Neural-Net Classification For Spatio-Temporal Descriptor Based Depression Analysis

Jyoti Joshi¹, Abhinav Dhall², Roland Goecke¹,², Michael Breakspear³,⁴ and Gordon Parker³

University of Canberra¹, Australian National University², University of New South Wales³, Queensland Institute of Medical Research⁴

jyoti.joshi@canberra.edu.au, abhinav.dhall@anu.edu.au, roland.goecke@ieee.org, mjbreaks@gmail.com, g.parker@blackdog.org.au

Abstract

Depression is a severe psychiatric disorder. Despite the high prevalence, current clinical practice depends almost exclusively on self-report and clinical opinion, risking a range of subjective biases. This paper focuses on depression analysis based on visual cues from facial expressions and upper body movements. The proposed diagnostic support system is based on computing spatio-temporal features from video sequences. Space Time Interest Points are computed for the videos for analysing the upper body movements and a temporal visual words dictionary is learned from them. Intra-facial muscle movement is captured by computing a LBP-TOP based codebook. Various neural-net classifiers are explored and compared with a SVM. The approach is evaluated on real-world clinical data from interactive interviews with depressed and healthy subjects.

1. Introduction

Depression is one of the most common and disabling mental disorders with a major impact on society. The landmark WHO 1990 Global Burden of Disease (GBD) report by Murray et al. [1] quantified depression as the leading cause of disability worldwide. The lifetime risk reported for depression is at least 15%. Despite its high severity, there currently exist no laboratory-based measures of illness expression, course and recovery. This compromises optimal patient care, compounding the burden of disability. The proposed framework is based on extracting spatio-temporal features from changes in affective state of a subject. Experiments are conducted on a real clinical database collected from subjects from the depression and bipolar clinics at the Black Dog Institute, Sydney, Australia.

Computer-based analysis of depression can be broadly categorised into the following main classes: video-only, audio-only, physiological measures only and multimodal systems. Researchers have been using physiological measures such as EEG [2], MRI [3], etc. to analyse the affective state of a person. There has been a considerable amount of work for audio-based depression analysis [4, 5].

The work proposed in this paper draws inspiration from the research in the facial expression analysis community. Facial expressions are dynamic in nature and convey a lot of information about the affective state of a person. Ellgring [6] proposed the hypothesis, based on clinical and systematic observations, that there is a significant amount of reduction in facial activity during depression and that increases with improvement of subjective well-being. McIntyre et al. [7], motivated by Ellgring’s work, analysed the facial response of the subjects by showing them a short video clip. They computed subject-specific Active Appearance Models (AAM) [8] and extracted shape features every fifth frame. Shape features from the chosen frames were combined and a Support Vector Machine (SVM) was used for classification. However, they only considered geometric (shape) features and performed frame level classification. It has been shown in the literature that temporal facial expression analysis provides more information than using static information alone [9].

Cohn et al. [10] explored the relationship between facial and vocal features and clinical depression detection. For the analysis of facial behaviour, manual Facial Action Coding System (FACS) [11] annotation and AAM [8] modelling were used. In the FACS-based method, parameters such as occurrence of Action Units (associated with depression), mean duration, ratio of onset to total duration and ratio of offset to onset phase were computed. For the method based on AAM mod-
elling, person-specific models were learned. They computed the frame-to-frame difference velocity based on the shape vectors for a fixed window of length $10s$. Vocal behaviour analysis was performed by pitch extraction and, thereby, the results were compared with facial behavioural changes.

The method proposed in this paper is based on extracting spatio-temporal features based on Space Time Interest Points (STIP) and Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) from videos and modelling them in a Bag-of-Words (BoW) framework. For classification, various neural network classifiers such as Probabilistic Neural Networks (PNN), Feed Forward Neural Networks (FFNN) and Restricted Boltzmann Machines (RBM) are explored and a comparison is made with a Support Vector Machine (SVM).

The proposed framework is based on bags of temporal words. LBP-TOP is a spatio-temporal texture descriptor. However, the XY-plane should be near the apex frame. Our proposed framework computes LBP-TOP based spatio-temporal words. [12] proposed a temporal word based framework where LBP-TOP was calculated for cuboids around interest points calculated from STIP for human action recognition. The framework proposed in this paper draws inspiration from [12]. However, video clips are divided into subclips for computing LBP-TOP for computing visual words based histograms.

In the case of STIP feature computation, in order to overcome the high memory requirement due to the high number of interest points generated by video samples in our database, a two-level clustering mechanism is proposed. Level one clustering is used for selecting the representative interest points. This is inspired by keyframe selection from [13], where the authors proposed a facial landmark point based clustering method for keyframe selection.

