Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes

Anastasia Pampouchidou, Kostas Marias, Manolis Tsiknakis, Panagiotis Simos, Fan Yang, Guillaume Lemaître, Fabrice Meriaudeau

Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes

Anastasia Pampouchidou[1], Kostas Marias[2], Manolis Tsiknakis[3],
Panagiotis Simos[4], Fan Yang[1], Guillaume Lemaître[1], Fabrice Meriaudeau[5]

***Abstract*— Depression is an increasingly prevalent mood disorder. This is the reason why the field of computer-based depression assessment has been gaining the attention of the research community during the past couple of years. The present work proposes two algorithms for depression detection, one Frame-based and the second Video-based, both employing Curvelet transform and Local Binary Patterns. The main advantage of these methods is that they have significantly lower computational requirements, as the extracted features are of very low dimensionality. This is achieved by modifying the previously proposed algorithm which considers Three-Orthogonal-Planes, to only Pairwise-Orthogonal-Planes. Performance of the algorithms was tested on the benchmark dataset provided by the Audio/Visual Emotion Challenge 2014, with the person-specific system achieving 97.6% classification accuracy, and the person-independed one yielding promising preliminary results of 74.5% accuracy. The paper concludes with open issues, proposed solutions, and future plans.**

I. Introduction

Depression is one of the most prevalent psychiatric disorders, typically associated with negative emotions, such as sadness and helplessness. Major depressive disorder (MDD) in particular is a foremost cause of disability, often leading to suicidal ideation. The emergence of depression is precipitated by a combination of diverse factors, including biological predispositions, environmental triggers, and psychological vulnerability traits. Screening instruments (self-report questionnaires) are helpful to document depression symptoms although responses often suffer from subjective bias. In addition, multiple assessments are typically required to support diagnosis. Modern technology has the means to deliver a non-obtrusive framework for steady follow-up on high-risk individuals. The present work introduces such a framework, beginning with a short review of selected approaches in Section II, and continues with the description of the proposed methodology in Section III. The data used for the validation of the presented approach are described in Section IV, followed by the experimental results in Section V. Findings are discussed in Section VI, closing with conclusions and future plans in Section VII.

[1]A. Pampouchidou, F. Yang and G. Lemaître are with Le2i Laboratory, University of Burgundy, Le Creusot, France. anastasia.pampouchidou@gmail.com, g.lemaitre58@gmail.com, fanyang@u-bourgogne.fr

[2]K.Marias is with the Foundation for Research & Technology - Hellas, Heraklion, Crete, Greece kmarias@ics.forth.gr

[3]M. Tsiknakis is with the Technological Educational Institute of Crete, Department of Informatics Engineering and with the Foundation for Research & Technology - Hellas, Heraklion, Crete, Greece tsiknaki@ics.forth.gr

[4]P. Simos is with the Department of Psychiatry, University of Crete, Heraklion, Crete, Greece akis.simos@gmail.com

[5]F. Meriaudeau is with Le2i Laboratory, University of Burgundy, Le Creusot, France and with CISIR, Electrical Engineering Department, Universiti Teknologi Petronas, Malaysia. fmeriau@u-bourgogne.fr

II. Related Work

Automatic depression assessment is an emerging domain, which recently started to attract the interest of the research community. Cohn et al. 2009 [1] detected depression from facial actions with the use of Active Appearance Models (AAMs) and vocal prosody with 79% accuracy. Cohn noted that AAMs limitation lies on the fact that they need to be tuned to person specific data before being applied. Joshi et al. 2013 incorporated speech analysis and body parts movement (independent and relative) along with facial activity, and achieved an accuracy of 91.7% [2]. Alghowinem et al. 2015 combined features extracted from eye movements, and head pose and movement, testing their approach on datasets from three different countries to establish for a cross-cultural equivalence [3].

