Can Body Expressions Contribute to Automatic Depression Analysis?

Jyoti Joshi[1], Roland Goecke[1,2], Gordon Parker[3] and Michael Breakspear[4,3]
[1] HCC Lab, Vision & Sensing Group, ESTeM, University of Canberra, Australia
[2] IHCC Group, RSCS, Australian National University, Australia
[3] University of New South Wales, Australia
[4] Queensland Institute of Medical Research, Australia
jyoti.joshi@canberra.edu.au, roland.goecke@ieee.org, g.parker@blackdog.org.au, mjbreaks@gmail.com

***Abstract*— Depression is one of the most common mental health disorders with strong adverse effects on personal and social functioning. The absence of any objective diagnostic aid for depression leads to a range of subjective biases in initial diagnosis and ongoing monitoring. Psychologists use various visual cues in their assessment to quantify depression such as facial expressions, eye contact and head movements. This paper studies the contribution of (upper) body expressions and gestures for automatic depression analysis. A framework based on space-time interest points and bag of words is proposed for the analysis of upper body and facial movements. Salient interest points are selected using clustering. The major contribution of this paper lies in the creation of a bag of body expressions and a bag of facial dynamics for assessing the contribution of different body parts for depression analysis. Head movement analysis is performed by selecting rigid facial fiducial points and a new histogram of head movements is proposed. The experiments are performed on real-world clinical data where video clips of patients and healthy controls are recorded during interactive interview sessions. The results show the effectiveness of the proposed system to evaluate the contribution of various body parts in depression analysis.**

I. INTRODUCTION

Affective sensing technology has come a long way since its start in the 1990s. With the advancement in computer vision technology, it is possible to build systems, which can predict neurological problems such as depression, pain and stress. This paper focusses on the analysis of major depressive disorders based on an automatic visual analysis of intra-facial movements and upper body gestures. We assess the contribution of different body parts and motions in three areas: face-only, entire upper body (including head) and head movements only.

Affect plays a vital role in our lives and is an integral part of human perception and communication. Emotions are not only responsible for cognitive functions such as rational decision making, perception and learning, but are also important for interpersonal communication [1]. Recent advances in affective sensing and related areas, e.g. automatic face tracking in videos, measuring facial activity, recognition of facial expressions, analysis of affective speech characteristics and physiological effects that occur as a result of affective state changes paired with the decreasing cost and increasing power of computing have led to an arsenal of prototypical affective sensing tools now being at our finger tips. We can employ these to tackle higher problems, e.g. supporting clinicians in the diagnosis and monitoring of mental health disorders, such as depression.

Depression is a common and disabling mental disorder, which has strong adverse effects on personal and social functioning as well as a significant economical impact. It has been quantified as the leading cause of disability worldwide by the landmark WHO 2004 Global Burden of Disease report by Mathers *et al.* [2]. The lifetime risk for depression has been reported to be at least 15% [3]. People of all ages suffer from depression, which is also a major cause for suicide. One prime reason for such a high suicidal rate is the absence of any objective measurement technique for the diagnosis of depression. All current assessment methods rely almost exclusively on patient-reported or clinical judgements of symptom severity, risking a range of subjective biases. This can be widely addressed, if depressed people consult physicians who have been provided with objective means to diagnose depression at early stages. Affective sensing technology can provide these objective means and assist physicians in both initial depression diagnosis and ongoing monitoring [4].

Disturbances in the expression of affect reflect changes in mood and interpersonal style, and are arguably a key index of a current depressive episode. This leads directly to impaired interpersonal functioning, causing a range of interpersonal disabilities, functioning in the workforce, absenteeism and difficulties with a range of everyday tasks (e.g. shopping). Despite its severity and high prevalence, there currently exist no laboratory-based measures of illness expression, course and recovery. This compromises optimal patient care, compounding the burden of disability. As healthcare costs increase worldwide, the provision of effective health monitoring systems and diagnostic aids is highly important. With the advancement in affective sensing and machine learning, computer aided diagnosis can and will play a major role in providing an objective assessment.

