Relative Body Parts Movement for Automatic Depression Analysis

Jyoti Joshi[†] Abhinav Dhall[‡] Roland Goecke[†‡] Jeffrey F. Cohn[*]
[†]University of Canberra [‡]Australian National University [*]University of Pittsburgh
jyoti.joshi@canberra.edu.au, abhinav.dhall@anu.edu.au, roland.goecke@ieee.org and jeffcohn@pitt.edu

***Abstract*—In this paper, a human body part motion analysis based approach is proposed for depression analysis. Depression is a serious psychological disorder. The absence of an (automated) objective diagnostic aid for depression leads to a range of subjective biases in initial diagnosis and ongoing monitoring. Researchers in the affective computing community have approached the depression detection problem using facial dynamics and vocal prosody. Recent works in affective computing have shown the significance of body pose and motion in analysing the psychological state of a person. Inspired by these works, we explore a body parts motion based approach. Relative orientation and radius are computed for the body parts detected using the pictorial structures framework. A histogram of relative parts motion is computed. To analyse the motion on a holistic level, space-time interest points are computed and a bag of words framework is learnt. The two histograms are fused and a support vector machine classifier is trained. The experiments conducted on a clinical database, prove the effectiveness of the proposed method.**

I. Introduction

Advances in affective computing have enabled researchers to model complex human behaviour. It is possible to build systems, which can predict neurological problems such as depression, pain and stress etc. Recently, affective computing techniques have been applied to depression detection [1], [2]. This paper focusses on depression detection based on the relative motion in body parts and holistic body movements. Depression is one of the most common and disabling mental disorders, which has strong adverse effects on personal and social functioning. It is quantified as the leading cause of disability worldwide by the landmark WHO 2004 Global Burden of Disease report by Mathers *et al.* [3]. This disorder can develop at any age. The lifetime risk for depression is reported to be at least 15% [4] and it is also a major cause for suicide. One prime reason for such a high suicidal rate is the absence of an objective measurement technique for 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 addressed, if depressed people consult physicians and physicians are provided with some means to diagnose depression in the early stages. Affective sensing technology can provide those objective means and assist physicians in the initial depression diagnosis as well as subsequent monitoring [5].

II. Related Work

Automatic depression analysis has recently gained attention in the affective computing research community. In an earlier work, Cohn *et al.* [6] used person-specific Active Appearance Models (AAM) [7], [8] 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 there was no attempt on combining different channel information.

Ellgring *et al.* [9] proposed that there is a significant decrease in facial activity in depression while it increases with the improvement of subjective well-being. Based on this hypothesis, McIntyre *et al.* [10] analysed the facial movements of the subjects when shown short video clips from movies, which had been rated to elicit various emotions in healthy control subjects [11]. The approach in [10] is based on extracting geometric features in every fifth frame via person-dependent AAM tracking.

The visual cues analysed in both [6] and [10] were derived from facial information only. Recently, emotion recognition from body movements and gesture analysis has attracted much attention from researchers in affective computing [12], [13], [14]. A detailed survey of various methods used for body expression recognition and analysis is presented in [15]. As reported in some of these works, body expressions and gestures are as significant a visual cue as facial expressions. During a depressive episode, patients suffer from psychomotor retardation [16]. This phenomenon is not only limited to facial activity. The amount and intensity of movements exhibited by patients are also another key observations for behaviour understanding. Thus, it is of interest to explore body movements and gestures for automatic depression analysis.

As our method is based on a body parts detector and Space-Time Interest Points (STIP) [17], it is not person-dependent as opposed to [6], [10]. This is important as for [6], [10], every time a new subject is added to the database, manual annotations of the fiducial landmark points are required to train a new person-specific AAM.

In our recent work, [18] a bag-of-words based approach was proposed for depression detection. The facial movements were analysed using a spatio-temporal descriptor. STIP was used for analysing the head and shoulder movement. The clinical database for this study contained 30 depressed subjects and 30 healthy controls (gender-balanced). The healthy controls had no history of depression or any psychological disease. However, in this paper, the clinical database is different in that there are highly depressed patients and mildly depressed patients but no healthy controls. Furthermore, the data corpus in [18] contained video recordings of the (partial) upper body

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.87

and face, which limited the analysis to the upper body only. One of the findings of our previous work [2] was that the head movement and upper body movement provide discriminative information for analysing depression.