Audio/Visual Emotion Challenges (AVEC) 2013 and 2014 [4] attracted several participations. In terms of the depression sub-challenge, the teams were required to predict individual scores on the Beck Depression Inventory based on their video recording. In one such approach, Senoussaoui et al. [5] grouped the subjects into two classes (depressed/non-depressed), as a preprocessing step, based on a 13/14 point cutoff, demonstrating accuracy of 82%.

III. Methodology

The aim of the proposed methodology is to detect depression based on video recordings. The first step is to initialize the face region at the first frame of the video, and then track it for the rest of the video using the Kanade-Tomasi-Lucas (KLT) tracker as described in [6]. For every frame, the extracted face region is processed by the Curvelet Transform [7], producing a pseudo-image (CurveFace). Curvelet transform extracts curvature information from an image; such information is useful, as different facial expressions can be differentiated from their curves (e.g. mouth corners are angled down in a sad expression). Local Binary Pattern descriptor (LBP) [8] is computed for each individual CurveFace in order to form the feature vector for the *Frame-based Classification* (see Fig.1).

However, apart from facial expression, curvature contains information on person-specific biometrics (e.g. different shapes of facial features, varying symmetry), as well

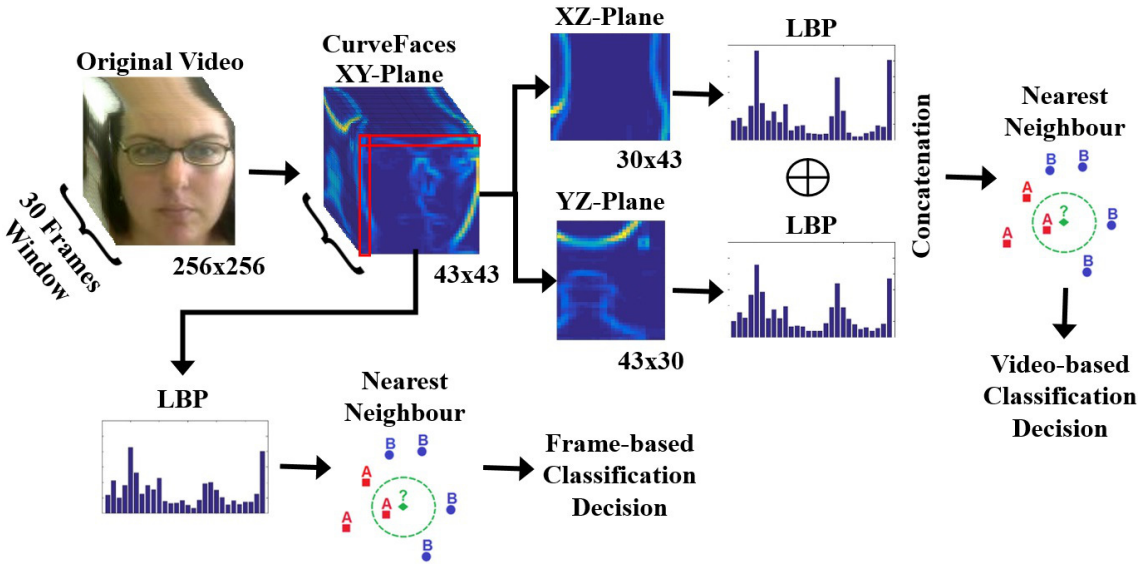

Fig. 1. Proposed framework with data flow

as occlusions (e.g. facial hair, eye-glasses). *Video-based Classification* was implemented in order to overcome this limitation. The idea is inspired from [9] "Local Gabor Binary Patterns in Three Orthogonal Planes" (LGBP-TOP), where frame-based Gabor transforms were combined for a window of frames, creating different planes in time (XZ and YZ); hereby Curvelet Transform is used instead of Gabor. Thus, for an overlapping window of frames (empirically set), each row and each column are taken over time, forming new planes which incorporate motion information; this is achieved by portraying the variation of the values over time. LBP descriptor is again computed on each plane.

