Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Int Conf Autom Face Gesture Recognit Workshops. Author manuscript; available in PMC 2016 July 21.

Published in final edited form as:

IEEE Int Conf Autom Face Gesture Recognit Workshops. 2015 May ; 1: .

Cross-Cultural Detection of Depression from Nonverbal Behaviour

Sharifa Alghowinem[1,7], **Roland Goecke**[2,1], **Jeffrey F. Cohn**[3,4], **Michael Wagner**[2,1], **Gordon Parker**[5], and **Michael Breakspear**[6,5]

[1]Australian National University, Research School of Computer Science, Canberra, Australia

[2]University of Canberra, Human-Centred Computing Laboratory, Canberra, Australia

[3]University of Pittsburgh, Pittsburgh, PA, USA

[4]Carnegie Mellon University, Computer Science at the Robotics Institute, Pittsburgh, PA, USA

[5]University of New South Wales, Sydney, Australia

[6]Queensland Institute of Medical Research, Brisbane, Australia

[7]Ministry of Higher Education: Kingdom of Saudi Arabia

Abstract

Millions of people worldwide suffer from depression. Do commonalities exist in their nonverbal behavior that would enable cross-culturally viable screening and assessment of severity? We investigated the generalisability of an approach to detect depression severity cross-culturally using video-recorded clinical interviews from Australia, the USA and Germany. The material varied in type of interview, subtypes of depression and inclusion healthy control subjects, cultural background, and recording environment. The analysis focussed on temporal features of participants‘ eye gaze and head pose. Several approaches to training and testing within and between datasets were evaluated. The strongest results were found for training across all datasets and testing across datasets using leave-one-subject-out cross-validation. In contrast, generalisability was attenuated when training on only one or two of the three datasets and testing on subjects from the dataset(s) not used in training. These findings highlight the importance of using training data exhibiting the expected range of variability.

I. INTRODUCTION

Clinical depression is a mood disorder with high prevalence worldwide, which can result in unbearable pain and disabling conditions that impair an individual’s ability to cope with daily life. The World Health Organization (WHO) lists depression as the fourth most significant cause of suffering and disability worldwide and predicts it to be the leading cause in 2020 [1], [2]. The WHO estimates that 350 million people worldwide are affected by depression [2]. Although clinical depression is one of the most common mental disorders, it

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

is often difficult to diagnose because it manifests itself in different ways and because clinical opinion and self-assessment are currently the only ways of diagnosis. This risks a range of subjective biases. According to the WHO [2], the barriers to effective diagnosis of depression include a lack of resources and trained health care providers. Moreover, evaluations by clinicians vary depending on their expertise and the depression screening instrument used (e.g. Quick Inventory of Depressive Symptoms-Self Report (QIDS-SR) [3], Hamilton Rating Scale for Depression (HRSD) [4], Beck Depression Inventory (BDI) [5]).

We believe that recent developments in affective sensing technology potentially enable an objective assessment. While automatic affective state recognition has been an active research area in the past decade, methods for mood disorder detection, such as depression, are still in their infancy. Our ultimate goal is to develop objective multimodal affective sensing approaches that support clinicians in diagnosis and monitoring of clinical depression. The question of generalisation across datasets is an important issue in this development.

Previous studies on emotion recognition in general and detection of depression in particular have investigated single datasets. By using a single dataset, many intervening variables are kept constant, such as emotion labels and categories, recording settings and environment. Therefore, generalising an emotion recognition system to other corpora might not result in a comparable outcome. An important step towards generalisability of an approach is to apply it in a cross-corpus context.[1] In general emotion studies, cross-corpus generalisation is a very young research area. To the best of our knowledge, only few studies have investigated method robustness on different environments [6], [7], [8]. Speech in particular is immensely affected by recording environment, such as varying room acoustics, and different microphone types and distances [8]. The video channel also has its obstacles regarding recording environment, such as illumination, cameras (focal point, type and distance), and video files (resolution, frame rate and dimensions). Therefore, when dealing with different datasets, several aspects have to be considered to control for variability between them.

The three datasets used in our study come from three different western cultures, which possibly introduces differences in nonverbal behaviour exhibited. They also differ in recording conditions and interviewers. Depression severity was assessed using either self-report or clinician-administered interviews. Diagnosis, in the current study, was defined using standard thresholds on the severity measures (e.g. a score of 15 or higher on the clinician-administered HRSD). Standard thresholds were calibrated with the Diagnostic and Statistical Manual of American Medical Association (DSM IV or V) [9]. While nonverbal behaviour or interpersonal communication varies cross-culturally, symptoms of depression in western societies are similar [10], [11], [12].

