Multimodal Depression Recognition with Dynamic Visual and Audio Cues

Lang He, Dongmei Jiang
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University (NPU)
Shaanxi Key Lab on
Speech and Image Information Processing
127 Youyi Xihu, Xi’an 710072, China
Email: jiangdm@nwpu.edu.cn

Hichem Sahli
NPU-VUB Joint AVSP Research lab
Department of Electronics & Informatics (ETRO)
Vrije Universiteit Brussel (VUB)
Pleinlaan 2, 1050 Brussels, Belgium
Interuniversity Microelectronics Centre (IMEC)
Kepeldreef 75, 3001 Heverlee, Belgium
Email: hsahli@vub.ac.be

Abstract—In this paper, we present our system design for audio visual multi-modal depression recognition. To improve the estimation accuracy of the Beck Depression Inventory (BDI) score, besides the Low Level Descriptors (LLD) features and the Local Gabor Binary Pattern-Three Orthogonal Planes (LGBP-TOP) features provided by the 2014 Audio/Visual Emotion Challenge and Workshop (AVEC2014), we extract extra features to capture key behavioural changes associated with depression. From audio we extract the speaking rate, and from video, the head pose features, the Space-Temporal Interesting Point (STIP) features, and local kinematic features via the Divergence-Curl-Shear descriptors. These features describe body movements, and spatio-temporal changes within the image sequence. We also consider global dynamic features, obtained using motion history histogram (MHH), bag of words (BOW) features and vector of local aggregated descriptors (VLAD). To capture the complementary information within the used features, we evaluate two fusion systems - the feature fusion scheme, and the model fusion scheme via local linear regression (LLR). Experiments are carried out on the training set and development set of the Depression Recognition Sub-Challenge (DSC) of AVEC2014, we obtain root mean square error (RMSE) of 7.6697, and mean absolute error (MAE) of 6.1683 on the development set, which are better or comparable with the state of the art results of the AVEC2014 challenge.

Keywords—depression recognition; spatio–temporal changes; global features; multi-modal fusion

I. INTRODUCTION

Depression and anxiety disorders are highly prevalent worldwide. Studies suggest that effective treatments for depression may be aided by the detection of the problems in its early stages. The affective computing community has shown a growing interest in developing new technologies to assist psychologists in the prevention and treatment of clinical depression, and various systems and studies have been proposed.

In analyzing the speech of depressed subjects, Low et al. [1], [2], [3], [4] have found that depressed subjects are prone to possess a low dynamic range for the fundamental frequency, a slow rate of speech, and a relatively monotone delivery. Consequently, they formulated subtle changes in speech characteristics (e.g. differences in the pitch, loudness, speaking rate, articulation, etc.) as indicators of depression. In the Audio Visual Emotion Challenge and Workshops (AVEC2013 and AVEC2014) [5], [6], low-level descriptors (LLDs) (such as energy, spectrum, and Mel frequency cepstrum coefficients-MFCC) based features were used as baseline audio features, which were proved to be effective for depression recognition. In [7], an i-vector based representation was computed on the extracted LLD features and proved to contain useful information for depression recognition. In [8], Mitra et al. explored a wide array of acoustic features that capture speech articulation, acoustic-phonetic information, spectral representation, speech modulation, vocal effort, rhythmicity, speech prosody, vowel stress, speech intensity, etc. Low-dimensional i-vector features were also computed from these features to convert the frame-level features to a waveform-level representation. Experimental results revealed that i-vector level fusion of low-level features can result in more accurate systems. In [9], Alghowinem et al. found that the average syllable duration was longer in depressed subjects, i.e. depressed subjects speak more slowly than non-depressed people. Therefore, in this paper, besides the LLD features provided by AVEC2014, we also use speaking rate as the audio feature for depression recognition.

Visual features considering body movements and gestures, subtle expressions and periodical muscular movements, have been also widely explored for depression analysis. For describing the dynamic facial appearance, AVEC2013 [5] adopted Local Phase Quantisation (LPQ) feature, while AVEC2014 adopted the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [6], [10] feature as the baseline visual features. In [11], Girard et al. investigated the relationship between nonverbal behavior and severity of depression using Facial Action Coding System (FACS) action units (AUs) and head pose. They found that when symptom severity was high, participants made fewer affiliative facial expressions (AUs 12 and 15), more non-affiliative facial expressions (AU 14), and diminished head motion. In [12] and [13], different representations of head pose were employed for depression classification, and promising results have been obtained.

