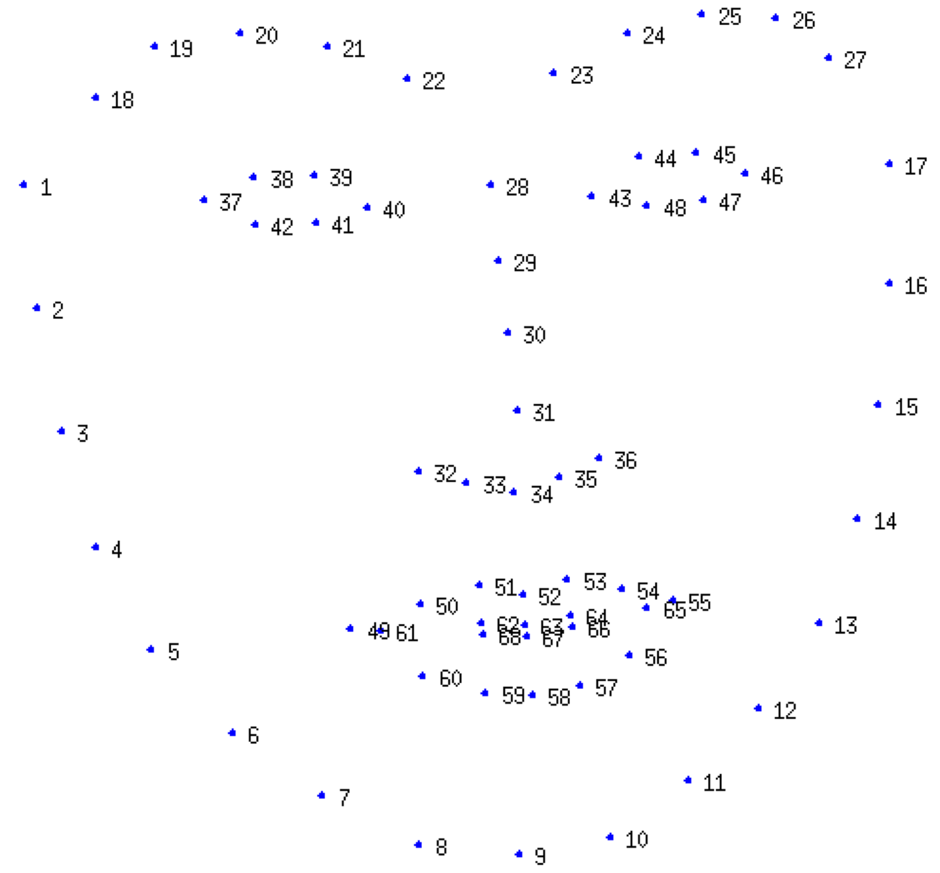

Fig. 1 2D facial landmarks as detected by OpenFace

landmarks considered are the ones which correspond to the facial features (eyes, eyebrows, nose-tip, and mouth), while the face outline is excluded.

This step was taken in order to emphasize inner facial movements and ignore the overall head movements. This is achieved by co-registering the involved landmarks using affine transformation before computing the LMHI, through alignment of the points corresponding to the temples, chin, and inner and outer corners of the eyes (landmarks $\{1, 9, 17, 37, 40, 43, 46\}$).

LMHI differs from the conventional MHI in that image intensities are not considered, but only the facial landmarks, which are detected in each frame. The adopted LMHI algorithm is similar to MHI, by maintaining the same H_i as in (1) and modifying Ψ_i as follows:

$$\Psi_i(x,y) = \begin{cases} 1 & (x,y) \in L_i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where L_i corresponds to the selected landmarks as detected in the ith frame.

3.2.3 *Gabor Motion History Image*

GMHI is another variant of MHI, where Gabor-inhibited images substitute original image intensities. The motivation for implementing this variant is that it focuses on the important details of the facial features and thus extracts the most relevant information. The motion representation algorithm is identical to the one described in Section 3.2.1, but the input image I is the result of the Gabor inhibition. The process of obtaining the Gabor-inhibited image is explained in detail below.

The Gabor wavelet at position (x, y) is given by:

$$\Psi_{\lambda,\theta,\phi,\sigma,\gamma}(x,y) = exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \phi\right) \tag{5}$$

with

$$\begin{aligned} x' &= x \cos\theta + y \sin\theta \\ y' &= -x \sin\theta + y \cos\theta \end{aligned} \tag{6}$$

where λ stands for the wavelength, θ for the orientation, ϕ for the phase offset, σ for the standard deviation of the Gaussian, and γ for the spatial aspect ratio [34].