We investigate generalising an approach to detect depression from nonverbal behaviour in a cross-cultural context.[2] The approach extracts features from eye activity and head pose modalities, which are then investigated individually and when fused using both feature and

[1]Cross-dataset generalisation is often also referred to as domain transfer.

[2]A common problem in this early stage of a fairly new research area is that it is not clear what terminology best suits the work presented in this paper. In the literature, “cross-corpus” (cross-dataset) is used when the investigated datasets have a similar data

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

hybrid fusion methods. We explore several approaches to generalisability by evaluating different training and testing combinations within and between datasets. We hypothesise that the approach has the ability to generalise within individual datasets and when the classifier is trained on varied observations.

II. BACKGROUND

A few studies in recent years have investigated the automatic detection of depression using pattern recognition techniques on either audio or video input, or multimodal input. Several studies investigated depressed speech characteristics and classifications using prosody features (e.g. [13], [14], [15], [16], [17]) and speech style (e.g. [18]) in adults. Recognising depression from video data has also been investigated, including facial activities and expressions (e.g. [15], [19], [20], [21]), general movements and posture of the body [22], [23], head pose and movement [20], [23], [24], and gaze and eye activity [22], [25]. Moreover, most previous studies on the automatic detection of depression have only investigated a single modality. Multimodal detection of depression is a new research area, with only a few studies investigating fusion techniques for this task [26], [27], [28], [29].

Using multiple depression datasets for generalisation is particularly hard to investigate due to challenging differences in recording environment, recording procedure and depression evaluation, not to mention ethical, clinical, and legal reasons regarding acquiring and sharing such datasets. To the best of our knowledge, only [13], [14] have used different datasets to collect their depression speech samples. In both studies, a preprocessing procedure was performed to compensate for possible differences in recordings. These studies raised concerns of such recording environment differences affecting the results even after feature normalisation methods. Each dataset was used as a separate class, which might affect the classification results, where the classifier might separate the classes based on their recording environment characteristics, not the actual class label.

III. Depression Datasets

Three datasets are used in this study: Black Dog Institute depression dataset (BlackDog) [30], University of Pittsburgh depression dataset (Pitt) [31], and Audio/Visual Emotion Challenge depression dataset (AVEC) [32]. The specifications and differences of these datasets are described here. For easier reference, Table I summarises and compares the selected subsets of each dataset.

The three datasets differ in several aspects that could affect the generalisation. We control for these differences by:

- Each dataset uses different depression screening instruments to measure depression severity. To use a common metric of depression, we converted the different metrics to their QIDS-SR equivalents using the conversion

collection and purpose. However, the datasets used here have several differences that, in our opinion, make the use of such terminology debatable. We chose the term “cross-cultural” to emphasise not only the technical differences but also the possibility of differences in how depression symptoms present themselves.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

table from [33] and categorised subjects based on the severity level of depression.

- While the BlackDog dataset compares depressed patients with healthy controls, both Pitt and AVEC datasets aim to monitor depression severity over time. To overcome this issue, we categorised the subjects in these datasets into two groups for a binary classification task: severe depressed (all datasets) vs. low depressed (AVEC and Pitt) + healthy controls (BlackDog).
- The data collection procedure for each dataset differs. BlackDog uses structured stimuli to elicit affective reactions, which include an interview of asking specific open ended questions, where the subjects are asked to describe events in their life that had aroused significant emotions to elicit spontaneous, self-directed speech and related facial expressions, as well as overall body language. The Pitt data collection procedure was conducted by interviews using the HRSD questions, where patients were interviewed and evaluated by clinicians. On the other hand, the AVEC paradigm is a human-computer interaction experiment containing several tasks including telling a story from the subject's own past (i.e. best present ever and sad event in the childhood). Therefore, in this study, we only analysed the childhood story telling from AVEC in order to match the interviews from the BlackDog and Pitt datasets.
- While BlackDog records one session per subject, AVEC and Pitt have multiple sessions per subject (up to four). In this study, we only select one session for each subject, thereby splitting the subjects into two groups: severe depressed vs. low depressed + healthy controls. We also aim to have a balanced number of subjects in each class to reduce classification bias towards the larger classes; hence, the relatively small number of selected subjects, but this is a common problem in similar studies.
- The duration of segments for each subject in each dataset varied. Therefore, to reduce variability from the length of subjects' segments, we extracted temporal features over the entire segments (Section IV-A).
- Recording environment and hardware differ for each dataset. The audio channel in particular is more vulnerable to the recording environment than the video channel (e.g. microphone distance, background noise, sampling rate). Therefore, we only focus on analysing nonverbal behaviour from eye activity and head pose.