II. RELATED WORK

Emotion analysis from facial expression recognition is a well-researched problem [5]. Over the past two decades, various geometric, shape, texture, static and temporal visual descriptors have been proposed for various expression analysis related problems (e.g. [6], [7], [8]). With advancement of affective sensing technologies, computer-based automatic

depression analysis has also become an active area of research. In one of the earliest works on automatic depression analysis, Facial Action Coding System (FACS)-based facial features and vocal features were explored by Cohn *et al.* [9]. In their work, person-specific Active Appearance Models (AAM) [10], [11] were used to automatically track facial features. Then, shape features were used to compute various parameters such as the occurrence of Action Units (AU) associated with depression, their mean duration, ratio of onset to total duration and ratio of offset to onset phase. Vocal features such as pitch were also explored and a comparison was made between facial and vocal analysis for depression detection, but no multimodal fusion of the two.

Following a proposition introduced by Ellgring [12] that there is a significant decrease in facial activity in depression while it increases with the improvement of subjective well-being, McIntyre *et al.* [13] analysed the facial movements of the subjects when shown short video clips from movies, which had been shown to elicit various emotions in subjects [14]. Similar to Cohn *et al.* [9], McIntyre *et al.* also trained person-specific AAMs and shape features were computed from every fifth video frame. The shape features were combined and classified at the frame level into either depressed or non-depressed via a Support Vector Machine (SVM). However, investigating the temporal cues of facial responses could be more beneficial for depression detection as facial activity is dynamic in nature. It has also been shown in the literature that temporal facial dynamics provide more information for facial expression analysis than using static information only [15].

Both [9] and [13] used person-specific AAM models. For a new subject, a new AAM model needs to be trained, which is both complex and time consuming. In contrast, the video analysis in our proposed framework is subject-independent. It has been shown in the literature that for dynamic facial expression analysis, temporal texture features perform better than geometric features [7]. Simple temporal features such as the mean duration of AUs [9] have been used. In our work, the sophisticated spatio-temporal descriptors known as Space-Time Interest Points (STIP) are applied, which have been used for incorporating temporal information.

Joshi *et al.* [16] proposed a spatio-temporal descriptor based approach for depression analysis. Facial dynamics were analysed by computing local binary pattern (LBP) descriptors and the upper body movements were analysed by computing STIP. Both the descriptors were further embedded in a bag of words framework. The experiments in [16] showed that upper body analysis, which includes the face/head and the shoulders, performs well. However, the contribution and effect of upper body parts versus face versus head movements was not analysed. In this paper, we explore any differences in the ability of different body parts to discriminate between depressed and non-depressed subjects.

Recently, emotion recognition from body movements and gesture analysis has attracted much attention from researchers in affective computing [17], [18], [19]. A detailed survey of various methods used for body expression recognition and analysis is presented in [20]. As reported in some of these works, body expressions and gestures are as significant a visual cue as facial expressions. Thus, it is of interest to explore body movements and gestures for automatic depression analysis. So far to the best of our knowledge, no attempt has been made to investigate head movements and overall upper body movements as a means to quantify depression. In this paper, we explore the contribution of upper body movements and head movements in depression analysis and compare it with information derived from facial dynamics.

In an interesting study, Boker *et al.* [21] looked at the effect of change in intensity of facial expressions and head movements during dyadic conversations. In their experimental setup, participants were video-conferencing with resynthesized avatars of confederates without any prior knowledge of the manipulation on the apparent live video. AAMs were used to attenuate the conversator's facial expression and head movements. This work gave some interesting results on adaptive facial behaviour in natural conversation. One of the finding was that when attenuation was employed, both conversants adapted by increasing the vigour of their head movements to elicit the expected response. It was also suggested in the Boker *et al.* paper that there is a close relationship between head movements (nods and turns) and facial expressions in dyadic conversations. Since the data used in our experiments is a recorded interview session between the subject and the interviewer, analysing the head movements and investigating their relationship with the facial expressions and upper body gestures will be an interesting contribution. As a first step, in this work, we analysed the overall contribution of (any) head movements in detecting depression in comparison to the contributions from upper body gestures and facial dynamics, respectively.