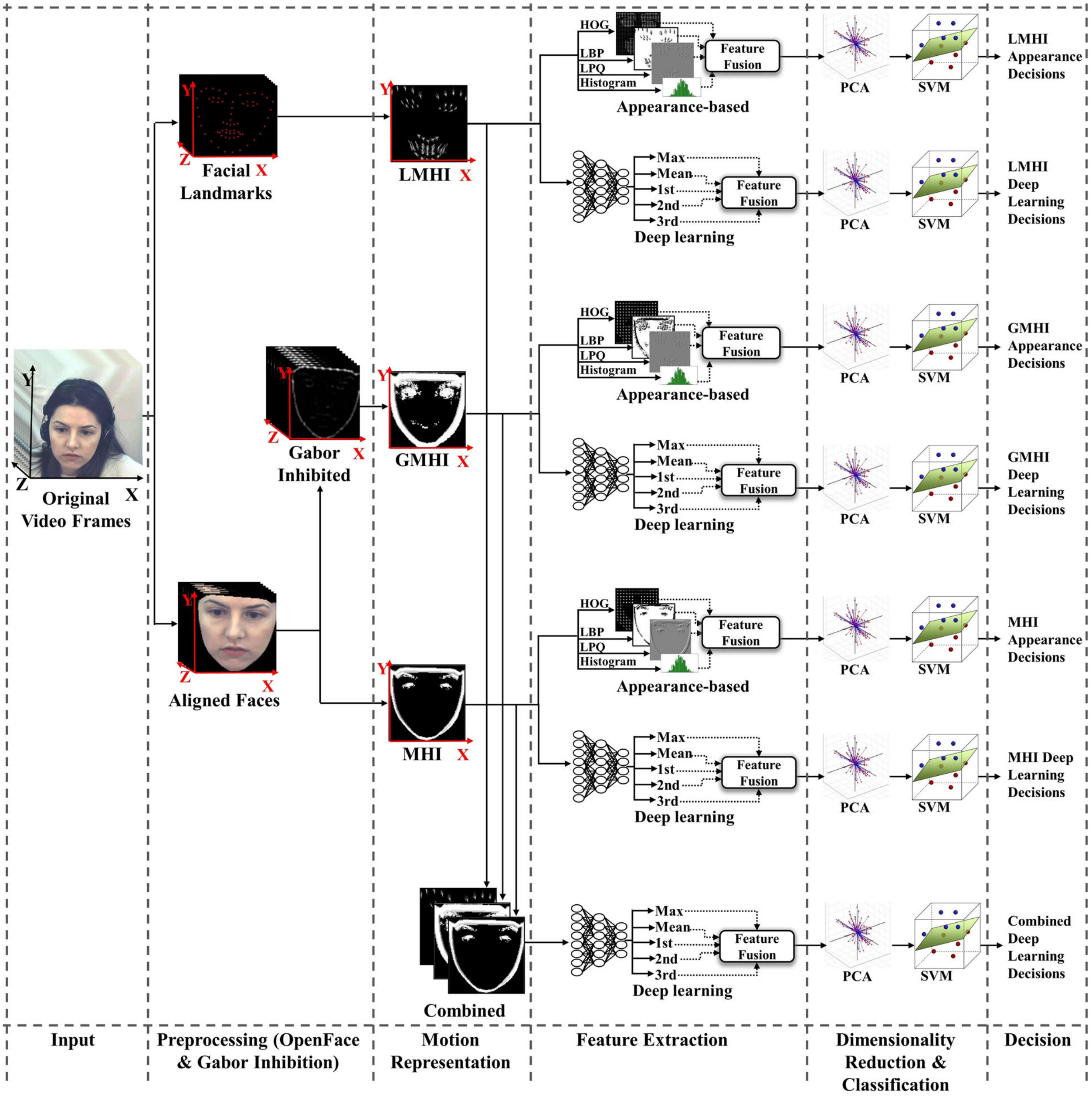
Original Video Frames
Facial Landmarks
Aligned Faces
Gabor Inhibited
LMHI
GMHI
MHI
Combined
HOG
LBP
LPQ
Histogram
Feature Fusion
Appearance-based
Max
Mean
1st
2nd
3rd
Deep learning
PCA
SVM
LMHI Appearance Decisions
LMHI Deep Learning Decisions
GMHI Appearance Decisions
GMHI Deep Learning Decisions
MHI Appearance Decisions
MHI Deep Learning Decisions
Combined Deep Learning Decisions
X
Y
Z
Input
Preprocessing (OpenFace & Gabor Inhibition)
Motion Representation
Feature Extraction
Dimensionality Reduction & Classification
Decision

Fig. 2 Process flow of the proposed algorithm. Dashed arrows in feature extraction indicate that features are considered individually or in combination with feature fusion. MHI: Motion History Image, LMHI: Landmark Motion History Image, GMHI: Gabor Motion History Image, HOG: Histogram of Oriented Gradients, LBP: Local Binary Patterns, LPQ: Local Phase Quantization, PCA: Principal Components Analysis, SVM: Support Vector Machine

The input image is usually filtered with many wavelets for multiple orientations and wavelengths. The energy filter response is obtained by combining the convolutions obtained from two different phase offsets ($\phi_0 = 0$ and $\phi_1 = \pi/2$) using the $L2$-norm. Background texture suppression is applied on the filter response, by removing a DoG-filtered image from the original response for each orientation [27]. Finally, the mean response of Gabor filtering is used to combine the responses across the different orientations, resulting in the pseudo-image used to compute the GMHI. An example of applying the common Gabor and the Gabor-inhibited algorithms to an aligned face image is illustrated in Fig. 3, where the Gabor-inhibited image appears to be sharper and with less texture in uniform regions than the original Gabor response.