IV. METHOD

In this section, a brief overview of the approach used to investigate the cross-dataset generalisation of depression detection is given. Fig. 1 shows the general design and individual components.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A. Feature Extraction and Normalisation

1) Eye activity—To accurately detect eye activity (blinking, iris movements), the eyelids and iris must be located and tracked. To this end, we train and build subject-specific 74-points Active Appearance Models (eye-AAM). For each eye, horizontal and vertical iris movement, and eyelid movement are extracted as low-level features per frame (30 fps), following [25]. A total of 126 statistical features ("functionals") are calculated over the entire subject's segment to reduce inter-subject differences in segment lengths:

- Maximum, minimum, mean, variance, and standard deviation for all 18 low-level features mentioned earlier (5 × 18 features)
- Maximum, minimum, and average of duration of looking left, right, up and down, as well as of blink duration for each eye (3 × 2 eyes × 5 features)
- Closed eye duration rate and closed eye to open eye duration rate for both eyes (2 eyes × 2 features)
- Blinking rate for both eyes (2 eyes × 1 feature)

2) Head pose—To extract head pose and movement behaviour, the face has to be detected and tracked before a 3 degrees of freedom (DOF) head pose could be calculated. We use an optimised strategy of constrained local models (CLM) [34]. The CLM used for face detection contains 64-points around the face, which are projected into a 58-points 3D face model following [24] to extract 3-DOF head pose features, as well as their velocity and acceleration, giving a total of 9 low-level features per frame. Over the duration of each each subject's segment, a total of 184 statistical features are extracted, which are:

- Maximum, minimum, range, mean, variance, and standard deviation for all 9 low-level features mentioned earlier. (6 × 9 features)
- Maximum, minimum, range and average duration of head direction left, right, up and down, tilting clockwise and anticlockwise. (4 × 6 features)
- Head direction duration rate, and rate of different head directions for non-frontal head direction for all directions mentioned above. (2 × 6)
- Change head direction rate for all directions mentioned above. (1 × 6 features)
- Total number of changes of head direction for yaw, roll, pitch, and all directions. (1 × 4 features)
- Maximum, minimum, range, mean, variance, duration, and rate for slow, fast, steady, and continuous movement of yaw, roll, pitch. (7 × 3 DOF × 4 features)

Moreover, inspired by [8], corpus normalisation is used to eliminate the differences of features from different datasets before their usage in combination with other corpora using min-max normalisation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. Classification and Evaluation

We use Support Vector Machine (SVM) classifiers, which are discriminative models that learn boundaries between classes. SVM has been widely used in emotion classification [35] and is often considered the state-of-the-art, as it provides good generalisation properties. The classification is performed in a binary (i.e. severe depressed vs. low-/non-depressed) subject-independent scenario. LibSVM [36] is used for SVM implementation. To increase the accuracy of SVMs, the cost and gamma parameters are optimised via a wide range grid search for the best parameters using radial basis function (RBF) kernel. Selecting the training and testing sets for classification is performed by two methods:

1) Leave-one-subject-out cross-validation method—To mitigate the limitations of the relatively small amount of data and also to train the classifiers on varied observations (especially when using different dataset combinations), a leave-one-subject-out cross-validation (test on one subject's data, train on all other subjects' data in each iteration, where the number of iterations is equal to the number of subjects) was used on individual dataset classification as well as the combinations of the datasets without any overlap between training and testing data. This method could overcome overfitting the model on the training set, especially as the final selected SVM parameters generalise to all training observations. In other words, the common parameters that give the highest average training accuracy of all training sets in the cross-validation are picked, hence the need for a wide range search. We believe that this method of selecting the parameters reduces overfitting issues on the training set and, therefore, assists in generalising to different observations in each leave-one-out cross-validation turn.