In [14], Scherer et
al. proposed to use the vertical (head and eye) gaze directionality, smile intensity and average duration, as well as self-adaptors and leg fidgeting, as nonverbal behavior descriptors. Experimental results showed strong correlation of the proposed behavior descriptors with specific psychological disorders.

On the other hand, space-temporal interesting point (STIP) features [15], [16], which describe the spatio-temporal changes by taking into account of the movements from facial area, hands, shoulder, and head etc., has been employed in [12], [17], [18] for depression classification, and excellent performance has been obtained. In [19], Jain et al. adopted the dense trajectory feature by involving tracking densely sampled points from each frame using an optical flow algorithm and thereby capturing motion information of trajectories. However, experiments showed that dense trajectories and LLD features did not significantly reduce the mean absolute error (MAE) and root mean square error (RMSE) when the features are combined with LBP-TOP features. Another visual feature, the Divergence-Curl-Shear (DCS) descriptor, which encodes scalar first-order motion features on the basis of optical flow, and captures physical properties of the flow pattern, provided promising performance in action recognition [20], [21]. However, its application for depression recognition has not been explored yet.

Typical symptoms of depression can be well described by global variation information, e.g. disorder changes in facial expression and variations of voices, which can reflect depression suffers’ disability of relaxation, irritability, fretfulness, and slurred-slow-monotonous speech. Consequently, most approaches in depression analysis extract global feature vectors from the complete video under analysis by aggregating a large set of the above described local descriptors. Motion History Histogram (MHH) is a descriptive representation of temporal-series, which achieved good performance in human action recognition [22] and depression analysis [23], [24]. Moreover, it has been proved in [25] that MHH can capture more sufficient dynamic information than many well-known motion features like Motion History Image (MHI) [26]. To extract global features from frame-level video and audio cues, in this work, we use MHH on LGBP-TOP, head pose estimated from the image frames, and on the LLD features extracted from audio signals.

In [17] and [18], the bag of words (BOW) feature was adopted for global feature extraction in depression analysis, which quantizes the local features using a codebook obtained via clustering. It can be applied to the space-temporal interesting point features and the DCS descriptors to capture the global dynamic features without other extra requirements. Similar to BOW, Vector of Local Aggregated Descriptors (VLAD) [27] relies on a codebook to capture global features. VLAD aggregates local feature descriptors according to a locality criterion in the feature space. It has been successfully used for action recognition [21]. To our knowledge, this technique has not yet been utilized for depression recognition.

Most of the research work in depression recognition integrate vocal and visual cues [9], [17], [23], [28], [24], [29], [30], [19], [8], [31]. Experimental results showed that the performance is highly improved compared to the single modal depression recognition systems.

The contributions of this paper are twofold:

- To capture the key behavioural changes associated with depression, we adopt MHH estimated on LGBP-TOP and LLD audio features, BOW or VLAD estimated on STIP features, and BOW estimated on head pose and DCS descriptors, as feature inputs for depression recognition. In the experiments, different combinations of these features are used to estimate the Beck Depression Inventory-II (BDI-II) value of a subject, results demonstrate the benefits of using these features.
- Two multi-modal fusion schemes – feature fusion scheme and model fusion scheme are proposed and compared to improve the accuracy of depression recognition by integrating the multi-modal features.

Experiments are carried out on the training and development sets of the Depression Recognition Sub-Challenge (DSC) task of AVEC2014, and results are compared to the state of the art results of the AVEC2014 challenge. Results show that our proposed model fusion scheme, using the proposed audio-visual features, outperforms the state-of-art methods.

The remainder of the paper is organized as follows. Section II addresses the audio visual multi-modal fusion schemes. Section III and IV describe the proposed visual and audio features, respectively. Section V gives the global feature extraction methods. Experiments and results are analyzed in section VI. In section VII, we draw conclusions.

II. AUDIO VISUAL DEPRESSION RECOGNITION

To make complementary use of the audio and visual features, in this paper, we propose two multi-modal fusion schemes, and evaluate their performances on depression recognition. The first scheme, as shown in Figure 1, is a model fusion strategy using local linear regression (LLR) [32], in which each stream of the extracted audio and visual features is firstly processed via principle component analysis (PCA) to reduce the redundancy, then input into a separate epsilon support vector regression (e-SVR) model with intersection kernel, the predicted BDI-II values [5], [6] are linearly combined via a LLR model for the final recognition.

The second fusion scheme is the feature-based one. For each video sequence, the processed audio and visual features via PCA are concatenated into a high dimensional feature vector. For the sake of modeling the nonlinearity of depression regression, SVR with intersection kernel [33], is employed to predict the BDI-II scores.