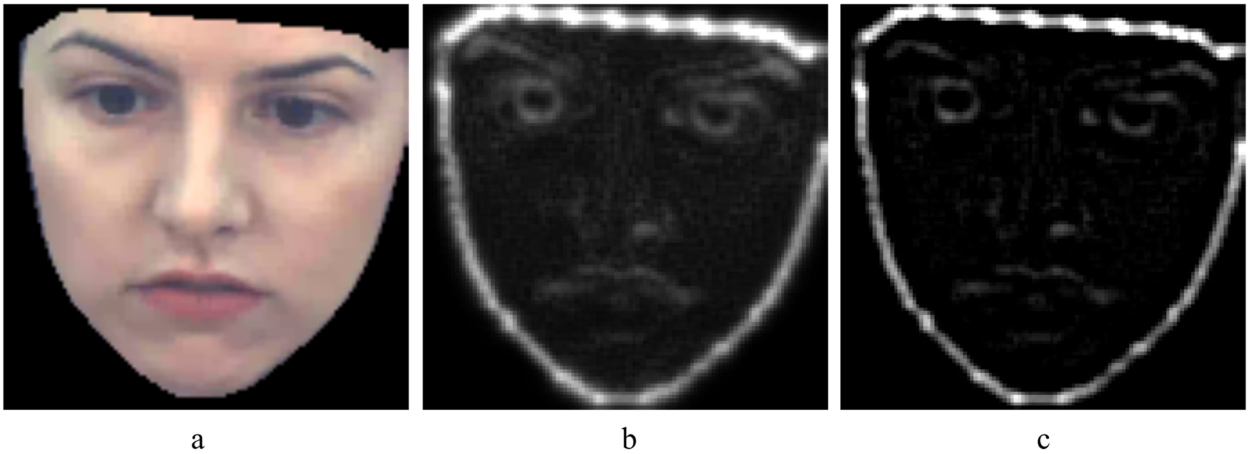

Fig. 3 Example of the Gabor-inhibited image process. **a** Original aligned face image, **b** Gabor reponse, **c** Gabor inhibited

3.3 Feature extraction

Feature extraction was implemented in the present work using two alternative approaches. The first employs appearance-based descriptors, popular in facial image analysis, while the second presents a preliminary attempt to address the problem based on deep learning methods. In both cases, the features were extracted from motion images, instead of the original video recordings.

3.3.1 Appearance-based descriptors

The appearance-based descriptors employed in the present work include the Histogram of Oriented Gradients (HOG), the Local Binary Patterns (LBP), and the Local Phase Quantization (LPQ). Additionally, the combined histogram, mean. and standard deviation of the motion-image gray values are also considered as a single descriptor [Hist-Mean-Std]. Specifically for the histogram, zero values (absence of movement) are disregarded, and only the bins of the remaining 255 gray values are considered, resulting in a 1×257 feature vector in addition to mean and standard deviation. The rest of the descriptors are explained next and illustrated in Fig. 4 for each motion image.

Histogram of oriented gradients (HOG) [35] entails counting gradient orientations in a dense grid. Each image is divided into uniform and non-overlapping cells; the weighted histogram of binned gradient orientations for each cell is computed and subsequently combined to form the final feature vector. HOG results in a 1×6084 feature vector.

Local Binary Patterns (LBP) [36] entails dividing the image into partially overlapping cells. Each pixel of the cell is compared to its neighbors to produce a binary value (pattern). The resulting descriptor is a histogram which represents the occurrence of different patterns. LBP for two sets of {radius, neighborhood} results to feature vectors of size 1×59 for {1,8} and size 1×243 for {2,16}.

Local Phase Quantization (LPQ) [37] is computed in the frequency domain, based on the Fourier transform, for each pixel. Local Fourier coefficients are computed, while their phase information results in binary coefficients after scalar quantization. The final descriptor corresponds to the histogram of the binary coefficients, and it consists of 1×256 features.

3.3.2 Visual Graphic Geometry

Visual Graphic Geometry (VGG) is a CNN variant proposed by Simonyan and Zisserman [38]. Using VGG, they achieved 92.7% top-5 test accuracy on the ImageNet Dataset, which comprises of over 14 million images in 1000 classes. The microarchitecture of VGG16 can be seen in Fig. 5.

The RGB image, with pixel values ranging between 0 and 255, is normalized by subtracting the mean pixel value. The input to VGG (a fixed-size 224×224 RGB image) passes through a stack of conv. layers, where the very small filters are of receptive field size 3×3 to capture the notion of left/right, up/down, and center. The convolution stride is fixed to 1 pixel; the spatial padding of a conv. layer input is such that the spatial resolution is preserved after convolution, i.e., the padding is 1 pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2. A stack of conv. layers (which has a different depth in different architectures) is followed by three fully connected (FC) layers: the first two have 4096 channels each and the third performs 1000-way ILSVRC classification