Human action recognition is a very active field of research. Various approaches have been proposed mainly either based on interest points [19] or on human pose estimation [20], [21]. Local motion is analysed in [19] using STIPs. These have given good results on complex databases such as the Hollywood database [19]. A limitation of these approaches is the lack of spatial information. [19] tried to address this by using overlapping spatio-temporal windows for computing bags of words, which adds partial spatial information. However, due to the high degree of freedom of human body parts, this may not always hold.

In an interesting work [21], part locations are computed using parts-based detectors and are used to create a polar histogram. They provide good results on the KTH human action recognition database. Part detection based approaches work well when a full or half body is present. Their limiation lies with the problem of double counting, which is induced in part detectors due to self-occlusion [20]. For techniques such as [21], it is important to get robust detection as input. However, the detection is not always accurate for human action recognition databases such as the Hollywood database [19]. Xu *et al.* [22] found that polar histograms alone can be ambiguous for some motion types. To overcome this, they fused the HOG-3D descriptor with the polar histogram.

Inspired from our previous findings in automatic depression analysis [18], [2], [1] and recent work in the field of human action recognition [19], [21], we propose a system, which fuses the holistic body motion based bag-of-words approach with the parts detector based body parts motion patterns to overcome the limitations of [19], [21].

The **key contributions** of the paper are:

1) We investigate the depression problem from the angle of body parts motion analysis. Earlier approaches [6], [10], [18], [2] have looked at either face dynamics or body parts such as head movement. However, the relationships between the parts are not explored in them.
2) We propose a bimodal system, which overcomes the ambiguity in polar histograms and helps maintaining the spatial information by combing holistic body motion information with it.

III. Data

The analysis is performed on two subsets of a clinical dataset collected at the University of Pittsburgh, USA. The diagnosis of the participants was done using a structured clinical interview [23] to identify a Major Depressive Disorder (MDD) [24]. The participants identified with MDD were interviewed and valuated using the Hamilton Rating Scale of Depression (HRSD) [25] to determine the severity of depression over the course of treatment. These interviews were recorded using four hardware synchronized analogue cameras, two of which positioned on each participant's left and right, focussing on face and shoulders. The third one recorded the interviewer's face and shoulders and the fourth one is used to record a full body view of the participant. For the present work, only videos form the fourth camera capturing a full body view were digitized at a frame rate of 29.97 fps. The top-most image in Figure 1 illustrates an example video feed used for analysis.

There are two subsets; Subset I consists of 12 participants (66.7% females and 33.3% males) and contains 2 videos for each participant, one where the participant has high severity of MDD, *i.e.* HRSD$>= 15$ initially and another one where the participant has shown low symptom severity, i.e. HRSD$<= 7$ in consequent interviews. The age ranges from 21 to 60 years with a mean age of 42 years. The interview clip length varies from 177s to 1023s with an average duration of 566s. The other subset, Subset II, consists of 36 participants (63.9% females and 36.1% males) with 18 participants belonging to the class with high symptom severity (HRSD $>= 15$) and 18 different participants fell into the low symptom severity (HRSD $<= 7$) class. The age range was from 19 to 60 years, with an average age of 41 years. The average clip duration was 555s.

There is a uniform distribution of samples belonging to the severely depressed class and the mildly depressed class in both the subsets. The population across the classes in subset II is mutually exclusive, *i.e.* a participant appears in just one class. In contrast, subset I has the interesting property that all the participants are in both classes, once highly depressed and once mildly depressed. It would be very intriguing to see the body movement pattern changes in an individual picked up by an automatic framework.

IV. Method

Given an input video $\mathcal{V}$ with N frames, the Mixture of PartS (MoPS) based human body detector [26] is employed on every second of $\mathcal{V}$ to obtain an approximate location of various body parts in a frame. Then, Relative Part Movement (RPM) is computed for different parts considering the torso as a reference point (see Section IV-A). The motion pattern for each part is combined for full body movement analysis. For a holistic body movement analysis (see Section IV-B), STIP are computed on the entire video sequence to capture overall movement by the body. The output from the parts based detector is used to define a probable Region Of Interest (ROI). Key interest points are selected from STIPs detected in the ROI and are further embedded in a bag of features framework. Figure 1 shows the flow of the approach.