The proposed work introduces Pairwise Orthogonal Planes. The norm for approaches based on Orthogonal Planes is to take all descriptors from each of the three planes (XY, XZ, YZ) and concatenate them all together, producing a vast vector of thousands of features. XY planes are considered only for the *Frame-based Classification*, for each frame separately, while for the *Video-based Classification* XZ and YZ planes are considered in pairs of two. This way, the plane corresponding to the first row is combined with the plane corresponding to the first column, second row with second column, (...), and the last row with the last column. This modification has the advantage of preserving the motion information in both axis, with a considerably shorter feature vector. The classifier used in both cases is the well-known Nearest Neighbour. The proposed framework is illustrated in Fig.1 and Fig.2, and the specific parameters of the different algorithms are explained in Section V. With a careful observation of the XZ and YZ planes, along with the sequence of CurveFaces, the motion patterns formed for the first row and column can be observed in both X and Y axes respectively.

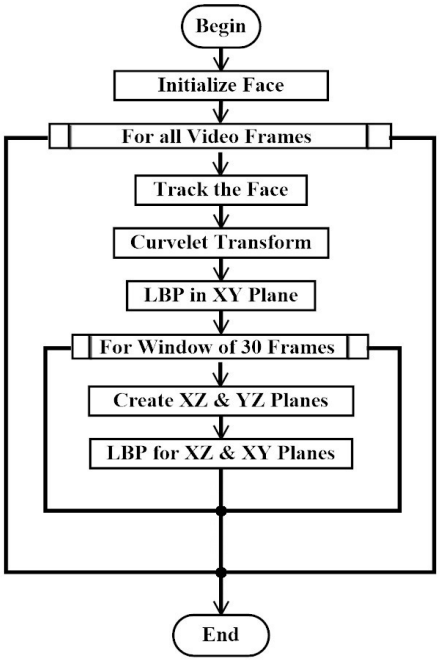

Fig. 2. Flowchart for the feature extraction procedure

IV. Data Organization

The proposed method was validated on the dataset provided by AVEC'2014, which is the only freely available dataset of facial videos annotated for depression symptomatology. Volunteer participants were recorded by a webcam while executing the following tasks presented in Power Point: a) FreeForm task, where participants answered questions such as: Discuss a sad childhood memory, and b) NorthWind task, where they read aloud an excerpt from a fable. The recordings were split into three partitions: training, development, and test set, of 150 NorthWind-Freeform pairs, totaling

300 recordings. Depression labels for the test set were not released to the public. Labels are known for 200 recordings, organized in 4 subsets: 50 FreeForm-Development + 50 FreeForm-Training + 50 NorthWind-Development + 50 NorthWind-Training. In the present work subsets were merged.

The purpose of the Challenge was to predict participants' score on the Beck Depression Inventory (BDI). The standard cut-off scores of the BDI are: 0-9 (minimal depression), 10-18 (mild depression), 19-29 (moderate depression), and 30-63 (severe depression). Participants who completed the tasks at different times completed the BDI as many times.

V. Experimental Results

In order for any false classification to be attributed solely to the feature extraction method, 100% accurate face detection had to be established; thus a semi-automatic face detection algorithm was implemented. Face region was manually initialized, and then tracked with the KLT, which is set to fail and to be reinitialized if the tracked points are below a threshold (20 points). Tracking fails when face goes out of the field of view, because of occlusions (e.g. hand in front of the face), or when illumination becomes too inadequate even for a human observer to distinguish facial features. Such issues are met in about 20 videos out of the total 200. The extracted facial region is resized to 256x256 pixels, followed by the Curvelet transform, which can be performed for different values of Orientation and Scale [7]. Here both parameters are set to 1, resulting to a 43x43 CurveFace. LBP descriptors are extracted for two sets of [Radius, Neighbourhood] [8]; LBP1=[1,8] and LBP2=[2,16]. LBP1 gave a 10 bin histogram, and LBP2 an 18 bin histogram; the two were concatenated to a 28 element feature vector for each frame in the *Frame-based Classification*.