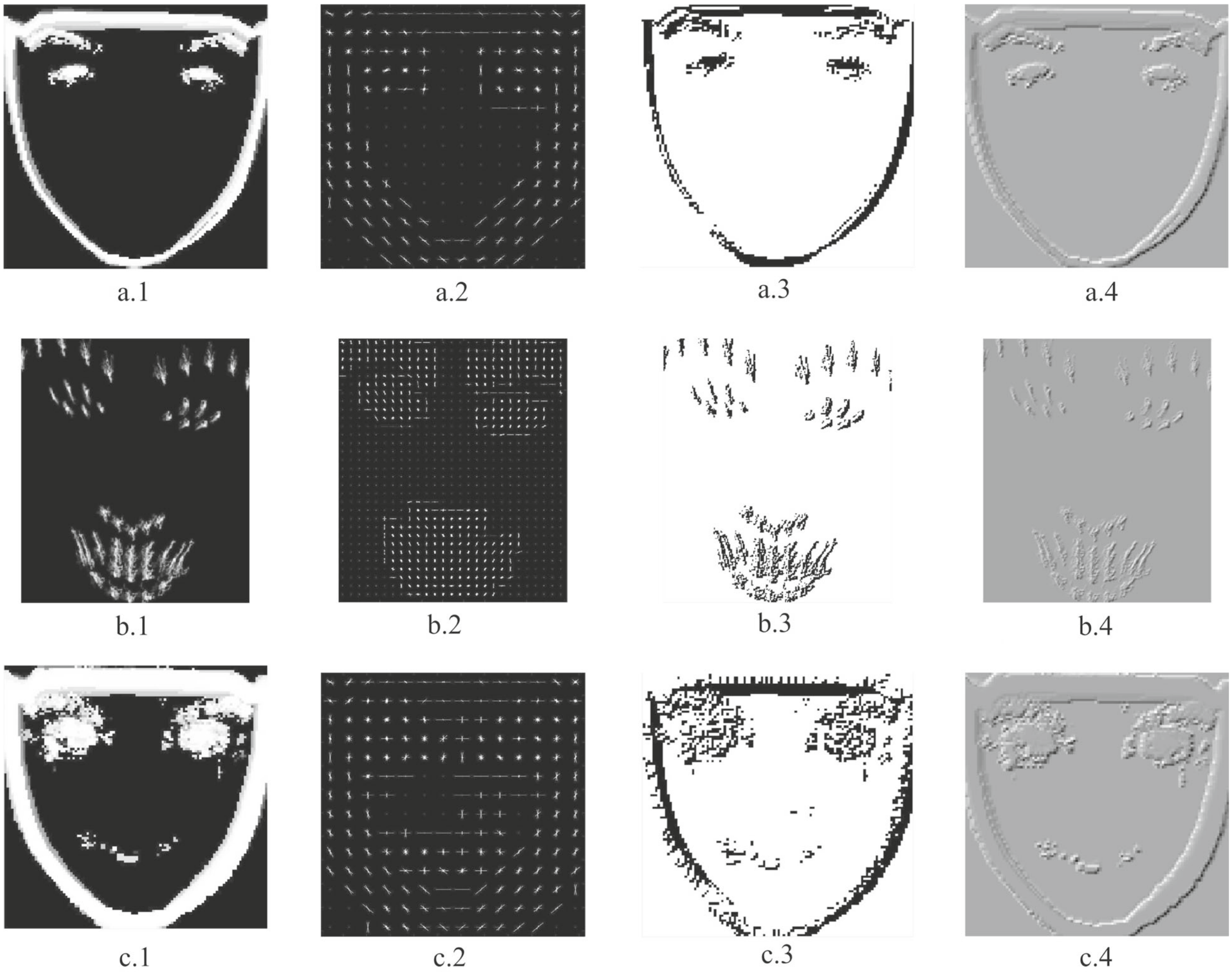

Fig. 4 Different motion representations with corresponding feature responses. **a.1** MHI, **a.2** MHI-HOG, **a.3** MHI-LBP, **a.4** MHI-LPQ, **b.1** LMHI, **b.2** LMHI-HOG, **b.3** LMHI-LBP, **b.4** LMHI-LPQ, **c.1** GMHI, **c.2** GMHI-HOG, **c.3** GMHI-LBP, **c.4** GMHI-LPQ

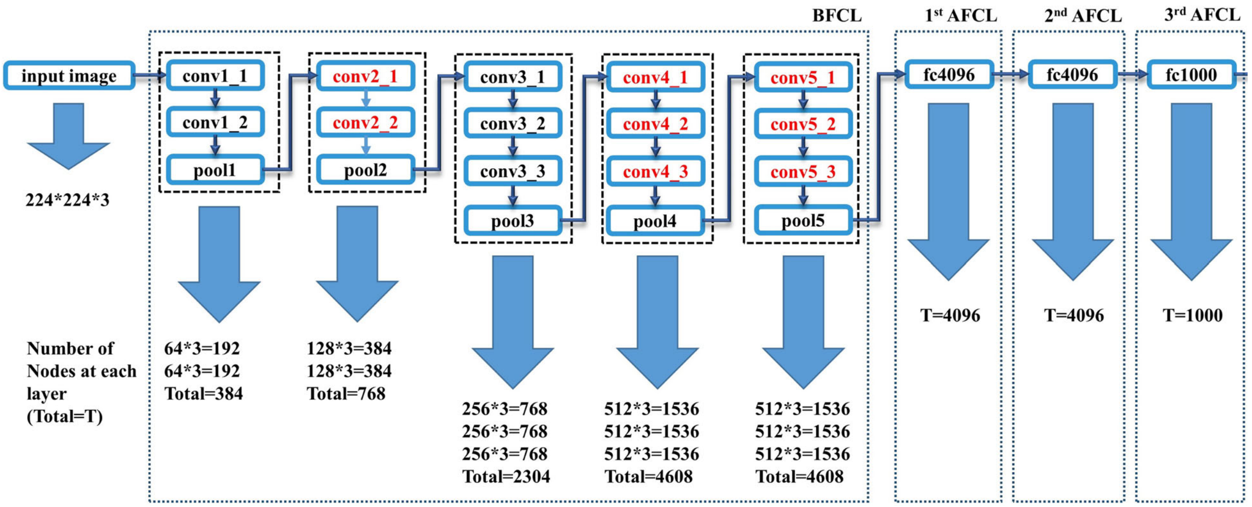

Fig. 5 Microarchitecture of VGG16. BFCL: Before fully connected layer, AFCL: After fully connected layer

and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks. All hidden layers are characterized by non-linearity afforded by ReLU [30].