The adopted features in Figure 1 are described in the following sections.

III. VISUAL FEATURES

Integration of different behavioral cues is of particular importance in improving the recognition accuracy of depression severity. In this paper, apart from the baseline LGBP-TOP feature provided by AVEC2014, we adopt various visual
features including dynamic information of head and body movements, and verify their validity in the multi-modal depression recognition experiments.

A. Divergence-Curl-Shear (DCS) Descriptors

Divergence-Curl-Shear descriptors, capturing physical properties of the flow pattern, encode scalar first-order motion features, namely the motion divergence, curl and shear [21]. Firstly, four local first-order differential scalar quantities are computed on the optical flow field of the video, including the divergence, the curl and the two hyperbolic terms. According to [21], they can be formulated by the first-order derivatives of the optical flow at the point \( p_t \) as

\[
\text{div}(p_t) = \frac{\partial (u(p_t))}{\partial (x)} + \frac{\partial (v(p_t))}{\partial (y)}
\]

(1)

\[
\text{curl}(p_t) = -\frac{\partial (u(p_t))}{\partial (y)} + \frac{\partial (v(p_t))}{\partial (x)}
\]

(2)

\[
\text{hyp}_1(p_t) = \frac{\partial (u(p_t))}{\partial (x)} - \frac{\partial (v(p_t))}{\partial (y)}
\]

(3)

\[
\text{hyp}_2(p_t) = \frac{\partial (u(p_t))}{\partial (y)} + \frac{\partial (v(p_t))}{\partial (x)}
\]

(4)

where \((u(p_t), v(p_t))\) denotes the optical flow vector at the point \( p \) at time \( t \). Then the shear quantity can be written as

\[
\text{shear}(p_t) = \sqrt{\text{hyp}_1^2(p_t) + \text{hyp}_2^2(p_t)}
\]

(5)

The divergence is related to axial motion, expansion and scaling effects, the curl to rotation in the image plane. The two hyperbolic terms express the shear of the visual flow corresponding to more complex configuration. Following the ideas of [21], we consider all possible pairs of kinematic features, namely (\text{div}, \text{curl}), (\text{div}, \text{shear}) and (\text{curl}, \text{shear}). At each pixel, the orientation and magnitude of the 2-D vector corresponding to each of these pairs are calculated. Finally, as descriptors, 8-bin normalized histograms for each of the three feature pairs or components of Divergence-Curl-Shear, are obtained. In our implementation, the DCS descriptors are estimated in each frame, and all the histograms are linearly normalized to [0, 1].

B. Space-Temporal Interesting Point (STIP)

Space-Time Interests Points (STIP) detects salient points, where image values have sufficient local variation in both the space and time domain. Subsequently, two histograms, the Histogram of Gradients (HOG) and the Histogram of Flow (HOF), are calculated around an interest point in a fixed sized spatial and temporal window. The detection results are as shown in Figure 2. Recent studies have proved that STIP is robust to temporal misalignments within the spatio-temporal feature space [12], [17], [18], [28]. In our implementation, STIP features are estimated on each image frame.

C. Head Pose Features

Many computer vision approaches have been proposed for automatic head pose estimation, such as template matching, deformable models, geometry determination among local features [34], 3D cylinder fitting [35], and estimation of 3D head
pose by combining Active Appearance Model (AAM) with Pose from Orthography and Scaling with Iteration (POSIT) [36], [37] etc. Recently, Supervised Descent Method (SDM) has been proposed in [38] to improve the accuracy of 3D pose estimation. To obtain accurate head pose estimation, in this work, we adopt SDM, which has been implemented and integrated into the software Interface [39]. In our implementation, head pose is estimated for each image frame.

IV. AUDIO FEATURES

For audio features, apart from the low-level descriptor (LLD) baseline features provided by the AVEC2014 challenge, we propose using the speaking rate. The speaking rates, or the articulation rates, are extracted from the manually labelled intervals for further statistical analysis. Specifically, Praat script [40] is utilized to calculate the speech and articulation rates as well as the pause rate. When measuring the speech rate, pauses are included in the duration time of the utterance, while the articulation rate excludes pauses.

V. GLOBAL FEATURE EXTRACTION

For each audio or video sequence, the above descriptors are encoded into a single vector representation using MHH, BOW, or VLAD.