2) Separate train-test dataset method—In this method, one or two datasets are used for training and then the remaining dataset(s) for testing. The SVM parameters are selected based on the highest accuracy of the training set. This method could suffer from overfitting to the training set and might not generalise to the completely different testing set(s). We apply this method to investigate the generalisation ability of the depression detection method to unseen data. We hypothesise that when using different combinations of datasets, leave-one-out cross-validation results in a higher performance than the train-test method, because it trains over varied samples of combined datasets, which reduces model overfitting to the training set. However, both methods are investigated to shed more light onto cross-corpora generalisation.

We measure the performance of the approach in terms of Average Recalls (AR), as it considers the correct recognition in both groups (severe depressed vs. low-/non-depressed) and is more informative than usually reported accuracy. The AR is also called "balanced accuracy" and is calculated as the mean of sensitivity and specificity.

C. Feature Selection

Since irrelevant features lead to a high data dimensionality and may affect the performance of a system, feature selection techniques can overcome this by selecting relevant features. Generally, a subset of features can be selected using statistical function methods, filter

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

methods, search strategies, etc. Since our classification is done in a binary manner, using a simple T-test threshold to perform feature reduction is sufficient. The T-tests are obtained as a two-sample two-tailed T-test, assuming unequal variances with significance $p = 0.05$. Features that exceeded a statistical threshold set in advance by a t-value corresponding to an uncorrected p-value of 0.05 ($p < 0.05$) are selected for the classification problem in two approaches.

1) Variable set of features exceeding the t-statistic based on the combined training data—Features that exceeded the t-statistic are identified from the combined training data and selected on the testing data. Thus, the selected features might vary with each individual and combination of datasets. However, in this approach, with leave-one-out cross-validation, common features that exceed the t-statistic in all turns are selected. That is, using the training subjects in each turn, we apply T-test to all extracted functional features, then only these that commonly exceed the t-statistic in every turn are selected. Then, these common features are fixed and used for all leave-one-out cross-validation turns in the testing. Acknowledging the risk of the feature selection being based on seeing all observations, a fixed number of features in each turn of the leave-one-out cross-validation ensures a fair comparison between turns. On the other hand, in the train-test classification method, the features that exceed the t-statistic in the training set are selected on the testing set.

2) Fixed set of features exceeding the t-statistic—Unlike the variable set of features, we seek to find a fixed set of features that commonly exceed the t-statistic on all individual datasets and combinations of the datasets. This fixed feature set is used on all individual and combinations of the datasets to ensure a fair comparison and also to conclude a set of features that can generalise for the task of detecting depression. This set of features is selected based on a majority agreement of features that exceed the t-statistic on all individual datasets and combinations of the datasets (see supplementary material for list of features).

D. Fusion

Multimodal fusion of different modalities could improve the classification results as it provides more useful information compared to using a single modality. Moreover, fusion can be performed as pre-matching (early) fusion, post-matching (late) fusion, or a combination of both (hybrid) [37]. In this work, we employ a hybrid fusion method that fuses individual modality results with the feature fusion result to obtain a final decision. Feature fusion is implemented here by concatenating the previously selected features. A hybrid fusion employs the advantages of both early and late fusion strategies. A majority voting method is used for the hybrid fusion in a single stage. We strongly believe that hybrid (high-level) fusion overcomes classification errors from individual and feature (low-level) fusion modalities. Since the final hybrid fusion decision relies on the majority agreement of the classification decisions of the fused modalities, outlier classification errors are reduced and the confidence level of the final decision is increased. Moreover, feature fusion also relies on the correlation of the features that are fused. Lack of correlation might affect the final

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

feature fusion, even if individual modalities had high classification result, which hybrid fusion overcomes.

The following comparisons are presented: (1) variable vs. fixed feature set, (2) leave-one-out cross-validation vs. train-test classification, and (3) individual modalities (eye activity and head pose) vs. fused modalities (feature fusion and hybrid fusion). These are tested on different combinations of datasets and then compared with individual datasets. The classification results of individual datasets are shown in Table II and combined datasets in Table III and Fig. 2.

V. RESULTS

A. Classification Results of Individual Datasets

Datasets were used for the classification individually to test the generalisation ability of the approach (see Fig. 1) and to establish a baseline for each dataset to compare their individual results with dataset combination results.

All three dataset classifications performed an average of 80% AR for both feature selection methods, considered to be high, as classification results for all modalities are significantly above chance level. The classification results for each modality of the three datasets are comparable, which supports our hypothesis that the approach used has the ability to generalise when applied individually to each dataset. Comparing feature selection methods, using the variable feature set performed significantly better in most cases than when using the fixed feature set.