To summarise, the key contributions of the paper are:

1) We investigate the problem of depression analysis using vision-based methods. As previous work has been limited to faces only, we include upper body and head movements along with intra-facial dynamics for assessing the presence of depression in a subject.
2) We propose a novel histogram of head movements, which is computed over the rigid facial points in a face for every fifth frame for an entire video sequence. We show that head movements can be a strong cue to infer if a subject is depressed.
3) We evaluate the performance of upper body and facial interest points in a Bag of Words (BoW) framework. The experiments section below shows that the upper body analysis gives the best estimation of depression in real-world clinical depression database.
4) As the interview videos are relatively long, a large number of interest points is generated. To overcome this, key interest point selection is performed for both facial and upper body samples. BoW are computed on top of the key interest points. This reduced the computational complexity and made analysis well feasible within a clinically acceptable time frame.

Fig. 1. STIP visualisation. The top two rows show the STIP generated on the upper body visible in the video frames in our database. The bottom two rows show the STIP generated on the aligned facial frames for the same video. The yellow circles indicate the presence and intensity of interest points. It is evident from the comparison of the frames inside the two rectangles that upper body expressions tend to generate many interest points, which provide useful discriminative information.

The rest of the paper is organised as follows: the clinical database is discussed in Section III. We present the proposed upper body analysis method in Section IV. This section also introduces the proposed face processing pipeline and the novel histogram of head movements. Experiments and results are discussed in Section V. Finally, we conclude with a summary of the findings and contributions of this work in Section VI.

III. DATA

The clinical database used in this study was collected at the Black Dog Institute, Sydney, Australia, a clinical research institute focussing on mood disorders, including depression and bipolar disorder.[1] 60 subjects (30 males and 30 females) with an age range of 19-72yr were interviewed. Subjects included 30 healthy controls (mean age 33.9 ± 13.6yr) as well as 30 patients (mean age 44.3 ± 12.4yr) who had been diagnosed with severe depression (but no other mental disorders or co-morbid conditions).

Participants in the Black Dog research program first complete a computerised mood assessment program (MAP), which generates diagnostic decisions and a profile of personality, co-morbid conditions such as anxiety disorders, current functioning assessments, as well as current and past treatments, and a section on the aetiology of their depressive episode (e.g. family history; stressful life events). Following the MAP, the participants undergo a structured interview (MINI) that assesses current and past depression as well as hypo(manic) episodes and psychosis (both current and past) as per the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). If they are currently depressed and are deemed eligible for the ongoing study (unipolar depression and no history of psychosis), they will also be rated on the CORE measure of psycho-motor disturbance [22]. In the present study, only severely depressed patients (HAMD $>$ 15) were included. The recordings were made after their initial diagnosis and before the start of any treatment. Control subjects were carefully selected to have no history of mental illness and to broadly match the depressed subjects in age and gender.

The audio-video experimental paradigm contains several parts, including a read sentences task and an interview with the subjects, similar to [13]. In this study, we are interested in analysing the changes in facial muscles, movement in head and shoulders while responding to the interview questions. The length of the video clips varies from $183 - 1200$s. In an ideal situation, one would wish to have a larger dataset. However, this project is part of an ongoing study and more data is being collected. Similar limitations with the sample size have been reported by Ozdas *et al.* [23] and Moore *et al.* [24].

IV. METHOD

Let us assume we have an input video $\mathcal{V}$ containing N frames $\mathcal{F}_i$, $\mathcal{V} = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_N\}$. First, a face detector is applied to each frame. The resultant face blob $\mathcal{F}$ is further aligned based on facial fiducial points. Spatio-temporal features are then computed on the aligned face blob video clip $\mathcal{V}_f$ (Section IV-A). Further, to analyse the upper body movements along with facial dynamics, STIP are computed on the entire video $\mathcal{V}$ (Section IV-B). For both intra-facial and upper body movement analysis, key interest point selection is performed (Section IV-C) and a BoW is learned from selected key points (Section IV-D). In the case of face analysis, we term this a Bag of Facial (BoF) dynamics. In the case of upper body analysis, we call it a Bag of Body (BoB)

[1] http://www.blackdoginstitute.org.au/

expressions. Finally, head movement analysis is performed using three rigid fiducial points and a histogram to represent the frequency of head movements is proposed (Section IV-E).