For the *Video-based Classification* the algorithm moves one step ahead, by computing the XZ and YZ planes, for a window of 30 subsequent frames, with an overlap of 15 frames. Therefore, a set of 30 CurveFaces of 43x43, results in 43 XZ planes of 30x43, and 43 YZ of 43x30. LBP1 and LBP2 are again applied, to provide 43 pairs $\left(\bigcup_{i=1}^{43}(XZ_i \bigoplus YZ_i)\right)$ of LBP1$\oplus$ LBP2 descriptors. That is for every window 43 different feature vectors of 56 elements are being extracted, each of which is being treated as an individual sample for the classifier.

Following Senoussaoui's approach, in making the problem binary, all three different cut-offs were experimented. {minimal/mild}, {mild/moderate}, {moderate/severe}, in order to find the best for detecting depression. However, the 4 subsets were highly unbalanced, with the 'minimal' class having as many recordings as all the rest together. Consequently random data-sampling was used in order to keep equal number of samples from each class.

Two main sets of experiments took place, with either 20-Fold or Leave-One-Out cross validation methods. The 20-Fold classification took place for each individual sample, with the sets being partitioned 20 times, that is 20 different randomly selected train/test sets were used in order to validate the method. In the Leave-One-Out method, all recordings belonging to the same participant were excluded from the training process, and were just used for testing. There were 58 different participants in all 200 video recordings. For the Leave-One-Out method classification of the videos was based on the class that was attributed to the majority of the samples.

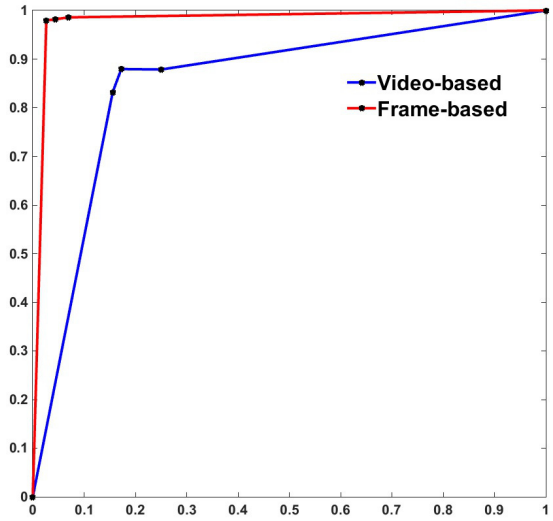

Fig. 3. ROC curve for both of the proposed algorithms, with the 3 different cut-offs (marker points), using the 20-Fold cross validation

TABLE I
Frame & Video Based Classification Results Accuracies %

	Leave-One-Out		20-Fold	
	Frame	Video	Frame	Video
Minimal/Mild	60.5	74.5	97.6	83.8
Mild/Moderate	59.0	63.5	96.9	85.4
Moderate/Severe	72.5	74.5	95.8	81.3

TABLE II
All four classes with the Frame-based algorithm and 20-Fold cross validation %

	Minimal	Mild	Moderate	Severe
Minimal	89.3	3.6	3.7	3.4
Mild	1.9	95.3	1.4	1.4
Moderate	1.3	1.2	96.1	1.5
Severe	0.8	1	0.8	97.4

Results for both *Frame-based* and *Video-based* classification algorithms, and all three cut-offs, for both cross validation methods are summarized in Table I. Table II presents the confusion matrix for all four classes, for the *Frame-based* algorithm with 20-Fold. Finally, Fig. 3 illustrates the Receiver Operating Characteristic (ROC) curve for both *Frame* and *Video* based algorithms, for all three different cut-offs, and with 20-Fold cross validation.