In the present work, a pre-trained VGG16 network was employed at different VGG layers, for each motion history image separately as well as combined in the form of an RGB image. The proposed method involves transfer learning, by employing the pre-trained VGG. It is applied on the motion history images in order to extract features before the 1st fully connected layer, as well as after the 1st, 2nd, and 3rd fully connected layers. Specifically, the D version of the network was chosen, as it has shown excellent results in related medical applications. The extracted features are subsequently used for classification purposes in the exact same manner as the appearance-based descriptors. The different implementations are explained in what follows.

Before fully connected layer (BFCL) provides the features to the fully connected layer of the VGG16, as shown in Fig. 5. Filter size is 14×14 with 512 kernels. Mean and max values are calculated from each filter, each resulting in a matrix of size 1×512 for each image.

After fully connected layer (AFCL) is the second approach. There are three fully connected layers in VGG16. Layers 1 and 2 operate on a feature matrix of size 1×4096, and layer 3 on a feature matrix of size 1×1000.

3.4 Dimensionality reduction and classification

In the present work, principal component analysis (PCA) was employed to achieve dimensionality reduction. PCA is one of the most popular methods for this purpose and is based on the linear transformation of the original feature vector, into a set of uncorrelated principal components. For a dataset of size $N \times M$ (i.e., N samples and M features), PCA identifies a $M \times M$ coefficient matrix (component loadings) that maps each data vector from the original space to a new space of M principal components. However, by properly selecting a smaller set of $K < M$ components, the dimensionality of the data can be reduced while still retaining much of the information (i.e., variance) in the original dataset.

In the present work, classification was based on a supervised learning model using Support Vector Machines (SVMs). SVM is a non-probabilistic binary linear classifier which, based on the training samples, attempts to identify an optimal hyperplane that maximizes the distance between the two classes.

4 Experimental results

In this section, the configuration of corresponding parameters of the employed algorithms is explained and their performance is evaluated.

4.1 Configuration of parameters

The analysis pipeline proposed here was tested on the benchmark dataset provided by AVEC’14. As mentioned in Section 2, AVEC’14 included data from two tasks: FreeForm and NorthWind and the dataset was partitioned into three subsets: training, development, and test. The BDI scores for each video-clip were provided only for the training and development sets, comprising 200 labeled recordings in total. In order to address the categorical depression assessment, a cutoff of 13/14 points on the BDI-II was set as in [14]. After applying the cutoff, the dataset was fairly balanced consisting of 96 individuals who scored high and 104 persons who scored low on BDI-II, henceforth labeled as “depressed” and “non-depressed” groups.

Regarding specific parameters of the motion representation algorithms, the value of ξ was set to 25 for MHI and to 8 for GMHI. These thresholds were chosen empirically, so that the static background did not present movement in the motion image. This effect was noted at lower ξ values, where differences in pixel intensity were attributed to illumination variations. Setting the appropriate threshold ensures that the movement represented by the motion images is meaningful and can be attributed solely to movements. LBP was tested with two sets of [radius, neighborhood], namely [1,8] and [2,16]. The Gaussian kernel was chosen for the SVM classifier, with the expected proportion of outliers in the training data set to 10%. The $\log(x/(1-x))$ transform function was applied. The number of principal components retained was selected empirically: $k = 100$ for the appearance-based descriptors and $k = 60$ for VGG. A variant of Leave-One-Out (LOO), the Leave-One-Subject-Out (LOSO), was used for cross-validation. LOSO was selected instead of LOO as a non-biased and person-independent method, given that the dataset contained more than one recordings of the same subject (ranging from 2 to 6).

4.2 Performance evaluation

Performance of the different configurations of the proposed algorithm is summarized in Tables 1, 2, 3, 4, and 5. Table 1 presents performance of the various appearance-based descriptors for each of the three different motion images. The descriptors were tested individually and combined with feature level fusion (concatenated). Table 2 presents the performance of VGG for the different configurations as explained in Section 3.3.2. The confusion matrix corresponding to the best-performing model (VGG feature fusion) is presented in Table 3.

Table 1 Experimental results employing appearance-based descriptors (F1-score %)

	LBP{1,8}[a]	LBP{2,16}[a]	HOG	LPQ	Hist + Mean + Std	Feature fusion
MHI	_[b]	_[b]	**81.9**	59.3	**81.9**	36.6
LMHI	_[b]	66.4	64.9	45.8	64.8	**72.7**
GMHI	_[b]	_[b]	**80.0**	69.8	**80**	74.0

The bold values corresponded to the highest performing approach

[a] LBP parameters in brackets correspond to {radius, neighborhood}, respectively

[b] Dash represents unavailable F1-score due to zero depressed individuals classified correctly

Additional performance metrics for the best performing model are reported in Table 4, whereas Table 5 compares the present findings to previously published results using similar datasets. It should be noted that the results presented by Senoussaoui et al. [14] and Alghowinem et al. [15] were obtained using different organizations of the AVEC datasets; thus, a direct comparison with the present results is not possible. In the cross-corpus approach of Alghowinem et al. [15] the used set was carefully selected from the original dataset (AVEC'13) in terms of the total number and duration of recordings per participant, in order to match the other two datasets. On the other hand, in [14], the algorithm was applied to the training dataset provided by the challenge organizers (AVEC'14) and tested on the development dataset. Although the AVEC dataset has been widely employed in approaches for continuous depression assessment, the aforementioned approaches, to the best of the authors' knowledge, are the only ones attempting categorical depression assessment on the specific dataset.