A. Face Analysis

The Viola-Jones object detector [25] is applied to each frame in a video $\mathcal{V}$. The resultant face blob $\mathcal{F}$ is used as a seed for facial feature extraction. Recently, parts-based models such as [26] have found success in facial landmark localisation. Their power stems from the representation of facial parts as nodes of a tree-graph and an individual template-based part detector. A distance transform is applied in a dynamic programming method for finding the facial parts' location in the scores of the parts-based detector templates. Learning individual part detectors and shape model leads to better generalisation for subjects.

Ideally, for a facial dynamics analysis system, one will want a subject-independent facial landmark detector. Chew *et al.* [27] also argue that subject-dependent facial parts methods such as AAMs [10] perform better than subject-dependent constrained local models [28]. However, if a proper descriptor is used on top of aligned faces from a subject-independent detector, the error in alignment can be compensated for. Moreover, Zhu and Ramanan [29] show the effectiveness of parts-based models over CLM and AAM for facial landmark localisation when there is a lot of head movement. Motivated by these arguments, the parts-based model of Everingham *et al.* [30][2] is used in our work to extract nine facial points, which describe the location of the left and right corners of both eyes, the centre point 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. To capture intra-facial movements only, STIP features are computed over the aligned face videos.

B. Upper Body Analysis

Along with the intra-facial muscle dynamics, we are interested in analysing the possible contribution of upper body movements to the classification of subjects into depressed or non-depressed. Lately, the STIP features [31] have received much attention in video analysis. It successfully detects useful and meaningful interest points in a video by extending the idea of the Harris spatial interest point detector to local structures in the spatial-temporal domain. Salient points are detected where image values have significant local variation in both the space and time dimensions. Two histograms of gradient (HoG) and flow (HoF) are calculated around an interest point in a fixed sized spatial and temporal window. STIP computes important spatio-temporal changes, which account for movements inside the facial area as well as outside (such as on the hands, shoulders and head movements overall).

[2]Zhu and Ramanan [29] report that their method works better than the approach of Everingham *et al.* [30]. However, because of the computational complexity and a comparable performance on our dataset, we employed the method of [30].

C. Key Interest Point Selection

An upper body video $\mathcal{V}$ gives K interest points (the STIP features). A total of $K = 4.8 \times 10^7$ interest points are computed from the 60 upper body video clips in our dataset. A video $\mathcal{V}_f$ of the face region only gives $K_f = 1.5 \times 10^7$ interest points. While this is a significant drop in the total number of detected interest points, the large number of interest points in both $\mathcal{V}$ and $\mathcal{V}_f$ is both computationally and memory wise non-trivial. In order to reduce the feature set size, the K-Means algorithm is applied to the K and K_f interest points of each $\mathcal{V}$ and $\mathcal{V}_f$ video, respectively, resulting in k and k_f cluster centres, respectively. These k and k_f cluster centres are the representative key interest points of a video, which is similar to key frame selection for emotion analysis [32]. The value for k and k_f is chosen empirically.

