Depression Analysis: A Multimodal Approach

Jyoti Joshi
Faculty of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia
Jyoti.Joshi@canberra.edu.au

ABSTRACT

Depression is a severe mental health disorder causing high societal costs. Current clinical practice depends almost exclusively on self report and clinical opinion, risking a range of subjective biases. It is therefore useful to design a diagnostic aid to assist clinicians. This project aims at developing a novel multimodal framework for depression analysis. In this PhD work, it is hypothesized that a multimodal affective sensing system can better capture what characterises a person's affective state than single modality systems. The project will explore facial dynamics, head movements, upper body gestures, EEG measures and speech characteristics related to affect, in subjects with major depressive disorders. Integrating the individual sensing modalities, a multimodal approach that show improved performance characteristics over single modality approaches will be developed.

Categories and Subject Descriptors

I.5 [**PATTERN RECOGNITION**]; I.5.4 [**Pattern Recognition**]: Applications; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Design, Human Factors, Experimentation

Keywords

Affective Computing; Depression Analysis; Multimodal framework

1. OBJECTIVE

This PhD project aims to solve the fundamental affective computing problem of developing robust non-invasive approaches for sensing a person's affective state with a mental health disorder such as depression and for further improving the robustness of such approaches through multimodal fusion. The focus of this research is (a) to investigate and develop multimodal affective sensing technology in general and (b) to validate the utility of such technology on the real-world example of detecting and monitoring mental health disorders, such as depression in particular. The proposed multimodal approach is aimed at underpinning a new generation of objective laboratory-style markers of illness expression. The central hypothesis is that a multimodal affective sensing system can better capture what characterises a person's affective state than single modality systems.

Depression and other mood disorders are common disabling disorders. Their impact on individuals and families is pro-found. The landmark WHO Global Burden of Disease (GBD) report by Mathers *et al.* [8] quantified depression as the leading cause of disability world-wide. Despite the high prevalence, current clinical practice depends almost exclusively on self-report and clinical opinion, risking a range of subjective biases. There currently exist no laboratory-based measures of illness expression, course and recovery, and no objective markers of end-points for interventions in both clinical and research settings. This compromises optimal patient care, compounding the burden of disability. As health care costs increases, the provision of effective health monitoring systems and diagnostic aides is highly important. With the advancement in affective sensing technology and machine learning, computer aided diagnosis can play a major role in providing an objective assessment. Thus, the aim is to develop robust multimodal affective sensing technology that gives clinicians additional tools for depression diagnosis and monitoring.

2. MOTIVATION & BACKGROUND

Depression is a severe psychiatric disorder that has a major impact on society and the workforce. The 1997 Australian Survey of Mental Health and Well-Being reported a 12-month prevalence rate of 6.3% for major depression [1]. The GBD report [8] quantified depression as the leading cause of disability in (10% of all years of life lost due to disability) and projected it to be the second-leading cause of disease burden by 2020. Disturbances in the expression of emotion reflect changes in mood and interpersonal style, and are arguably a key index of a current depressive episode. This leads directly to impairments in interpersonal functioning, causing a range of interpersonal disabilities including intimate relationships (and their loss), functioning in the workforce and difficulties with a range of everyday tasks (such as shopping). While these are a constant source of distress in affected subjects, they have arguably received insufficient research attention.

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

Affective state recognition has been an active field of research over a decade but with a limited attention towards applicability of these techniques for depression analysis. Researchers have also explored various multimodal approaches for better affect recognition. Zeng *et al.* [18] presented a thorough survey on all the existing approaches and outlined some of the challenges.

In one of the first works for automatic depression analysis, Cohn *et al.* [3] explored the relationship between FACS-based facial features, vocal features and clinical depression detection. They trained subject-dependent Active Appearance Models (AAM) [16] to automatically track facial features. The shape and appearance features after AAM fitting are further used to compute parameters such as the occurrence of action units (associated with depression), mean duration, ratio of onset to total duration and ratio of offset to onset phase. Vocal behaviour analysis was performed by pitch extraction and the results from two different modalities were compared. The audio and video features were not fused in this work.