Results are reported for F1-score, unless indicated otherwise, which is given by:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

where precision is given by:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

and recall by:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

Table 2 Experimental results employing VGG features, before and after fully connected layer (F1-score %)

	Before		After			Feature fusion
	Max	Mean	1st	2nd	3rd	
MHI	87.1	63.0	51.5	84.8	70.1	**87.4**
LMHI	64.0	**67.7**	56.6	18.2	66.4	64.0
GMHI	**85.7**	62.7	55.6	76.0	65.5	84.3
Combined	64.6	52.1	51.8	**74.7**	46.3	65.1

The bold values corresponded to the highest performing approach

Table 3 Confusion matrix for the best F1-Score of the proposed approach (VGG-MHI feature fusion)

Self-reported predicted	Non-depressed	Depressed
Non-depressed	**102**	2
Depressed	20	**76**

The bold values correspond to the correctly classified samples

for the confusion matric C:

$$C = \begin{bmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{bmatrix} \tag{10}$$

5 Discussion

In summary, the MHI approach outperformed the other motion images with 87.4% F1-score with the VGG feature level fusion, as it can be observed throughout Tables 1 and 2. Comparable classification results were obtained with the MHI (87.1%) and GMHI (85.7%), both for VGG before fully connected max (c.f. Table 2). The deep learning approach outperformed the appearance-based one as indicated by an improvement of the F1 score in the order of 5.5%. Given that GMHI performed only slightly lower than MHI, it remains promising for further exploration. Combination with different features, testing on a richer dataset, or other relevant applications, such as facial expression recognition, may provide more comprehensive insights on the true value of the GMHI algorithm.

The best performance achieved by appearance-based descriptors, as shown in Table 1, corresponded to an F1 score of 81.9% for HOG as well as for the combination of [Hist+Mean+Std] which, again, is comparable to previous reports. Although LMHI did not perform as well as other methods (72.7%), there may be room for improvement by supplementing it with additional feature descriptors and using different classifiers. It should be noted that head movements were not formally considered in the present work, although changes in the outline of the face during such movements may have been captured by both MHI and GMHI. Moreover, LMHI requires significantly less information (image frames vs. selected landmarks) and ensures participant anonymity by retaining only facial landmarks—an important attribute in studies with clinical samples.

Table 4 Additional performance metrics for the best-performing approach (VGG-MHI feature fusion/%)

F1	87.4
Accuracy	89.0
Sensitivity	79.1
Specificity	98.1
Precision	97.4
Cohen's Kappa	77.8
.95 Confidence interval	±8.6

Table 5 Comparison with previous published results (%)

Approach	Accuracy	Average recall	F1
Sennoussaoui et al. [14]	82.0	–	–
Alghowinem et al. [15]	–	81.3	–
Pampouchidou et al. [16]	74.5	–	–
Pampouchidou et al. [17]	–	–	58.6
Proposed (VGG feature fusion)	**89.0**	**88.6**	**87.4**

The bold values corresponded to the highest performing approach

Further, LBP did not perform well, for the given set of parameters, presenting null recognition. A potential improvement could include testing it for bigger radius and neighborhood, as the ones selected here may represent only micro-movement patterns. HOG and [Hist-Mean-Std] performed equally well and best, among the appearance-based descriptors, with the latter being preferable between the two due to its lower dimensionality (6084 versus 257). It should also be taken into consideration that the aligned images were of lower resolution than the original frame due to OpenFace preprocessing. Thus, representation in higher dimensions may improve performance.

The improvement of VGG-Max from 87.1 to 87.4% F1-score with VGG feature fusion (c.f. Table 2) is minor in comparison to the steep rise in dimensionality (from 512 to 10216 features); although eventually only 60 principal components were selected, PCA was performed on the full feature vector which is of higher computational cost. Along the same lines, appearance-based descriptors performed lower than VGG, but with significantly fewer features; [Hist+Mean+Std] for example comprised 257 (81.9%) features as compared to the 10216 features (87.4%) involved in the feature fusion of VGG. However, given that only the generic VGG was employed in the proposed work, based on its previously reported performance, it is highly probable that after training and tuning the network, the rate of correct depression assessment could improve significantly.

6 Conclusions

The present work introduced the GMHI, a novel variant of MHI and reported on the first application of LMHI [23] on the AVEC dataset. Another novelty of the proposed work is that categorical assessment of depressive symptomatology was performed using deep learning methods, for the first time on this dataset. Although performance achieved here outperforms related work, there is still room for significant improvements. Future work may attempt training and tuning the VGG, testing different versions of the network, assessing the performance of additional classifiers, as well as attempting decision fusion [39] for the different motion representations and feature combinations. Moreover, combining audio-based features [40] could potentially improve overall performance. Testing the two tasks (NorthWind and FreeForm) independently is also of interest. Furthermore, one of the greatest challenges is to test the performance of the proposed pipeline across different benchmark datasets, as well as on the dataset recently collected by our research team.

Funding
Anastasia Pampouchidou was funded by the Greek State Scholarship Foundation, under the scholarship program instituted in memory of Maria Zaousi.