D. Bag of Words

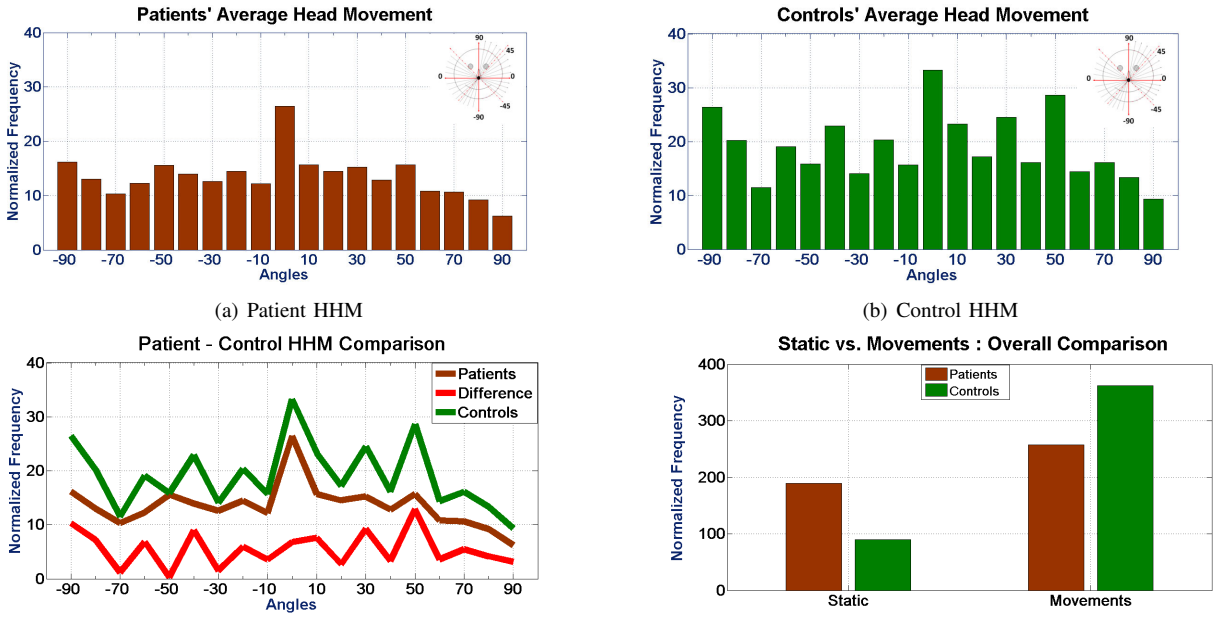
Bag of Words, originally from the natural language processing domain, has been successfully applied to image analysis problems [31]. It represents documents based on the unordered word frequency. In the problem described here, a video is a document in the BoW sense. Codebooks are computed in the work described here. First, a Bag of Body expressions is learnt by computing a codebook C by clustering the cluster centres k for each video $\mathcal{V}$. Then, a second codebook C_f is computed to create a Bag of Facial dynamics using cluster centres k_f for each video $\mathcal{V}_f$. The size of C and C_f is also chosen empirically.

E. Head Movement Analysis

As part of this study, we analysed the frequency of occurrence of overall head movements. We propose a Histogram of Head Movements (HHM), computed as follows: three rigid fiducial points representing the two outer corners of the eye l_l, l_r and the tip of the nose l_n (see Section IV-A) are selected. A histogram is computed for an input video V by first computing the slope of a line formed by the two corresponding points (l^t, l^{t+1}) in a pair of frames $(t, t+1)$ with the horizontal axis. For the experiments, the temporal distance between t and $t+1$ was 5 frames for computational efficiency. Other values are possible. Then, angle θ is deduced from the slope of this line. Formally, the HHM can be defined as:

$$HHM = \sum_{N} avg(\theta(l_l^t, l_l^{t+1}), \theta(l_r^t, l_r^{t+1}), \theta(l_n^t, l_n^{t+1}))) \quad (1)$$

Here, $\{l_l^t, l_r^t, l_n^t\}$ are the locations of three facial parts in the current frame and $\{l_l^{t+1}, l_r^{t+1}, l_n^{t+1}\}$ are the locations in the next fifth frame, resulting in three values for θ: $\theta(l_l^t, l_l^{t+1}), \theta(l_r^t, l_r^{t+1}), \theta(l_n^t, l_n^{t+1})$. An average over these is calculated every for each pair of video frames to normalise the error in the facial landmark detection, if any. HHM is then the histogram created with bins of width 10° in the range of $-90^\circ + 90^\circ$. The video clips in our dataset are of different length; therefore, individual, subject-specific HHMs were computed on all sub-sequences/sub-clips of length 60s (without overlap) of a subject's interview video. Finally, the



(a) Patient HHM

(b) Control HHM

(c) Patient-Control: HHM Comparison

(d) Patient Control: Overall Head Movement Comaprison

Fig. 2. Graphs (a) and (b) describe the Histogram of Head Movements (HHM) of patients and controls, respectively, normalised over time. The HHMs are computed over time intervals of 60s and averaging is applied. Graph (c) describes the comparison of HHM of patients and controls across the database. Graph (d) compares the frequency of occurrence of head movements versus a static head position. A t-test confirmed that the differences are statistically significant at the $p = 0.0025$ level.