## 2 Data Collection

The clinical data used in the experiments was collected at the Black Dog Institute\(^1\), Sydney, Australia as part of an ongoing study. 60 subjects were interviewed. In the interview, the subjects were asked to describe events in response to 8 different groups of questions.

In this paper, we are interested to analyse the changes in the affective state of a person in response to these questions. The subjects comprised of 30 controls (with no history of mental illness) and 30 (severely), with an equal number of males and females. The age range is 21-75yr. The duration of the interviews ranges from 208-1672s. In an ideal situation, one would wish for a larger dataset. As this project is a part of an ongoing study, more data is being collected. Similar limitations with the sample size have been reported by Ozdas et al. [4] and Moore et al. [5] and are a common problem in this area of research.

## 3 Method

Given an input video $V$ containing $N$ frames $F_j$, $V = \{F_1, F_2, ..., F_N\}$, first, a face detector is applied to each frame. The resultant face blob $F_j$ is further aligned based on facial fiducial points. Spatio-temporal features are then computed on the aligned face blob clip $V_j$. A BoW is learned from these computed features. The videos in our clinical database contain the upper body of the subjects. Not only the facial features inside the face, but also the head, hand and shoulder movements provide important information. Therefore, STIP are computed on the entire video $V$. This is further embedded in a BoW model. Figure 1 describes the flow of the system.

### 3.1 Face Analysis

The Viola-Jones object detector is applied to each frame in a video $V$. The resultant face blob $F$ is used as

\(^1\)http://www.blackdoginstitute.org.au/
a seed for facial feature extraction. Everingham et al.’s approach [14] is used to extract nine facial points, which describe the location of the left and right corners of both eyes, the tip of the nose, the left and right corners of the nostrils, and the left and right corners of the mouth. For aligning the faces, an affine transform based on these points is computed.

### 3.2 LBP-TOP

Intra-face muscle movements are computed using Local Binary Pattern (LBP) [15]. Specifically, LBP-TOP [15] is computed for a video. It considers patterns in three orthogonal planes: $XY$, $XT$ and $YT$, and concatenates the pattern co-occurrences in these three directions. The local binary pattern (LBP-TOP) descriptor assigns binary labels to pixels by thresholding the neighborhood pixels with the central value. Therefore, for a centre pixel $O_p$ of an orthogonal plane $O$ and its neighbouring pixels $N_i$, a decimal value $d$ is assigned

$$
d = \sum_{O} \sum_{i=1}^{k} 2^{i-1} I(O_p, N_i) \quad . \tag{1}
$$

For a video $V$ of length $l$, uniformly timed sub-clips of length $t$ were segmented. These sub-clips were used to compute LBP-TOP. Therefore, there are $l/t$ sub-clips and their corresponding LBP-TOP descriptors $d_{1:t}$.

### 3.3 Space Time Interest Points

Along with the intra-facial muscle dynamics, we are interested in upper body movements. Lately, STIP [16] has seen much attention in video analysis. It successfully detects useful and meaningful interest points in a video. It extends the idea of the Harris spatial interest point detector to local structures in the spatial-temporal domain. The salient points are detected where image values have significant local variations in both the space and time dimensions. Two histograms of gradient and flow are calculated around an interest point in a fixed sized spatial and temporal window. STIP computes the vital spatio-temporal changes which accounts for movements inside the facial area and outside (like hands, shoulders and head movements).

**Key-Interest Point Selection:** A video $V$ gives $K$ interest points. A total of $48 \times 10^6$ interest points are computed from the 60 video clips. This is both computationally and memory wise non trivial. In order to reduce the feature set size, K-Means algorithm is computed for each $V$. $K$ interest points give $K_c$ cluster centers. These $K$ key-interest points are the representative interest points of a video sample and this is similar to keyframe selection for emotion analysis [13]. The value $K_c$ is chosen empirically.

### 3.4 Codebook creation

Bag of Words, originally from the natural language processing domain, has been successfully applied to image analysis problems [16]. It represents documents based on the unordered word frequency. In the problem described here, a video is a document in the BoW sense. Two codebooks are computed in the work described here. Codebook $C_1$ is computed by clustering the cluster centres $K_c$ for each video (which are computed by clustering interest points of each video as described in Fig. 1). Codebook $C_2$ is computed by clustering the LBP-TOP descriptors for all videos. A hard word assignment is performed for computing two histograms from the two codebooks, respectively. The two histograms for a video are joined to form the final feature vectors. These are then used to learn the classifier defined in the next section.

### 3.5 Classifiers

Neural networks have been successfully employed to complex pattern classification problems in recent years. Neural networks are non-linear, data driven self-adaptive models. The underlying principle of a neural network is that it approximates to fit to any arbitrary function. Given the scarcity of data and the complexity of the problem discussed in this paper, neural network based classification is a viable direction to explore. Two basic neural network algorithms – Feed Forward Neural Network (FFNN) [17] and Probabilistic Neural Network (PNN) [18] – have been experimented with.