A. Relative Part Movement

We are interested in investigating the discriminative ability of body movements. We used the MoPS framework, which is an extension of the pictorial structure framework [27]. MoPS is the state-of-art body part detector and this motivated us to apply it here. It deals with the parts orientation and the MoPS then represents the parts of an object as a graph with n vertices $V = \{v_1, \ldots, v_n\}$ and a set of edges E. Here, each edge pair $(v_i, v_j) \in E$ encodes the spatial relationship between parts i and j. The human body is represented as a tree-graph. Formally, for a given image I, the MoPS framework computes a score for the configuration $L = \{l_i : i \in V\}$ of parts based on two models: an appearance model and a spatial prior model. The appearance model scores the confidence of a part-specific

template w_p applied to a location l_i. Here, p is a view-specific mixture corresponding to a particular head pose. $\phi(I, l_i)$ is the histogram of gradient descriptor [28] extracted from a location l_i. Thus, the appearance model calculates a score for configuration L and image I by computing confidence maps. The confidence maps are a part-specific detector response. The shape model learns the kinematic constraints between each pair of parts. During inference, the task is to maximise a score function:

$$Score(I, L, p) = App_p(I, L) + Shape_p(L) \qquad (1)$$

L is the part location l_i where $l_i = \{x_i, y_i\}$. In order to compute the motion pattern in the body parts, the torso centre $l_t = \{x_t, y_t\}$ is considered as the reference point. Polar coordinates for each part representing orientation and distance from torso centre are computed from the real pixel coordinate value. For a part i, the orientation θ_i is computed as $\arctan(\frac{y_i - y_t}{x_i - x_t})$ and the distances r_i as $\sqrt{(x_i - x_t)^2 + (y_i - y_t)^2}$. These values are computed for all the parts for every second of $\mathcal{V}$. The motion patterns are joined part-wise and a 2D histogram is computed based on the values of (θ_i, r_i) depicting overall movement by all the parts combined. The histogram represents the relative movement of the body parts in $\mathcal{V}$. Therefore, we call it the *Relative Parts Movement (RPM)* histogram. RPM proved to be discriminative for both the subsets used in the experimental validation. Figure 2 describes the different movement patterns captured by RPM for same participant in two different states of MDD.

B. Holistic Body Movement

The STIP framework proposed by Laptev *et al.* [17] has been widely used for many computer vision problems such as human action recognition [19]. Building on the concept of the 2D Harris corner point detector, it identifies salient points in spatial-temporal domain. A significant local variation in image value in both, space and time domain results in an interest point. Further, Histogram of Gradient (HOG) and Histogram of Flow (HOF) are computed for each detected interest point in a fixed spatial and temporal window. STIP capture almost every small change in the video $\mathcal{V}$ including all the subtle movements exhibited by patients.

As seen in the example Figure 1, the background in the video sequences of the data subset is static. However, some of the frames captured noise due to inadvertent interference of the interviewer (who is sitting right next to the camera recording the patient). Although, the interference is generally not occluding the patient, it causes the generation of some spurious interest points, which do not represent movement by the patient. This issue is addressed by using the output from MoPS based human body detector to define an approximate ROI. STIP detected outside the ROI are excluded from further analysis.

The total number of interest points detected on the video sequences of both the data subsets are quite large in number. To make further analysis memory and computationally inexpensive, a key-interest point selection approach, similar to concept frame selection in affect analysis [29], is used. For a video $\mathcal{V}$, a total of K detected interest points are clustered using the Approximate Nearest Neighbour (ANN) algorithm and