HHM for a subject was generated by averaging all the subsequence level HHMs. Figure 2 (a) is the average HHM of all the patients in the dataset and Figure 2 (b) represents the average HHM of all the controls in the dataset. As hypothesised, the depressed subjects have fewer head movements. This is evident from Figure 2 (c), which illustrates the comparison between the HHM of patients and controls. The red line in the plot shows the amount of difference in the frequency of occurrence of overall head movements between patients and controls across the whole dataset. Figure 2 (d) describes the presence or absence of any head movement in the dataset for patients and controls. Observing the results, it can be clearly stated that patients show fewer movements over time with a larger number of static head positions than controls, who exhibit more head movements.

V. EXPERIMENTS AND RESULTS

There are a total of 60 subjects in the dataset. A leave-one-subject-out protocol was followed. Let us discuss the implementation details of the methods. The original spatial resolution of the video frames of $\mathcal{V}$ was 800×600 pixels. The videos were downsampled to 320×240 pixels for computational efficiency. The STIP features were first computed on the entire interview video sequence for each subject. The Harris 3D corner detector wass used as the interest point detector. For computing the HoG, the spatial window size was set to 3 and the temporal window size for HoF to 9. The total number of interest points generated from STIP computation on all the controls' and patients' videos was 4.8×10^7. Any type of further processing on all the detected interest points together is computationally and time-wise expensive. In order to overcome this issue, the K-Means algorithm was applied to the STIP features of each $\mathcal{V}$ with four different values of cluster centres k *viz.* 1000, 1500, 2000 and 2500. Then, a BoB is computed over selected key interest points. Various codebook sizes C in the range [200-750] were experimented on.

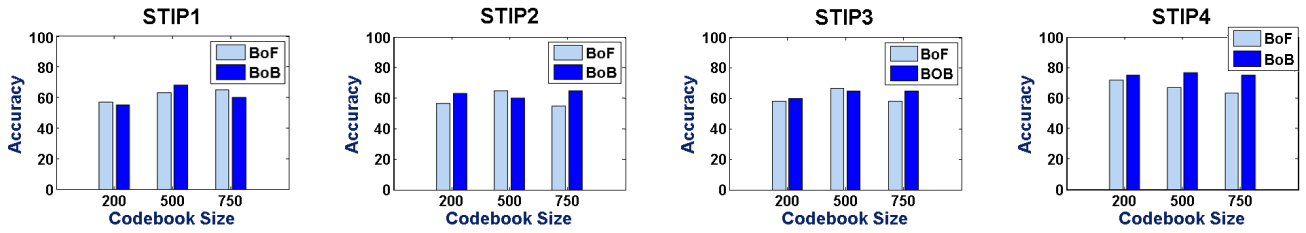
Configurations		STIP1	STIP2	STIP3	STIP4
Bag of Facial Dynamics	Acc. (%)	65.0	65.0	66.7	71.7
	F1-Score	0.64	0.73	0.69	0.73
Bag of Body Expressions	Acc. (%)	68.3	65.0	65.0	**76.7**
	F1-Score	0.71	0.68	0.7	0.8

TABLE I

BEST CLASSIFICATION ACCURACIES AND F-SCORE MEASURES FOR DIFFERENT CONFIGURATIONS OF STIP FOR THE BAG OF BODY EXPRESSIONS (BOB) AND BAG OF FACIAL DYNAMICS (BOF).

The spatial resolution of the face only video frames of $\mathcal{V}_f$ was 90×80 pixels. Similar steps to $\mathcal{V}$ were followed to compute STIP on $\mathcal{V}_f$, using a similar configuration with Harris 3D as interest point detector, a spatial window size of 3 for HoG and a temporal window size for HoF of 9. Again, due to the large number of interest points, which in this case was 1.5×10^7, K-Means clustering was performed with cluster centre values $k_c = 1000$, $k_c = 1500$, $k_c = 2000$ and $k_c = 2500$. Furthermore, a BoF was computed for sizes C_f = [200-750]. Figure 3 shows the effect of changing codebook size for various STIP configurations in BoB and BoF.