FFNN is one of the simplest forms of neural networks consisting of series of layers. In this paper, a multi-layer network is employed where the information travels from the input layer through hidden layers to output layers. A supervised learning algorithm, scaled conjugate gradient [19], is used for training the network. PNN is a special form of FFNN, which combines FFNN and Bayesian networks. Other than input and output nodes, PNN has two more nodes – pattern node and summation node. The pattern node consists of gaussian functions formed considering given set of data points as centres. The summation node simply sums all the inputs from the pattern layer corresponding to the category from which the pattern is selected.

Further, we also explore Restricted Boltzmann Machine (RBM) [20] for its advantages over the basic neural network algorithms. In traditional neural networks, the weighted connections between the multiple layers
are evaluated by calculating the back-propagation error. RBM are stochastic neural networks, in which binary visible units are connected to binary hidden units using symmetrically weighted connection.

4 Experimentation and Results

The original video frame resolution is $800 \times 600$ pixels. For face detection, both frontal and profile face models available in openCV\(^2\) are computed in a cascade. The Region of Interest (ROI) generated by the face detector is used as seed for a pictorial structure and the face blob size is set to $70 \times 70$ pixels. For STIP, harris 3D interest point detector is used. The spatial window size for computing HOG is set as 3 and the temporal window size for HOF as 9. The videos are down sampled to $320 \times 240$ pixels. Two different values of cluster centers $K_c = 2500$ and $K_c = 5000$ are experimented for clustering $K$ interest points for each video $V$. For computing LBP-TOP, the sub-clip sizes are experimented for $t = 3s$ and $t = 6s$. For both visual codebooks $C_1$ and $C_2$, results for three different cluster size $C_s = [200, 500, 750]$ are evaluated. Figure 2, describes the performance of the four classifiers based on varying cluster sizes of the codebook.

For SVM, a Radial Basis Function is used and the cost and gamma parameters are selected empirically using a grid search. For RBM, the number of hidden variables is chosen empirically from a range of $2 - 100$. For PNN, various spread values are experimented ranging from $1 - 100$ and for FFNN, the range of hidden layers experimented ranges from $10 - 500$. Table 1 describes the performance comparison for these classifiers and the two descriptors and their fused variants. Here, $STIP_1$ means STIP with level one clusters size $K_c = 2500$ and $STIP_2$ is STIP with level one cluster size $K_c = 5000$. For LBP-TOP, $LBP_1$ is the configuration with clip length $t = 6s$ and $LBP_2$ is the configuration with clip length $t = 3s$. Further, four possible descriptor combinations: $STIP_1 + LBP_1$, $STIP_1 + LBP_2$, $STIP_2 + LBP_1$ and $STIP_2 + LBP_2$ are also experimented.

In our experiments, RBM performs the best classification for the features $STIP_1$ with $C_s = 750$ and hidden variables as 16. The maximum accuracy attained by $STIP_1$, $C_s = 750$ is 88.3\%, closely followed by the combination of $STIP_2$ and $LBP_2$: 86.7\%. These results confirm the argument that upper body gestures along with intra-facial dynamics provide robust cues for depression classification. It is evident that adding LBP-TOP to STIP does not always give an increase in performance. In fact in some configurations for RBM, it is decreased. We attribute this to two factors: LBP-TOP in its default formulation (for facial expression analysis) assumes the $XY$ frame close to apex, which is not always true in our case. Multiple $XY$ frames in each sub-clip can be helpful. Secondly, STIP also considers significant movements in the face, therefore fusing more facial dynamics information may be becoming redundant and given the small size of the database and higher dimension of the feature, it can lead to the ‘curse of dimensionality’.

5 Conclusion

Depression is a major disabling disorder. The paper proposes a subject independent framework for visual cues based depression recognition via spatio-temporal descriptors. LBP-TOP descriptors are computed on the

---

\(^2\)http://opencv.willowgarage.com/
aligned faces in the videos to capture intra-facial dynamics. In order to extract spatio-temporal facial dynamics, the computed descriptors are used to learn a codebook of spatio-temporal words. STIP are computed on the videos and in order to reduce the memory requirements, key interest points are selected for each video by clustering at the video level. The key interest point selection trick made the problem tractable on a normal computer. Further, a codebook is computed from the video level key interest points. A clinical database collected at the Black Dog Institute is used in the experiments. A leave-one-out experiment protocol is used. The accuracy results prove the effectiveness of the proposed framework. As part of future ongoing work, audio features will be fused with the framework and face pose handling mechanisms will be added to the system.

### References