For the eye activity modality, the classification results are consistently high for all three datasets, which implies that eye activity is a strong characteristic to differentiate severe depressed from low-/non-depressed behaviour. Similar to BlackDog and Pitt, the lowest result in the AVEC dataset was obtained for the head pose modality. While for both BlackDog and Pitt it performed at 80% AR, for AVEC it performed at 69% AR. This is expected as the task is based on a human-computer interaction scenario with limited head movements. Interestingly, the performance of the head pose modality decreased significantly for BlackDog when using the fixed feature set. The decrease is due to the set of fixed features containing only two features out of seven that exceed the t-statistic for the BlackDog dataset for head modality.

Feature fusion classification performances varied when comparing with the fused modalities. Feature fusion results either were similar or slightly improved from individual modality results in most cases. Two exceptions were found for the AVEC dataset, where feature fusion (1) decreased from the higher result for the variable feature set and (2) had a catastrophic result where the fusion result was lower than the lowest individual modality for the fixed features. This variation in feature fusion performance is not statistically significant. It could be due to several reasons, such as differences in signal quality, depression diagnosis or data collection procedure. Further investigations are needed.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Like feature fusion, hybrid fusion using majority voting of decisions from individual modalities and from feature fusion had a slightly varied outcome compared to the modalities that it fuses for each dataset. In most cases, hybrid fusion results were either similar or slightly lower than the highest modality result, with one exception of a slight improvement (Pitt dataset with fixed features). This variation could imply varied decisions for each subject in the investigated modalities. That is, modalities had varied agreement/disagreement for the same subject. However, as with feature fusion, the classification results of the hybrid fusion were not catastrophic, giving a stronger confidence in the hybrid fusion decisions compared to individual modality decisions.

In general, applying the approach on the datasets individually had high performance, which (1) implies that eye and head modalities and their fusion results had distinguishing characteristics to detect severe depression from low or no depression, and (2) supports the hypothesis the approach is able to generalise to different datasets. Moreover, using the variable feature set, where features that exceed the t-statistic ($p < 0.05$) are selected on each dataset individually, performed better than using a fixed feature set selected based on the majority of features that exceed the t-statistic on dataset combinations.

B. Classification Results of Combination of Datasets

The goal here is to investigate the generalisation ability of the approach to combinations of datasets that differ in several aspects (see Table I), assessing the flexibility and scalability of the approach. Classifications on dataset combinations are performed in two methods: leave-one-subject-out and train-test methods (cf. Section IV-B). Classification results are presented in Table III and Fig. 2, respectively.

1) Leave-one-subject-out cross-validation of dataset combinations—Comparing classification results from individual datasets (see Table II) with the results for dataset combinations (Table III), none of the dataset combination results showed an improvement. Moreover, all classification results of dataset combinations are catastrophic (i.e. statistically lower than the lowest classification results of individual datasets) when using the variable feature set. On the other hand, when using a fixed set of features, the classification results of the dataset combinations were higher than the lowest result of the individual datasets in most cases (three exceptions: eye activity modality for both BlackDog + AVEC and all three datasets, and head pose modality with AVEC + Pitt). This finding suggests that the fixed feature set has a stronger generalisation ability than the variable feature set. A reduction or at least no improvement was expected. Moreover, having higher results from dataset combinations than the lowest result of individual datasets (not catastrophic) is considered a good performance given the differences between the datasets, which supports the generalisability claim of the approach and the selected features.

Several combinations of the three datasets were used for classifying severely depressed subjects from low-/non-depressed subjects to identify, which combination of datasets generalises best. In general, the classification results of dataset combinations in the leave-one-out method are high, performing on average at 70% AR, which implies that this method generalises to different combinations of datasets. The combination of BlackDog + AVEC

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

performed the lowest in most modalities for both feature selection methods. On the other hand, the combinations of BlackDog + Pitt, AVEC + Pitt and all three datasets had reasonably comparable classification results. That could imply that the Pitt dataset, the common dataset in these three combinations, has stronger generalising characteristics.

Comparing feature selection methods, the fixed feature set performed better than when using the variable feature set in most cases. Three exceptions for this finding (eye activity modality for BlackDog + Pitt, head pose modality for AVEC + Pitt, and feature fusion for all three datasets) where using the variable feature set was slightly better than the fixed feature set. That suggests that even when the features are selected based on the specific dataset combination (using the variable feature set), they have a lower generalisation ability in the classification problem when combining two or more datasets than when using a fixed feature set. Moreover, since the fixed feature set is selected based on the majority of features that exceed the t-statistic in all individual and combined datasets, it has more generalisation power for the classification problem when combining two or more datasets than using the variable feature set.