Authors' contributions
AP was the main contributor to the design and overall project coordination, implemented the MHI, LMHI, and GMHI algorithms, and produced the initial draft of the manuscript. MP implemented the Gabor inhibition and wrote the corresponding part of the manuscript. AM performed the feature extraction and contributed to algorithm testing. MA implemented and ran the feature extraction for CNN and wrote the corresponding part of the manuscript. CMV carried the preprocessing step. SS advised on the correct use of PCA and on future plans for improving performance. FM had the overall supervision on the image processing and machine learning methods. MT, PS, KM, FY, and FM had the overall academic supervision. PS, MT, KM, MP, SS, FY, and FM modified the content of the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Le2i Laboratory, University of Burgundy, Le Creusot, France. [2]Institute of Computer Science, Foundation for Research & Technology - Hellas, Heraklion, Crete, Greece. [3]Department of Informatics Engineering, Technological Educational Institute of Crete, Heraklion, Crete, Greece. [4]CISIR, Electrical Engineering Department, Universiti Teknologi PETRONAS, Malaysia. [5]Division of Psychiatry, School of Medicine, University of Crete, Heraklion, Crete, Greece.

Received: 2 June 2017 Accepted: 24 August 2017
Published online: 06 September 2017

References

1. World Health Organization. http://www.who.int/mental_health/management/depression/en. Accessed 8 July 2017
2. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Publishing, Washington, 2013)
3. MB First, *Structured Clinical Interview for DSM-IV-TR Axis I Disorders: Patient Edition*. (Biometrics Research Department, Columbia University, 2005)
4. AT Beck, RA Steer, R Ball, WF Ranieri, Comparison of beck depression inventories-IA and-II in psychiatric outpatients. J. Pers. Assess. **67**(3), 588–597 (1996). doi:10.1207/s15327752jpa6703_13. PMID: 8991972. Accessed 8 July 2017
5. A Pampouchidou, K Marias, M Tsiknakis, P Simos, F Yang, F Meriaudeau, in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. Designing a framework for assisting depression severity assessment from facial image analysis. (2015), pp. 578–583. doi:10.1109/ICSIPA.2015.7412257
6. H Ellgring, *Non-verbal Communication in Depression*. (Cambridge University Press, New York, 2007)
7. PH Waxer, Therapist training in nonverbal communication. I: nonverbal cues for depression. J. Clin. Psychol. **30**(2), 215–218 (1974)
8. A Bobick, J Davis, in *Applications of Computer Vision, 1996. WACV '96., Proceedings 3rd IEEE Workshop On*. Real-time recognition of activity using temporal templates, (1996), pp. 39–42. doi:10.1109/ACV.1996.571995