A. Motion History Histogram (MHH) Representation

MHH is implemented following the approach described in [41], but the difference is that here we use MHH for statistical quantifications of the frame-level features rather than images. Let \( \{ f(e, k), e = 1, \ldots, N, k = 1, \ldots, K \} \) be the set of \( K \) frames of features with dimension \( N \), and \( \{ D(e, k), e = 1, \ldots, N, k = 1, \ldots, K \} \) be a binary matrix, in which if \( f(e, k) - f(e, k - 1) \) is greater than a threshold, \( D(e, k) = 0 \), otherwise \( D(e, k) = 1 \). We also define the patterns \( \{ P_i, 1 \leq i \leq M \} \) in the \( \{ D(e, k), k = 1, \ldots, K \} \) sequences, based on the number of connected “1”s, e.g. \( P_m = 010 \) with “0”s as the start and end values and \( m \) successive “1”s in between. \( I(e) \) is a frame index standing for the number of the starting frame of a new pattern on dimension \( e \). At the beginning, \( I(e) = 1(e = 1, \ldots, N) \), meaning that a new pattern starts from frame 1 for each feature dimension. While \( \{ D(e, I(e)), \ldots, D(e, k) \} \) forms \( P_i(1 \leq i \leq M) \), MHH \( H(e, i) \) increases by 1, and \( I(e) \) will be updated to \( i = k \). The total pattern number \( M \), which represents the patterns of movement, should be defined before starting the computing process. In our work, different values of \( M \) are set up to get the best depression recognition results. In our implementation, MHH is applied to the LLD features of audio, and LGBP-TOP features as well as head pose features of video, denoted as Audio-MHH, LGBP-MHH, and Head-MHH, respectively.

B. Bag of Words Features (BOW)

Producing a BOW representation is performed by first extracting local descriptors, as described in the previous sections. Clustering of the descriptors is then performed by a k-means quantizer with \( K \) clusters. In our implementation, BOW is applied to the STIP features, DCS features and head pose features, denoted as STIP-BOW, DCS-BOW, and Head-BOW, respectively.

C. Vector of Local Aggregated Descriptors (VLAD)

Similar to BOW, VLAD [21] relies on a codebook \( C = \{ c_1, c_2, \ldots, c_K \} \) of \( K \) centroids learned by k-means. For each class \( c_i \), the representation is obtained by summing up the differences \( x - c_i \) of the vectors \( x \) assigned to \( c_i \), thereby producing a vector representation of length \( d \times K \), where \( d \) is the dimension of the local descriptors. Finally, the obtained sums of all classes are concatenated to form a feature vector and then normalized using \( L2 \) norm. In our implementation, VLAD is applied to the STIP features and head pose features, denoted as STIP-VLAD and Head-VLAD, respectively.

VI. EXPERIMENTS AND ANALYSIS

A. AVEC2014 Challenge and Baselines

We apply the model fusion scheme and feature fusion scheme on the Depression Recognition Sub-Challenge (DSC) of AVEC2014. As described in [6], the challenge uses a subset of the AVEC2013 audio-visual depression corpus, which is formed of 150 videos of task oriented depression data recorded by a webcam and a microphone in a human-computer interaction scenario. The recordings in the AVEC2014 subset consist of only 2 of the 14 tasks present in the original recordings, to allow for a more focused study of affect and depression analysis. Both tasks are supplied as separate recordings, resulting in a total of 300 videos (ranging in duration from 6 seconds to 4 minutes 8 seconds).

The two selected tasks are as follows:

- Northwind - Participants read aloud an excerpt of the fable "The North Wind and the Sun", spoken in the German language.
- Freeform - Participants respond to one of a number of questions such as "What is your favourite dish?", etc., again in the German language.

For each video, the severity level of depression is labeled with the Beck Depression Inventory-II (BDI-II) value [6]. We should notice here that the BDI-II scores here are self-reported depression symptom severity scores, and not the measures of diagnosed clinical depression. 150 Northwind-Freeform pairs of 84 subjects, totalling 300 task recordings, are split into three partitions: a training, a development, and a test set. Tasks are split equally over the three partitions. Care is taken to have similar distributions in terms of age, gender, and depression levels for the partitions. There is no session overlap between partitions, i.e. multiple task recordings taken from the same original clip would be assigned to a single partition. As the ground truth BDI-II labels of the test set are not known to us, in our experiments, we use the training sets of the Freeform dataset and Northwind dataset to train the models, and the development sets to verify the performance. The baseline recognition results with the baseline audio visual features on AVEC2014 are given in [6]: on the development set, the root mean square error (RMSE) between the estimated
BDI-II values and the ground-truth labels is 9.26, while on the test set, the RMSE is 10.859 and the mean absolute error (MAE) is 8.857.