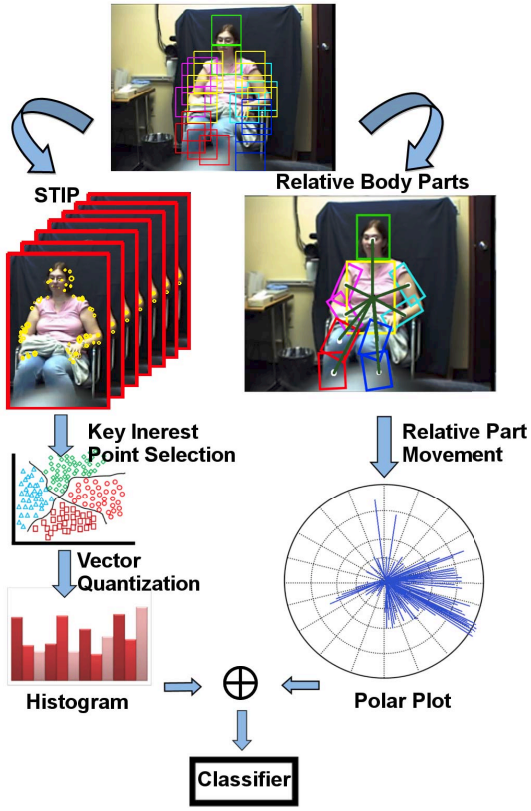

Fig. 1. Flow of the proposed system. Given a video containing a subject, body parts are detected using [26]. STIP are computed on the body window. Key interest points are chosen and vector quantisation is performed and a histogram is generated. For analysing the relative part movements, parts centres are computed and their relative position is calculated with respect to the torso centre. A polar histogram is computed depicting motion patterns. The two histograms are concatenated and an SVM model is used to infer the label.

are represented by k cluster centres. These k spatio-temporal descriptors from all the video sequences in the data subsets are used to create a spatio-temporal Bag of visual Features (BOF). The concept of BOF has been widely used in image processing [19]. It is derived from the Bag-of-Words approach, originated in natural language processing domain. To learn a visual vocabulary C, clustering is further performed on k cluster centres. The size of k and C is chosen empirically.

C. Fusion

The RPM histogram provides relative movement of parts for the entire body. The intra occlusion in human body may lead to the problem of double counting in MoPS [30]. Xu *et al.* [22] reported in their work that polar histograms might give ambiguous results in some motion types. These shortcomings are prone to propagate error in RPM. Overall body movements captured by the STIP framework gather very subtle changes in the consecutive frames. However, at times a few number of detected interest points across a small area may not always give a global representation of body movement. Thus, to overcome any ambiguity and to equilibrate any missing information, overall body movements captured by the STIP framework will

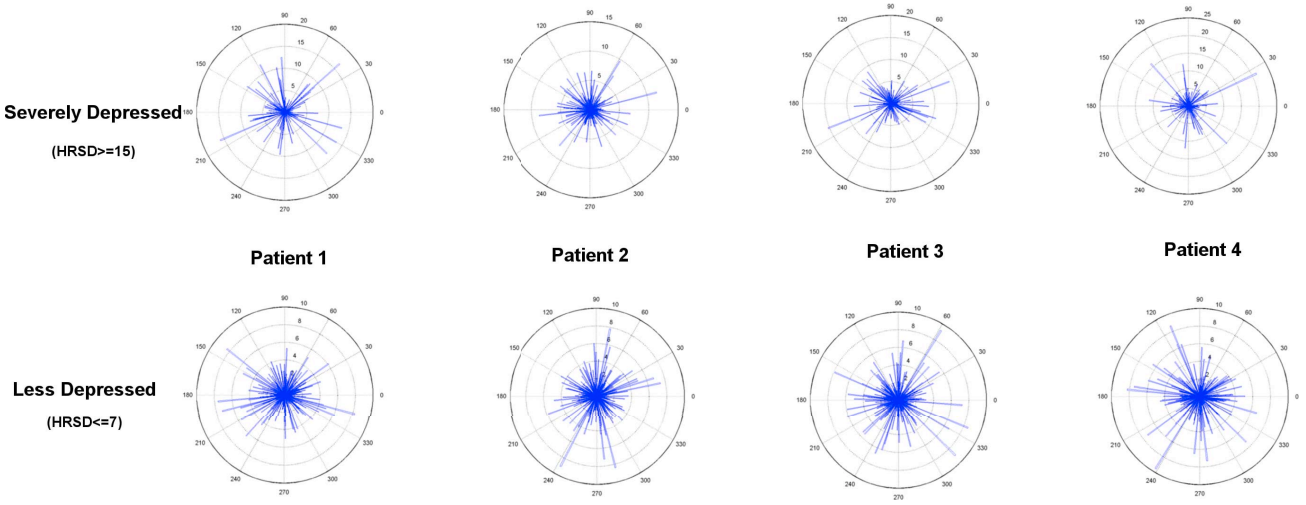

Fig. 2. The figure describe the RPM histograms for four subjects who were severely depressed at one point in time and have shown improvement over the course of treatment. Each column shows two plots depicting motion patterns, which belong to the same participant in two different states. It is visually apparent that as the participants' HRSD score decreases, their RPM shows higher activity.

be fused with RPM histograms. As found in our previous study [1], decision fusion performs best for combining information for multiple modalities. Thus, we use decision fusion to merge different channels. A second stage SVM classifier is learnt on the decisions from RPM histogram and holistic body movement analysis to produce the final result.