According to the hypothesis proposed by Ellgring [5], depression leads to a remarkable drop in facial activity, while with the improvement of subjective well-being facial activity increases. Considering Ellgring's hypothesis as a starting point, McIntyre *et al.* [10] analysed the facial response of the subjects when shown a short video clip. Like Cohn *et al.* [3], subject-specific AAMs were trained and shape features were computed from every fifth video frame. The shape features were combined and classified at the frame level by the means of Support Vector Machine (SVM). However, this work just analysed the static features where as facial activity is quite dynamic in nature. It has been shown in the literature that temporal facial dynamics provides more information than using static information alone [2]. A limitation of both [3] and [10] is their use of subject-specific AAM models. For a new subject, a new AAM model needs to be trained, which is both complex and time consuming.

Speech analysis has also been used for depression analysis [12]. Research into possible indicators of affective disorders, such as depression, has explored subtle changes in speech characteristics. Depression patterns of speech have been recognised for many years, finding differences in the pitch, loudness, speaking rate, and articulation [12]. It has been found that depressed subjects have a lower dynamic range for the fundamental frequency than normal subjects, which increases after treatment [14]. Moreover, it has been found that depressives have a slower rate of speech and a relatively monotone delivery when compared to normal speaking patterns as well as lacking in significant expression [11]. Also, there is convincing evidence that sadness and depression are associated with a decrease in loudness [17]. In addition, mel-frequency cepstral coefficients (MFCC) were investigated [4] and their classification results were significant for detecting depression.

Researchers have been using physiological measures such as EEG [15], MRI [6] *etc.* to analyse the affective state of a person. It is evident from the literature that EEG [15] and other psychophysiological signals [9] can effectively help in analysing someone's affective state. But so far, to the best of our knowledge, there has not been any attempt to integrate these individual modality techniques to establish profound multimodal approaches for depression analysis.

3. CHALLENGES

Data Collection - Collecting relevant data is one of the biggest challenges involved in this project. The project is in collaboration with the Black Dog institute [1], a medical research institute in Sydney, where a mood/affect induction experimental setup is used to collect data. During the session, the participant's facial and vocal expressions are recorded while they watch a video, a set of images and are interviewed with a set of questions to elicit a certain emotion. A common problem is the limited number of the participants. Depressed participants are selected who are diagnosed with depression but no other mental disorders and medical conditions. Control subjects are carefully selected to have no history of mental illness and to match the depressed subjects in age and gender. This is an ongoing study and the number of participants is expected to further grow. Due to the sensitive nature of data and privacy issues, one cannot share and expect other research groups working in similar area to share the data. To analyse subjective well-being and improvement/worsening of participants over time, it is also required to have multiple sessions with the same participants, which again is not an easy task. Furthermore, this data is spontaneous in nature as opposed to current main stream facial dynamics research data, which is typically acted.

Feature Selection - Choosing the right features, which capture the expression information and have high discriminative power, is very important for robust affect analysis and a non-trivial task. For vision based facial expression analysis, features such as the Scale-Invariant Feature Transform (SIFT) and the Pyramid of Histogram of Oriented Gradients (PHOG) have been used recently and have been found to give promising results. In this project, the associated temporal information for the facial expressions, head pose/movements and shoulder movements will also be explored. In that case, the feature vectors should be able to capture the spatio-temporal variation. For audio based affect analysis, both linguistic and paralinguistic cues are taken into account. However, finding an optimal feature set that can best describe depression in the audio modality is quite a challenging task.

Curse of Dimensionality - Extensive long interviews, which are initially aimed at capturing different emotions from the participant, result in very long feature vectors. Dimensionality reduction methods such as Principal Component Analysis (PCA), Latent Discriminant Analysis (LDA), Fixed Component Analysis (FCA), or Kernel PCA (KPCA) need to be applied to the extracted features so as to use the highly discriminating features and omitting the non-contributing features.

Robust Face Tracking - Robust face tracking is a very well researched problem. It has met with success but there is a scope of extending the current state of the art AAM [16] to 3D for achieving pose and illumination invariance. 3D morphable model exist but are computationally very expensive. For vision based depression diagnosis while analysing facial expressions, robust face tracking plays a key role.