VI. Discussion

In properly evaluating the present results, issues regarding the data set need to be stressed. Having to work on a highly unbalanced dataset can become an issue for classification problems, and down-sampling is not guaranteeing that the selected samples are the most representative. Furthermore, when working with visual data, acquisition conditions are of great importance, as they often affect processing tasks such as face detection. To this end, illumination, image resolution, frame rate, are some crucial factors to be considered. It is also important to stress that the participants in the AVEC dataset are not diagnosed patients but just volunteers, with BDI scores varying significantly through the different recordings of individual participants. This means that a simple score on an instrument is not necessarily enough to establish the diagnosis of depressed; in reality it is possible that some of the subjects ranked as severely depressed could just be under great stress. Thus, a clinical interview would be more reliable for building a robust dataset.

In terms of the proposed algorithm, keeping the feature vector considerably low can be counted as an achievement, since similar approaches in this area tend to have huge feature vectors. More specifically, it is actually usual to have feature vectors of thousands, and then need to use dimensionality reduction algorithms, while hereby the *Frame-based* approach has a vector of 28 features, and the *Video-based* of 56 features. This trait of the proposed algorithm is providing the potential for real-time performance, an important attribute for clinical applications.

Moreover, the different cross validation methods, 20-Fold and Leave-One-Out, can also be considered as Person-specific and Person-independent, respectively. This is due to the fact that for the 20-Fold the samples were being divided randomly, not guaranteeing that samples from the same participant were not being used in both training and testing; it is however ascertained that the same sample was never included in both sets. Having such a high performance for the 20-Fold (with accuracy as high as 97.6% for the binary classification problem, and a sensitivity of 97.4% for the four-way classification) suggests that the proposed method could perform very well as a personalized system, by simply requiring a calibration in order to obtain a person-specific baseline, which is a practice generally applied in medicine (e.g. even heart rate baselines can vary significantly among different individuals). The Leave-One-Out method tested the functionality of the system without prior knowledge for a given individual; results of this approach might not be as high as the person-specific ones, yet they are indeed promising with 74.5% accuracy being well above chance, which shows that there is true potential in this algorithm. It can also be observed that the difference in performance (person-specific vs. person-independent) is achieved by the different algorithms. This could be attributed merely to the fact that the *Frame-based* approach keeps information related to biometrics and static expressions that tend to vary among different people, while the *Video-based* considers the movements that take place within the region of the face or even the movements of the face itself, which can be generalized more easily. Finally regarding the different cut-offs, the best results were obtained by separating the samples into {minimal} and {mild, moderate, severe}, which means that total absence of signs related to depression was highly distinct from even mild ones.

VII. Conclusions & Future Work

The present work proposed two different algorithms, both of low dimensionality, and considerably high performance for the person-specific one, which functions almost flawlessly. In the person-independent approach, on the other hand, there is room for improvement; different classifiers, combination of additional descriptors, or even combining the two approaches (*Frame* and *Video* based) are some of the ideas that could be implemented in order to increase the classification accuracy, as well as to provide a comparative study. Finally, in order to be able to test the algorithms developed with more robust data, a data acquisition campaign with clinically diagnosed patients, is being designed and planned by our group.

References

[1] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De La Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[2] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on MultiModal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[3] Sharifa Alghowinem, Roland Goecke, Jeffrey F Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 1, pp. 1–8.

[4] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[5] Mohammed Senoussaoui, Milton Sarria-Paja, João F Santos, and Tiago H Falk, "Model fusion for multimodal depression classification and level detection," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 57–63.

[6] Carlo Tomasi and Takeo Kanade, *Detection and tracking of point features*, School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[7] Emmanuel Candes, Laurent Demanet, David Donoho, and Lexing Ying, "Fast discrete curvelet transforms," *Multiscale Modeling & Simulation*, vol. 5, no. 3, pp. 861–899, 2006.

[8] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[9] Timur R Almaev and Michel F Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 356–361.