In our experiments, to alleviate the possible over-fitting on the development set, we train the models with different parameter sets for many times and do depression recognition experiments on the development set, the parameters, which obtain the consistent (comparable) and good results on both the training set and development set, are chosen as the final parameters.

B. Single Modal Depression Recognition

We firstly carry out depression recognition (BDI-II estimation) experiments using the single feature stream, on the Northwind task and the Freeform task, respectively. In the process of extracting the features, some parameters have significant influence on the results, such as the size $K$ of the codebooks in BOW and VLAD, as well as the threshold motion level $M$ for MHH. Therefore we conduct a few trial-and-errors attempts to find the optimum parameters required for each feature. Table I summarizes the recognition results by RMSEs, along with the optimal parameters required for particular features. As it can be seen, for the Northwind dataset, the proposed LGBP-MHH feature obtains the best estimation results. While for the Freeform dataset, the LGBP-mean feature, used as the AVEC2014 baseline feature, obtains the best results. Among the proposed features of this paper, the DCS-BOW feature obtains the lowest RMSE on both Northwind and Freeform database, followed by the Audio-MHH feature.

C. Multi-modal Depression Recognition

For evaluating the multi-modal depression recognition schemes with the proposed features, for each dataset, we adopt the features which obtained the optimal results in single modal recognition (Table I). For example, for the LGBP features, we adopt LGBP-MHH feature for the Northwind dataset and LGBP-mean for the Freeform dataset. While for the STIP features, we adopt STIP-BOW for Northwind and STIP-VLAD for Freeform. Head-BOW features are adopted for both Northwind and Freeform datasets.

For each dataset, we rank the features which obtained good performances in the single modal recognition task, and input them into the feature fusion or modal fusion scheme. Estimation results of BDI-II on the development set of Northwind, with different combinations of the top ranked features, are shown in Table II, and those on the development set of Freeform are given in Table III. From Table II, one can notice that on the Northwind dataset: 1) the model fusion scheme obtains overall better performance than the feature fusion scheme. 2) for the feature fusion scheme, the performance firstly improves with the gradual increasing of feature types. When the inputs contain the LGBP-MHH, DCS-BOW, Audio-MHH, and STIP-BOW features, the system obtains the best performance. But when the Head-BOW feature and Speaking rate feature are further added, the performance will reduce. For the model fusion scheme, the best performance is obtained with LGBP-mean and DCS-BOW features, while for the feature fusion scheme, the best performance is obtained with all the proposed features as inputs. Moreover, the model fusion scheme obtains much better performance than the feature fusion scheme.

We further calculate the RMSEs and MAEs on the estimation results of the Northwind-Freeform pairs. For the feature fusion method, we calculate on the estimation results from feature combination C for Northwind and those from feature combination A1 for Freeform. While for the model fusion method, RMSE and MAE are calculated on the estimation results from feature combination B for Northwind and those from feature combination E1 for Freeform. Results are shown in Table IV. One can see that the model fusion scheme obtains very promising results, with RMSE reduced to 7.6697 from the baseline result of 9.26, and MAE to 6.1683 from 7.2902 of the feature fusion method.

Comparison of our best results from the model fusion
Depression is a serious psychological disorder. Computer-aided technologies have been investigated to assist psychologists in the prevention and treatment of clinical depression. In this paper, to improve the accuracy of automatic depression recognition from audio and visual cues, we propose several features to capture key behavioural changes associated with depression, such as head pose features, STIP features and Divergence-Curl-Shear descriptors, which allow describing body movement, spatio-temporal changes of the video, as well as optical flow characteristics. Moreover, motion history histogram (MHH), bag of words (BOW) and vector of local aggregated descriptors (VLAD) are performed on these features for global feature extraction. Finally, a feature fusion scheme and a model fusion scheme are proposed. Experimental results show that the model fusion scheme with the proposed audio visual features obtains better or comparable results than the state of the art results of the AVEC2014 challenge, with RMSE reduced to 7.6697 from 9.26 of the baseline result given by AVEC2014. In our future work, we will explore better feature representation methods, and more powerful regression models to further improve the accuracy of depression recognition.

ACKNOWLEDGMENT
This work is supported by the framework of the National Natural Science Foundation of China (grant 61273265) and the VUB Interdisciplinary Research Program through the EMO-App project.

REFERENCES