Person Independent Model - Being able to create a person-independent face model, which can generically fit to any unseen data, is another very crucial point involved in this research. The biggest problem posed by working with

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

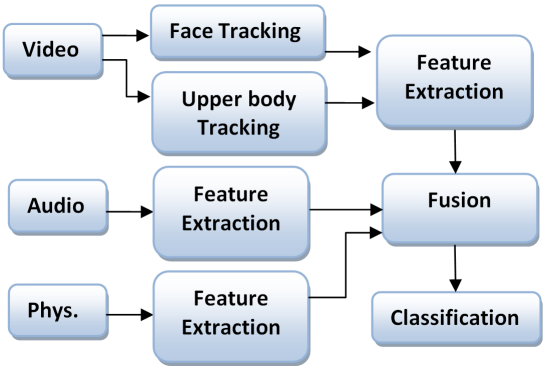

Figure 1: Flow of the proposed system.

person-dependent models is that for every new participant, one needs to train the system, which is highly time consuming and computationally expensive. One possible way of addressing this issue is to be able to train the algorithms on a large amount of data (which is also challenging to obtain). Moreover, new machine learning methods of using unlabelled data for learning robust system will be explored.

4. APPROACH & METHODOLOGY

The objective is to first investigate various uni-modal techniques for depression analysis. Depression analysis methods can be broadly classified into four main categories: a) Vision based, b) Audio based, c) Physiological measures based, and d) Fusion based methods. Initially, individual channels will be analysed for their discriminative ability and later various fusion methods will be experimented on for finding the most discriminative combination of the three different channels. Figure 1 outlines the proposed framework.

4.1 Vision Based

As vision based expression analysis is a well studied problem, various facial expression analysis methods and their aptness for depression analysis will be explored. Studies by psychologists have shown that depressed people show fewer facial movements and also display reluctance to participate in social activities. This can be analysed through head pose/movements and body gestures. Integration of these different behavioural cues seems to be of particular importance when judging such complex affective states.

Facial Expression Analysis - Facial expressions are dynamic in nature and convey a lot about the affective state of a person. So far, virtually all research in automatic facial expression recognition (FER) has been on static, single images. This completely leaves the temporal aspect of facial expressions consisting of onset, apex and offset aside, yet for human observers, this is an important clue in correctly judging someone's facial expression. In recent years, the Local Binary Patterns (LBP) [19] and Local Phase Quantisation (LPQ) [13] classes of visual descriptors have been widely employed for FER. Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) [19] considers patterns in three orthogonal planes: XY, XT and YT, and concatenates the pattern co-occurrences in these three directions. The LBP-TOP descriptor assigns binary labels to pixels by thresholding the neighbourhood pixels with the central value assuming that the XY plane is always near to the apex of expression. Problematic is that this assumption does not always hold true, particularly in the case of spontaneous expression analysis over $30 - 40$min long video recordings. LBP-TOP will be extended by addressing this key issue.

Head Pose/Movements - It is worthwhile to investigate features such as optical flow and features based on Space Time Interest Points (STIP) [7], which have found much attention in temporal video analysis recently to analyse head pose and movements.

Upper Body Gesture Analysis - Pose estimation is a classic problem in computer vision. Analysing the pose and upper body movements of subjects can provide useful information. Pose estimation based on pictorial structures and similar methods will be explored. Further, pose dynamics will be fused with face information.

4.2 Audio Based

In the audio modality, the focus is on prosodic patterns in the speech signal. To this end, existing freely available software tools, such as Praat[2] will be employed to analyse range, mean, median, and variability of the pitch, pitch contour, intensity / energy of the speech signal, speech rate and duration, and lag in responding to stimuli and questions. Initially, the approach will be based on the most basic conceivable emotion / mental state recognition system, with design choices in accordance with simple baseline systems of this kind commonly found in the literature.

4.3 Physiological Measures Based

Human electroencephalography (EEG) measures electrical activity generated by the brain. The use of EEG has enabled researchers to study regional brain activity and brain function, in particular various human cognitive and emotional processes. Recently, EEG information has also been recorded as part of data collection and will be explored for depression detection.

4.4 Integration of Uni-Modal Techniques

The primary objective is to validate the proposed affective sensing approaches by characterising the visual, audio and physiological data as a function of severity and the presence or absence of melancholic features to better understand disturbances in major depression. The individual sensing modalities will be integrated to develop multimodal approaches with improved performance characteristics over uni-modal approaches. This will be achieved by understanding the co-variation between the affective data and the behavioural, disability and neuro-cognitive data using standard multi-variate techniques, such as logistic regression and partial least squares.