For the eye activity modality, similar to individual datasets, the classification results of dataset combinations in the leave-one-out cross-validation are consistently high, especially when using the fixed feature set. This finding supports, once again, our claim that eye activity has strong distinguishing characteristics in depression detection. On the other hand, the head pose modality performed lower than the eye activity modality, yet the classification results were above chance level in all but one case of dataset combinations. This finding implies that head pose holds useful information for the depression classification task.

We fuse eye activity and head pose modalities via feature and hybrid fusion. Feature fusion improves the classification results compared to the individual modality results that it fuses in all but one case. This combination is BlackDog + Pitt where the classification result was slightly lower than the highest modality but not catastrophic. The improvements in feature fusion results suggest that eye activity and head pose modality features are correlated and complement each other on the task of detecting depression. Hybrid fusion employs early and late fusion by combining the decisions from individual modalities with decisions from feature fusion (Fig. 1), which increases the confidence level of the final decision. In dataset combinations using the leave-one-out cross-validation, hybrid fusion either keeps or slightly improves the classification results from the individual modality results that it fuses in most cases, especially for a fixed feature set. The exception for this is when using the variable feature set, the hybrid fusion classification results slightly decreased form the highest individual modality results. These findings suggest that fusing individual modalities increases both the recognition rate and the confidence level of the final decision.

In general, the classification results on dataset combinations in leave-one-out performed considerably better, even with the dataset differences. We believe that is due to the classifier learning from varied observations from each dataset, reducing the effect of overfitting the model to specific observation conditions. We also believe that the extracted temporal

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

features in general, and the selected features in particular, are robust to different recording conditions.

2) Separate train-test classifications of dataset combinations—Since the fixed feture set performed better than using the variable feature set for the generalisation to dataset combinations with the leave-one-out method, only the fixed feature set is used for the train-test method. The classification results of generalisation using the train-test method and a fixed set of features are shown in Fig. 2.

In general, the classification results when using one or two datasets for training and using the remaining dataset(s) for testing are mostly at or lower than chance level with a few exceptions. That is expected as, unlike the leave-one-out cross-validation method with dataset combinations, the classifier on the train-test method is trained on certain observations of dataset(s), which risks overfitting that reduces the classifier's ability to generalise to separate and different dataset observations (unseen data).

Comparing different train-test dataset combinations, the only combination that has a reasonably above chance level classification result is the AVEC + Pitt dataset combination used for training and the BlackDog dataset for testing. This indicates that when using AVEC + Pitt datasets, the classifier is trained on varied observations where the model is able to generalise to the BlackDog dataset observations. These variations might be due to (1) the classification problem for both AVEC and Pitt being to classify severe depression from low depression and, therefore, the model is trained on wide depression ranges, which might reduce the effect of overfitting, (2) the number of females in the AVEC + Pitt combination is more than half the total number of subjects (47 females out of 70 subjects). It has been reported that women amplify their mood when depressed [38] and, therefore, the AVEC + Pitt combination model is trained on easily distinguishable observations, or (3) the differences in recording conditions and collection procedures, which made it flexible to generalise to the BlackDog recording conditions.

Eye activity classification results were higher than the classification results of the head pose modality in the train-test combinations in all cases with one exception. This finding suggests a higher ability for the eye activity to generalise to different datasets than the head pose modality, which supports our view that eye activity has strong distinguishing characteristics to detected depression.

As with previous classification problems, feature and hybrid fusion were investigated with the train-test dataset combinations. Feature fusion improves the classification results from the highest classification result of individual modalities in most cases of classifying different train-test dataset combinations. This finding suggests that eye activity and head pose features are correlated and, therefore, complement each other. On the other hand, hybrid fusion led to no improvement on classification results from the individual modalities that it fuses in most train-test combination classifications, yet the results were not catastrophic.

To summarise, generalising using the train-test method for classification of dataset combinations performed very low compared to the leave-one-out method, which might be

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

caused by overfitting, as the model is trained on specific conditions that prevent it from generalising to different observations.

By investigating the generalisability of the approach to different dataset combinations using the leave-one-out cross-validation and train-test methods for classification, we conclude that when the classifier is trained on varied observations, the effect of overfitting, which is the main obstacle for cross-dataset generalisation, could be reduced and, therefore, the model has a better ability to generalise to new observations than a classifier trained on specific observations. That is, the more variability in the training observations, the better the generalisability to the testing observations.