From here on, STIP1 means the STIP features computed on $\mathcal{V}$ and $\mathcal{V}_f$ with cluster centre $k = k_f = 1000$, respectively. STIP2 has cluster centres $k = k_f = 1500$, STIP3 has cluster



(a) STIP1 (b) STIP2 (c) STIP3 (d) STIP4

Fig. 3. The four graphs describe the accuracy comparison between depression detection by BoB (Bag of Body expression) and BoF (Bag of Facial dynamics) for different configurations of STIP. Here, STIP1 means STIP with level one cluster size $k = k_f = 1000$, STIP2: $k = k_f = 1500$, STIP3: k = k_f = 2000 and STIP4: k = k_f = 2500

centres $k = k_f = 2000$ and STIP4 has cluster centres $k = k_f = 2500$.

A non-linear SVM was used for classification. A radial basis function was used and the parameters, cost and gamma, were selected using an extensive grid search. Table I reports the best accuracy and F1 scores for depression detection using different configuration of STIPs on BoB and BoF. It is evident from the results shown in the table that the best performance was by BoB for STIP4 with codebook size $C = 500$, which gave F1=0.8 and 76.7% accurate classification. For the same configuration of STIP, the best accuracy given by BoF was 71.7% for codebook size $C_f = 200$ and F1=0.73. These results clearly show the significance of using body expressions alongside facial movements than using face information only.

Other STIP configurations also show similar behaviour. For STIP1, the performance of BoB was 68.3% and F1=0.71 for codebook size $C = 500$. The performance of BoF was 65.0% accurate results and F1=0.64 for the same value of codebook size. STIP2 had the same accuracy results for both BoB and BoF as 65% for $C = 750$ and $C_f = 500$, respectively. In STIP3, BoF performs slightly better with accuracy results as 66.7% with F1=0.69 for $C_f = 500$, closely followed by BoB with an accuracy of 65% and F1 measure as 0.7 for C equals to C_f. The slight decrease in performance of BoB for this particular configuration can be attributed to the key-interest points selection step. The K-means algorithms is sensitive to initialisation and at times has problems, when stuck in a local minima. [31] ran K-means 8 times to get the maximum results and assumes that a globally optimum solution will be reached. Another solution is to average out the results. We will address this issue as part of future work.

As described in Section IV-A, a face blob $\mathcal{F}$ was extracted from $\mathcal{V}$. The detected face blob $\mathcal{F}$ was used as seed to extract nine facial points (Section IV-A). For the head movement analysis, the HHMs for all the patients and controls were computed (see Section IV-E). A t-test was performed to statistical validate the difference between the average patients' HHM and average controls' HHM shown in Figure 2 (a) and (b). The difference between the two is statistically significant with $p = 0.0025$ for α = 0.05. An SVM classifier was used to classify between depressed patients and healthy controls. The classification based on overall head movements resulted in an accuracy of 71.7%. This is a very significant outcome of the experiment as the best result for the Bag of Facial dynamics (Table I) was also 71.7%. This shows that analysing overall head movements alone can also significantly help in depression recognition using affective sensing technology. Figure 2 represents the frequency of occurrence of head movements in patients and controls.

One limitation of the current framework is that the face analysis was done on a holistic level. As discussed in [33], the spatial structure is important for face analysis. As part of the future work, a Bag of Words framework based on spatial pyramids will be explored. Further, salient region detection will be performed such that features can be made more discriminative. Head movement analysis performed in this work was subjected to scale variation, which will be addressed by including a 3D component in the analysis. The dataset analysed in this study had 60 subjects, but more subjects have been added recently to the dataset at the Black Dog Institute. Future experiments will be conducted on the extended dataset. Moreover, recently, methods such as domain transfer and transfer learning [34] have been widely used and have shown promising outcome in overcoming the problems of a lack of labelled training data and the intra-class variation. It will be interesting to see the application of these methods to our problem.

VI. CONCLUSIONS

Depression is a common and disabling mental health disorder. The absence of objective measurement technique makes the diagnosis and treatment difficult. Our research is concerned with developing affective sensing techniques, which supports psychologists in the initial diagnosis and the ongoing monitoring during treatment. In this paper, we studied the contribution of different parts to the recognition of depression. We explored upper body expressions and gestures, head movements and facial dynamics to classify between depressed and controls. STIP features were computed to detect upper body movements and intra-facial movements. Bags of Words were created for body expressions and facial movements separately. Moreover, head movement analysis was performed by assessing the displacement of rigid facial

points and a histogram of head movements was generated. The results of our experiments evidently show that body expressions, gestures and head movements can be as significant a visual cue as facial expressions alone for depression detection.