5. PROGRESS TO DATE

In one of the initial work, visual cues for depression analysis have been investigated. A holistic model for upper body movements and facial dynamics is explored. STIP are computed for the videos for analysing the upper body move-

[2] `http://www.fon.hum.uva.nl/praat/`

ments and a temporal visual words dictionary is learned from it. Intra-facial muscle movement is captured by computing LBP-TOP on the aligned face. Due to a large number of interest points, the size of the feature vector grows very quickly. This problem is addressed by defining multiple spatio-temporal grids in a bag of visual words approach inspired by [7] for LBP-TOP and STIP separately.

Initial experiments in the direction of multimodal depression analysis are also performed by exploring visual and audio cues and fusing them to improve the performance of the system. For the video data analysis, intra-facial muscle movements and the movements of the head and shoulders are analysed by computing STIP. In addition, various audio features are computed. Next, a bag of visual features and a bag of audio features are generated. Fusion methods at feature level, score level and decision level have been compared.

As part of ongoing research, the work of McIntyre *et al.* [10] is being extended by scaling it to a larger dataset, investigating use of temporal features and using person independent AAM tracking in the existing framework.

6. RESEARCH TIMELINE

First six months are for survey of existing technologies and performing initial experiments looking at the visual cues mainly facial dynamics for depression analysis. In later half of first year, the plan is to explore temporal information at both feature extraction and classification level. In second year, contribution of other visual cues such as body pose and head movements in depression analysis will be investigated. Other modalities viz. audio and physiological measures will be analysed and different fusion methods will be explored in the final year.

7. ACKNOWLEDGEMENTS

I would like to thank Dr. Roland Goecke, primary supervisor and panel chair, and Prof. Michael Wagner for their invaluable guidance and support. I would also like to thank collaborators at Australian National University and Queensland Institute of Medical Research, Australia. This PhD project is funded by Australian Research Council discovery project grant DP110103767.

8. REFERENCES

[1] Mental health and wellbeing: Profile of adults. Technical report, Australian Bureau of Statistics, Australia, 1997.

[2] Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, May 2005.

[3] 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)*, pages 1–7, 2009.

[4] N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An Investigation of Depressed Speech Detection: Features and Normalization. In *Proc. Interspeech*, 2011.

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

[6] K. Krishnan, J. Hays, and D. Blazer. MRI-defined vascular depression. *American Journal of Psychiatry*, 154(4):497–501, Apr. 1997.

[7] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2008)*, 2008. DOI: 10.1109/CVPR.2008.4587756.

[8] C. Mathers, T. Boerma, and D. M. Fat. The global burden of disease: 2004 update. Technical report, WHO Press, Switzerland, 2004.

[9] M. Mauri, V. Magagnin, P. Cipresso, L. Mainardi, E. N. Brown, S. Cerutti, M. Villamira, and R. Barbieri. Psychophysiological signals associated with affective states. In *32nd Annual International Conference of the IEEE, Engineering in Medicine and Biology Society*, pages 3563–3566, 2010.

[10] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear. An Approach for Automatically Measuring Facial Activity in Depressed Subjects. In *Proc. ACII2009*, pages 223–230, Sept. 2009.

[11] 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*, volume 55, pages 96–107, 2008.

[12] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts. Voice acoustic measures of depression severity and treatment response collected via interactive voice response technology. *Journal of Neurolinguistics*, 20(1):50–64, 2007.

[13] V. Ojansivu and J. Heikkilä. Blur Insensitive Texture Classification Using Local Phase Quantization. In *Proc. 3rd International Conference on Image and Signal Processing*, Lecture Notes in Computer Science, pages 236–243. 2008.

[14] 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.

[15] H. Peng, B. Hu, Q. Liu, Q. Dong, Q. Zhao, and P. Moore. User-centered depression prevention: An EEG approach to pervasive healthcare. In *Proc. PervasiveHealth*, pages 325–330. IEEE, May 2011.

[16] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, 2009.

[17] K. R. Scherer. *Vocal assessment of affective disorders*, pages 57–82. Lawrence Erlbaum Associates, 1987.

[18] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, Jan. 2009.

[19] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.