VI. CONCLUSIONS

Intending to ultimately develop an objective multimodal system that supports clinicians in diagnosis and monitoring of clinical depression, we investigated generalisability of an approach that extracts nonverbal temporal patterns of depression to cross-cultural datasets. Assuming similar depression symptoms, we apply a depression detection approach on three different datasets (BlackDog, Pitt, AVEC) individually and combined to investigate generalisability and scalability.

These three datasets differ in several aspects including collection procedure and task, depression diagnosis test and scale, cultural and language background, and recording environment. To reduce the differences (1) similar tasks of the collection procedure in each dataset were selected, containing spontaneous self-directed speech, (2) the classification problem was cast as a binary problem (i.e. severe depressed vs. low depressed/healthy controls), (3) functional features were extracted over the entire duration of each subject's segment to reduce duration variability, (4) nonverbal behaviour from eye activity and head pose modalities were investigated because they are less dependent on the recording conditions, and (5) normalisation of the extracted features was applied to reduce recording environment and setting differences.

Although, the eye activity modality has performed better than the head modality for both individual and combined datasets, fusing these modalities in feature and hybrid fusion improved the AR in most cases. We conclude that the eye activity modality has a distinguishing characteristic for detecting depression and also suggest that the two modalities are correlated and complement each other.

In general, applying the approach on individual datasets and their combinations in a leave-one-out cross-validation led to considerably high performance. This supports the hypothesis of the generalisability of the approach to dataset combinations even with the several differences between the datasets. Moreover, we believe that the extracted temporal features are robust to different recording conditions. However, the performances of using one or two datasets for training and the remaining dataset(s) for testing were at chance level, which might be due to overfitting to the training set. We conclude that when the classifiers are trained on varied observations, they have a stronger ability to generalise to new observations than when trained on observations with less variability.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In future work, we will extend the analysis to include datasets from non-western cultures, e.g. Arabic cultures, to further investigate cultural differences in depression expression, which is relevant to evaluating automated depression analysis approaches in different cultural settings.

Acknowledgments

This research was funded in part by the Australian Research Council Discovery Project grant DP130101094.

References

1. W. World Health Organization. The world health report 2003: shaping the future. World Health Organization; 2003.
2. Mathers, C.; Boerma, J.; Fat, D. The Global Burden of Disease: 2004 Update. Geneva, Switzerland: WHO; 2008.
3. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, et al. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. Biological psychiatry. 2003; 54(5):573–583. [PubMed: 12946886]
4. Hamilton M. A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry. 1960; 23(1):56.
5. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. Journal of personality assessment. 1996; 67(3):588–597. [PubMed: 8991972]
6. Schuller B, Zhang Z, Weninger F, Rigoll G. Using multiple databases for training in emotion recognition: To unite or to vote? INTERSPEECH. 2011:1553–1556.
7. Lefter, I.; Rothkrantz, LJ.; Wiggers, P.; Van Leeuwen, DA. Text, Speech and Dialogue. Springer; 2010. Emotion recognition from speech by combining databases and fusion of classifiers; p. 353-360.
8. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G. Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Trans on Affective Computing. 2010; 1(2):119–131.
9. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Vol. 4. American Psychiatric Association; 1994. ApaEditors, Ed
10. Tseng W. Handbook of cultural psychiatry. International Review of Psychiatry. 2001; 14:71–3.
11. Singer K. Depressive disorders from a transcultural perspective. Social Science & Medicine (1967). 1975; 9(6):289–301.
12. Ruchkin V, Sukhodolsky DG, Vermeiren R, Koposov RA, Schwab-Stone M. Depressive symptoms and associated psychopathology in urban adolescents: a cross-cultural study of three countries. The Journal of nervous and mental disease. 2006; 194(2):106–113. [PubMed: 16477188]
13. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE transactions on bio-medical engineering. Jul; 2000 47(7):829–37. [PubMed: 10916253]
14. Ozdas, A.; Shiavi, RG.; Silverman, SE.; Silverman, MK.; Wilkes, DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. Vol. 9. Vanderbilt University, Department of Biomedical Informatics; Nashville, TN 87215 USA: 2004. asli.ozdas@vanderbilt.edu, Tech. Rep
15. Cohn, JF.; Kruez, TS.; Matthews, I.; Yang, Y.; Nguyen, MH.; Padilla, MT.; Zhou, F.; De la Torre, F. Detecting depression from facial actions and vocal prosody. 3rd International Conference on Affective Computing and Intelligent Interaction; sep 2009; p. 1-7.
16. Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: Features and normalization. Interspeech. 2011:2997–3000.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