V. Experiments and Results

The video rate of the data in our clinical database is 30 Hz and spatial resolution of video frame N is 640×480 pixels. Please note that, as mentioned in Section III, the experiments are performed on two subsets of data: Subset I has the same 12 participants in both the classes and Subset II contains a total of 36 different participants; 18 in each class.

To estimate the human body position, the MoPS is computed on every 30th frame. This framework computes and gives the location of 26 body parts. It is argued in [20] that the larger number of parts leads to accurate detection. However, in our study, the objective is to analyse overall body movement and depict varied motion patterns that can distinctively identify severely depressed patients. Thus, relative patterns of all the body part's motion are fused into a single histogram. The bounding boxes in Figure 1 represent the detected body parts. Now further for computing RPM, 9 body parts relative to the torso centre are considered. These body parts consist of the upper and lower parts of left and right arm, head, and upper and lower parts of left and right leg (Fig. 1 second column). The central location of 9 parts is transformed into polar coordinates (θ_i, r_i). For computing the RPM histogram, several bin sizes are evaluated for orientation $\theta \in (0,2\pi)$ and distance r to obtain an optimal performance.

In the clinical database, the duration of the interview recording varies for every participant. Thus, to perform a fair comparison of overall body movement depicted by the RPM histogram normalisation is performed. First a histogram vector is generated for every minute and is finally normalized for the entire sequence to generate the final RPM histogram. The resultant vector is further used to classify between severely depressed and mildly depressed populations.

For the holistic body movement analysis, STIP is computed for entire clip length for each participant in both the subsets. Harris 3D point detector is used for computing STIP. Around each detected interest point, HOG with a spatial window size of 3 and HOF with a temporal window size of 9 is computed. The total number of interest points generated in subset I is 4.02×10^6 and in subset II is 5.62×10^6. To compute the key-interest points ANN is applied to the interest points with different cluster sizes k varying between 2500, 5000, 7500 and 10000. An ANN implementation from [31] is used. Then, a BOF is computed over the key-interest points. The optimum size C of the visual dictionary is evaluated empirically for different values of C ranging between [200-1000]. Let us name

Configurations ->		STIP1	STIP2	STIP3	STIP4
SUBSET I	Acc. (%)	75.0	**83.3**	79.1	70.8
	F1-Score	0.70	**0.85**	0.83	0.76
SUBSET II	Acc. (%)	80.5	**91.7**	83.3	83.3
	F1-Score	0.81	**0.91**	0.80	0.80

TABLE I. Best classification accuracies and F-score measures for different configurations of STIP for subset I and subset II for holistic body movement analysis.

Approach		SUBSET I	SUBSET II
RPM	Acc. (%)	75.0	94.4
	F1-Score	0.70	0.94
STIP	Acc. (%)	83.3	91.7
	F1-Score	0.85	0.91
RPM + STIP	Acc. (%)	**87.5**	**97.2**
	F1-Score	**0.89**	**0.97**

TABLE II. This table compares the best classification accuracies and F-score measures from RPM and STIP based approach for subset I and subset II. The last row shows the increase in the performance on fusion.