9. M Valstar, B Schuller, K Smith, F Eyben, B Jiang, S Bilakhia, S Schnieder, R Cowie, M Pantic, in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. AVEC '13*. AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge (ACM, New York, 2013), pp. 3–10. doi:10.1145/2512530.2512533. http://doi.acm.org/10.1145/2512530.2512533. Accessed 8 July 2017
10. M Valstar, B Schuller, K Smith, T Almaev, F Eyben, J Krajewski, R Cowie, M Pantic, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. AVEC '14*. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge (ACM, New York, 2014), pp. 3–10. doi:10.1145/2661806.2661807. http://doi.acm.org/10.1145/2661806.2661807. Accessed 8 July 2017
11. M Valstar, J Gratch, B Schuller, F Ringeval, D Lalanne, M Torres Torres, S Scherer, G Stratou, R Cowie, M Pantic, in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. AVEC '16*. Avec 2016: Depression, mood, and emotion recognition workshop and challenge (ACM, New York, 2016), pp. 3–10. doi:10.1145/2988257.2988258. http://doi.acm.org/10.1145/2988257.2988258. Accessed 8 July 2017
12. J Gratch, R Artstein, G Lucas, G Stratou, S Scherer, A Nazarian, R Wood, J Boberg, D DeVault, S Marsella, D Traum, A Rizzo, L-P Morency, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. The Distress Analysis Interview Corpus of Human and Computer Interviews (LREC, Reykjavik, 2014), pp. 3123–3128
13. T Baltrušaitis, P Robinson, LP Morency, in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Openface: an open source facial behavior analysis toolkit, (2016), pp. 1–10. doi:10.1109/WACV.2016.7477553
14. M Senoussaoui, M Sarria-Paja, JaF Santos, TH Falk, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. AVEC '14*. Model Fusion for Multimodal Depression Classification and Level Detection (ACM, New York, 2014), pp. 57–63. doi:10.1145/2661806.2661819. http://doi.acm.org/10.1145/2661806.2661819
15. S Alghowinem, R Goecke, JF Cohn, M Wagner, G Parker, M Breakspear, in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Cross-cultural detection of depression from nonverbal behaviour, vol. 1, (2015), pp. 1–8. doi:10.1109/FG.2015.7163113
16. A Pampouchidou, K Marias, M Tsiknakis, P Simos, F Yang, G Lemaitre, F Meriaudeau, in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes, (2016), pp. 3835–3838. doi:10.1109/EMBC.2016.7591564
17. A Pampouchidou, O Simantiraki, C-M Vazakopoulou, C Chatzaki, M Pediaditis, A Maridaki, K Marias, P Simos, F Yang, F Meriaudeau, M Tsiknakis, in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Facial Geometry and Speech Analysis for Depression Detection (IEEE, Jeju, Korea, 2017)
18. MAR Ahad, JK Tan, H Kim, S Ishikawa, Motion history image: its variants and applications. Machine Vision and Applications. **23**(2), 255–281 (2012). doi:10.1007/s00138-010-0298-4
19. M Valstar, M Pantic, I Patras, in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*. Motion history for facial action detection in video, vol. 1, (2004), pp. 635–6401. doi:10.1109/ICSMC.2004.1398371
20. H Meng, D Huang, H Wang, H Yang, M Al-Shuraifi, Y Wang, in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. AVEC '13*. Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression (ACM, New York, 2013), pp. 21–30. doi:10.1145/2512530.2512532. http://doi.acm.org/10.1145/2512530.2512532
21. H Pérez Espinosa, HJ Escalante, L Villaseñor-Pineda, M Montes-y-Gómez, D Pinto-Avedaño, V Reyez-Meza, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. AVEC '14*. Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition: INAOE-BUAP's Participation at AVEC'14 Challenge (ACM, New York, 2014), pp. 49–55. doi:10.1145/2661806.2661815. http://doi.acm.org/10.1145/2661806.2661815
22. A Jan, H Meng, YFA Gaus, F Zhang, S Turabzadeh, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. AVEC '14*. Automatic Depression Scale Prediction Using Facial Expression Dynamics and Regression (ACM, New York, 2014), pp. 73–80. doi:10.1145/2661806.2661812. http://doi.acm.org/10.1145/2661806.2661812
23. A Pampouchidou, O Simantiraki, A Fazlollahi, M Pediaditis, D Manousos, A Roniotis, G Giannakakis, F Meriaudeau, P Simos, K Marias, F Yang, M Tsiknakis, in *6th International Workshop on Audio/Visual Emotion Challenge. AVEC '16*. Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text (ACM, Amsterdam, 2016), pp. 27–34. doi:10.1145/2988257.2988266
24. Y-L Tian, T Kanade, JF Cohn, in *Handbook of Face Recognition*, ed. by SZ Li, AK Jain. Facial Expression Analysis (Springer, New York, 2005), pp. 247–275
25. B Fasel, J Luettin, Automatic facial expression analysis: a survey. Pattern Recognition. **36**(1), 259–275 (2003). doi:10.1016/S0031-3203(02)00052-3
26. A Cruz, B Bhanu, NS Thakoor, in *2013 IEEE International Conference on Image Processing*. Facial emotion recognition with anisotropic inhibited Gabor energy histograms, (2013), pp. 4215–4219. doi:10.1109/ICIP.2013.6738868
27. C Grigorescu, N Petkov, MA Westenberg, Contour detection based on nonclassical receptive field inhibition. IEEE Trans. Image Process. **12**(7), 729–739 (2003). doi:10.1109/TIP.2003.814250
28. Y LeCun, Y Bengio, G Hinton, Deep learning. Nature. **521**, 436–444 (2015). doi:10.1038/nature14539
29. O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, A Karpathy, A Khosla, M Bernstein, AC Berg, L Fei-Fei, ImageNet large scale visual recognition challenge. Int. J. Comput. Vision. **115**(3), 211–252 (2015). doi:10.1007/s11263-015-0816-y
30. A Krizhevsky, I Sutskever, GE Hinton, in *Proceedings of the 25th International Conference on Neural Information Processing Systems. NIPS'12*. ImageNet classification with deep convolutional neural networks (Curran Associates Inc., USA, 2012), pp. 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257
31. A Teixeira Lopes, E de Aguiar, AFD Souza, T Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017). doi:10.1016/j.patcog.2016.07.026
32. H Dibeklioğlu, Z Hammal, JF Cohn, Dynamic multimodal measurement of depression severity using deep autoencoding. IEEE J. Biomed. Health Inform. **PP**(99), 1–1 (2017). doi:10.1109/JBHI.2017.2676878
33. Y Zhu, Y Shang, Z Shao, G Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Trans. Affective Comput. **PP**(99), 1–1 (2017). doi:10.1109/TAFFC.2017.2650899
34. M Dahmane, J Meunier, S D'Mello, A Graesser, B Schuller, J-C Martin, in *Continuous Emotion Recognition Using Gabor Energy Filters* (Springer, Berlin, Heidelberg, 2011), pp. 351–358. doi:10.1007/978-3-642-24571-8_46
35. N Dalal, B Triggs, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Histograms of oriented gradients for human detection, vol. 1, (2005), pp. 886–8931. doi:10.1109/CVPR.2005.177
36. T Ojala, M Pietikainen, T Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Machine Intell. **24**(7), 971–987 (2002). doi:10.1109/TPAMI.2002.1017623
37. V Ojansivu, E Rahtu, J Heikkila, in *2008 19th International Conference on Pattern Recognition*. Rotation invariant local phase quantization for blur insensitive texture analysis, (2008), pp. 1–4. doi:10.1109/ICPR.2008.4761377
38. K Simonyan, A Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. S Sfakianakis, ES Bei, M Zervakis, in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. Stacking of network based classifiers with application in breast cancer classification (Springer, 2016), pp. 1079–1084
40. O Simantiraki, P Charonyktakis, A Pampouchidou, M Tsiknakis, M Cooke, in *Proc. Interspeech 2017*. Glottal source features for automatic speech-based depression assessment, (2017), pp. 2700–2704. doi:10.21437/Interspeech.2017-1251