VII. ACKNOWLEDGMENTS

The research reported in this paper was in part funded by the Australian Research Council Discovery Project grants DP110103767 and DP130101094.

REFERENCES

[1] R. Picard, *Affective Computing*. Cambridge (MA), USA: MIT Press, 1997.

[2] C. Mathers, T. Boerma, and D. M. Fat, "The global burden of disease: 2004 update," WHO Press, Switzerland, Tech. Rep., 2004.

[3] R. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K. Merikangas, A. Rush, E. Walters, and P. Wang, "The Epidemiology of Major Depressive Disorder: Results From the National Comorbidity Survey Replication (NCS-R)," *The Journal of the American Medical Association*, vol. 289, no. 23, pp. 3095–3105, Jun. 2003.

[4] M. Prendergast, *Understanding Depression*. Australia: Penguin, Mar. 2006.

[5] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *CVPR (2)*, 2005, pp. 568–573.

[6] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, "Evaluating AAM Fitting Methods for Facial Expression Recognition," in *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*, ser. ACII'09, 2009, pp. 598–605.

[7] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2007.

[8] A. Dhall, A. Asthana, and R. Goecke, "Facial expression based automatic album creation," in *ICONIP*, 2010.

[9] J. F. Cohn, T. S. Kreuz, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting Depression from Facial Actions and Vocal Prosody," in *Proc. Affective Computing and Intelligent Interaction (ACII2009)*, 2009, pp. 1–7.

[10] G. Edwards, C. Taylor, and T. Cootes, "Interpreting Face Images Using Active Appearance Models," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition FG'98*. Nara, Japan: IEEE, Apr. 1998, pp. 300–305.

[11] J. Saragih and R. Goecke, "Learning AAM fitting through simulation," *Pattern Recognition*, 2009.

[12] H. Ellgring, *Nonverbal communication in depression*. Cambridge University Press, 2008.

[13] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear, "An Approach for Automatically Measuring Facial Activity in Depressed Subjects," in *Proc. ACII2009*, Sep. 2009, pp. 223–230.

[14] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," in *Cognition and Emotion*, 1995, pp. 87–108.

[15] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions," *Psychological Science*, pp. 403–410, 2005.

[16] J. Joshi, A. Dhall, R. Goecke, M. Breakspear, and G. Parker, "Neural-net classification for spatio-temporal descriptor based depression analysis," in *Proceedings of the International Conference on Pattern Recognition*, ser. ICPR'12, 2012.

[17] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1334–1345, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.jnca.2006.09.007

[18] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, ser. ACII '07, 2007, pp. 71–82.

[19] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, june 2008, pp. 1 –6.

[20] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *Affective Computing, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2012.

[21] S. M. Boker, J. F. Cohn, B. . Theobald, I. Matthews, T. R. Brick, and J. R. Spies, "Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3485–3495, 2009.

[22] G. Parker and D. Hadzi-Pavlovic, *Melancholia: A Disorder of Movement and Mood*. New York, USA: Cambridge University Press, 1996.

[23] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Analysis of fundamental frequency for near term suicidal risk assessment," in *Proc. IEEE International Conference on Systems Man and Cybernetics*, 2000.

[24] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," in *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, 2008, pp. 96–107.

[25] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[26] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," *IJCV*, 2005.

[27] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan, "In the pursuit of effective affective computing: The relationship between features and registration," *IEEE TSMC B*, pp. 1006–1016, 2012.

[28] J. M. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *ICCV*, 2009.

[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[30] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video," in *Proceedings of the British Machine Vision Conference*, 2006, pp. 899–908.

[31] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR'08, 2008, pp. 1–8.

[32] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *IEEE AFGR2011 workshop FERA*, 2011.

[33] I. Biderman and P. Kalocsais, "Neurocomputational bases of object and face recognition." in *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 1997, pp. 1203–1219.

[34] M. Liu, S. Li, S. Shan, and X. Chen, "Enhancing expression recognition in the wild with unlabeled reference data," 2012.