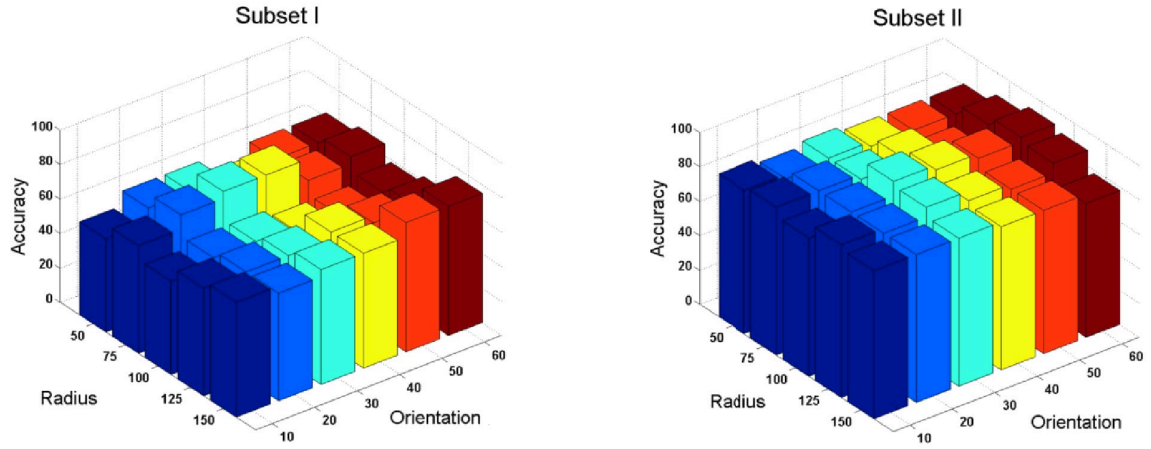

Fig. 3. The figure demonstrate the variation in the performance of RPM histograms with change in bin size of distance (r) and orientatoin (θ) for Subset I and Subset II.

STIP features computed on $\mathcal{V}$ with cluster centre $k = 2500$ as **SITP1**; STIP with cluster centre $k = 5000$ as **STIP2**; STIP with cluster centre $k = 7500$ as **STIP3**; and STIP with cluster centre $k = 10000$ as **STIP4**.

A non-linear Support Vector Machine (SVM) and a leave-one-out approach is used for the classification. A radial basis function kernel is used. The parameters: cost and gamma are selected using an extensive grid search. Table I reports the best performance based on the STIP approach on both the subsets. Please note that in the Table I, STIP2 configuration yields the best performance on Subset I as well as Subset II. In the case of Subset I, STIP2 ($k = 5000$) with codebook size $C = 500$ performs best with 83.3% classification accuracy and an F1-Score as 0.85. For the other data part, Subset II, STIP2 ($k = 5000$) with codebook size C = 750 gives the highest performance with 91.7% accuracy and a F1-Score value of 0.91. Further increase in the number of key-interest points, $k = 7500$ and $k = 10000$, the performance saturates and gradually drops for Subset I. However, for Subset II, it decreases and saturates beyond $k = 5000$. In the case of Subset I, for STIP1; best performance is, accuracy = 75% and F1-Score = 0.7 with codebook size $C = 200$. STIP3; is accuracy= 79.1% and F1-Score = 0.83 for codebook size $C = 750$ and lastly for STIP4 the reported highest performance is with codebook size $C = 750$.

Figure 3 presents the variation in classification accuracy for a change in the bin size of orientation and distance. We find that the RPM histogram performs best with 75 bins for distance and 40 orientation bins, *i.e.* 75×40 giving a classification accuracy of 75% and a F1-Score of 0.7. The best performance on Subset II is 94.4% classification accuracy and a F1-Score of 0.94 with 100 distance and 30 orientation bins, *i.e.* 100×30 (the same is reported in the first row of Table II). It is noteworthy that for subset II, where there are different participants across the classes, the RPM histogram performs *better* than the STIP approach, giving a 94.4% classification accuracy. However, the same is not true for subset I. It can be argued that the RPM histogram depicts relative motion by parts in the body holistically, whereas STIP are very sensitive to any subtle motion exhibited by the participant. When we are observing the same participant in different depression severity situations, these subtle changes are more distinctive to assess the state.

The RPM for four different participants is shown in Figure 2. The upper row in the table depicts the normalised relative motion patterns while the participants were severely depressed with HRSD>= 15. The bottom row presents the normalised relative motion patterns of the same participants while they were less depressed (with a HRSD<= 7). It is evident from the the figure that the motion patterns of the participants with better health are 'denser' and, hence, depict more movement as compared to when they were severely depressed.

Furthermore, the results from the RPM histogram and the STIP approach are fused by learning another non-linear SVM on the decisions from the single channels. The last row in Table II shows the increase in the individual components after fusion. The accuracy on Subset I increases to 87.5% and F1-Score to 0.89. It is almost 12% (absolute) and 4% (absolute) increase as compared to RPM and STIP approaches respectively. On Subset II, the fused performance exceeds the performance of both the channels. It gives 97.2% accuracy and F1-Score = 0.97, which is an increase of almost 3% (absolute) and 5% (absolute) as compared to RPM and STIP approaches respectively.

VI. Conclusions and Future Work

In this paper, a human body parts based relative and local motion pattern (STIP) based depression detection framework has been proposed and evaluated. Relative parts movement histograms are computed by analysing the relative movement of body parts with respect to the torso. The radius and orientation are fused in a polar histogram and a support vector based classifier is learnt. The experiments show that the RPM has good discriminative ability. To augment the system, a bag-of-words based framework is created, which analyses the intra-facial dynamics and local motion in any body part. The fusion of the two modalities shows their complementarity and the effectiveness of a bimodal system.

As part of ongoing work, we plan to create part-specific bag-of-word based dictionaries. Using a spatial pyramid and part-specific dictionaries, an explicit spatial constraint can be applied. In the current system, the relative parts movement histogram is accumulated to create one histogram. As a natural next step, individual part-specific polar histograms should be computed. It will be interesting to explore different fusion scenarios, as part-specific histograms may contain overlapping information due to occlusion.

REFERENCES

[1] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal Assistive Technologies for Depression Diagnosis and Monitoring," *Springer, Journal on Multimodal User Interfaces*, 2013.

[2] J. Joshi, R. Goecke, M. Breakspear, and G. Parker, "Can body expressions contribute to automatic depression analysis?" in *Proceedings of the International Conference on Automatic Face and Gesture Recognition FG2013*, 2013.

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

[4] 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.

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

[6] 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*, ser. ACII'09, 2009, pp. 1–7.

[7] 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.

[8] J. Saragih and R. Goecke, "Learning AAM fitting through simulation," *Pattern Recognition*, vol. 42, no. 11, pp. 2628–2636, 2009.

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

[10] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear, "An Approach for Automatically Measuring Facial Activity in Depressed Subjects," in *Proc. Affective Computing and Intelligent Interaction*, ser. ACII'09, Sep. 2009, pp. 223–230.

[11] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, Jan 1995.

[12] 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

[13] 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 ACII2007*, 2007, pp. 71–82.

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

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

[16] W. Tryon, *Activity Measurement in Psychology and Medicine*, ser. Applied Clinical Psychology Series. Springer, 1991.

[17] I. Laptev and T. Lindeberg, "Space-time Interest Points," in *International Conference on Computer Vision (ICCV)*. IEEE, 2003, pp. 432–439.

[18] 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 ICPR2012*, 2012, pp. 2634–2638.

[19] 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 CVPR2008*, 2008, pp. 1–8.

[20] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011, pp. 1385–1392.

[21] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition," *Pattern Recognition*, pp. 2562–2572, 2012.

[22] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and J. J. Corso, "Combining skeletal pose with local motion for human activity recognition," in *Articulated Motion and Deformable Objects - 7th International Conference, AMDO 2012, Port d'Andratx, Mallorca, Spain, July 11-13, 2012. Proceedings*, 2012, pp. 114–123.

[23] M. First, M. Gibbon, R. Spitzer, L. Benjamin, and J. Williams, *User's Guide for the Structured Clinical Interview for DSM-IV Axis II Personality Disorders: SCID-II*. American Psychiatric Press, Incorporated, 1997.

[24] *Diagnostic and Statistical Manual of Mental Disorders Fourth Edition*, 4th ed. Washington, DC: American Psychiatric Association, 1994.

[25] M. Hamilton, "Development of a Rating Scale for Primary Depressive Illness," *British Journal of Social and Clinical Psychology*, vol. 6, no. 4, pp. 278–296, 1967.

[26] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts." in *Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2011, pp. 1385–1392.

[27] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, ser. CVPR'05, 2005, pp. 886–893.

[29] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *IEEE Automatic Face and Gesture Recognition 2011 workshop FERA*, 2011.

[30] I. Radwan, A. Dhall, J. Joshi, and R. Goecke, "Regression based pose estimation with automatic occlusion detection and rectification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 121–127.

[31] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.