17. Scherer, S.; Stratou, G.; Gratch, J.; Morency, L-P. Investigating voice quality as a speaker-independent indicator of depression and ptsd. Proceedings of Interspeech; 2013; 2013.
18. Trevino AC, Quatieri TF, Malyska N. Phonologically-based biomarkers for major depressive disorder. EURASIP Journal on Advances in Signal Processing. 2011; 2011(1):1–18.
19. Maddage, M.; Senaratne, R.; Low, L-S.; Lech, M.; Allen, N. Video-based detection of the clinical depression in adolescents. Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE; IEEE; 2009. p. 3723-3726.
20. Stratou, G.; Scherer, S.; Gratch, J.; Morency, L-P. Automatic non-verbal behavior indicators of depression and ptsd: Exploring gender differences. Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on; IEEE; 2013. p. 147-152.
21. Joshi, J.; Dhall, A.; Goecke, R.; Breakspear, M.; Parker, G. Neural-net classification for spatio-temporal descriptor based depression analysis. 21st International Conference on Pattern Recognition (ICPR); 2012.
22. Scherer, S.; Stratou, G.; Mahmoud, M.; Boberg, J.; Gratch, J.; Rizzo, A.; Morency, L-P. Automatic behavior descriptors for psychological disorder analysis. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on; IEEE; 2013. p. 1-8.
23. Joshi, J.; Goecke, R.; Breakspear, M.; Parker, G. Can body expressions contribute to automatic depression analysis?. International Conference on Automated Face and Gesture Recognition (FG); 2013.
24. Alghowinem, S.; Goecke, R.; Wagner, M.; Parkerx, G.; Break-spear, M. Head pose and movement analysis as an indicator of depression. Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII; 2013. p. 283-288.
25. Alghowinem, S.; Goecke, R.; Wagner, M.; Parker, G.; Breakspear, M. Eye movement analysis for depression detection. IEEE International Conference on Image Processing (ICIP); 2013. p. 4220-4224.
26. Cummins, N.; Joshi, J.; Dhall, A.; Sethu, V.; Goecke, R.; Epps, J. Diagnosis of depression by behavioural signals: a multimodal approach. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge - AVEC 13; ACM Press; 2013. p. 11-20.
27. Meng, H.; Huang, D.; Wang, H.; Yang, H.; AI-Shuraifi, M.; Wang, Y. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge; ACM; 2013. p. 21-30.
28. Scherer, S.; Stratou, G.; Morency, L-P. Audiovisual behavior descriptors for depression assessment. Proceedings of the 15th ACM on International conference on multimodal interaction; ACM; 2013. p. 135-140.
29. Joshi J, Goecke R, Alghowinem S, Dhall A, Wagner M, Epps J, Parker G, Breakspear M. Multimodal assistive technologies for depression diagnosis and monitoring. Journal on Multimodal User Interfaces. 2013
30. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. Proc FLAIRS-25. 2012:141–146.
31. Yang Y, Fairbairn CE, Cohn JF. Detecting depression severity from intra- and interpersonal vocal prosody. IEEE Trans on Affective Computing. 2013
32. Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; Pantic, M. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge; ACM; 2013. p. 3-10.
33. Inventory of Depressive Symptomatology (IDS) & Quick Inventory of Depressive Symptomatology (QIDS). [Online]. Available: http://www.ids-qids.org/index2.html
34. Saragih, JM.; Lucey, S.; Cohn, JF. Face alignment through subspace constrained mean-shifts. IEEE 12th International Conference on Computer Vision; Ieee; 2009. p. 1034-1041.
35. Zeng Z, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans on PAMI. 2007; 31(1):39–58.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

36. Chang, CC.; Lin, CJ. Libsvm: a library for svm. 2001. 2006-03-04[http:www.csic.ntu.edu.tw/rcjlin/papers/lib.svm

37. Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal fusion for multimedia analysis: a survey. Multimedia systems. 2010; 16(6):345–379.

38. Nolen-Hoeksema S. Sex differences in unipolar depression: Evidence and theory. Psychol. 1987; (101):259–282.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

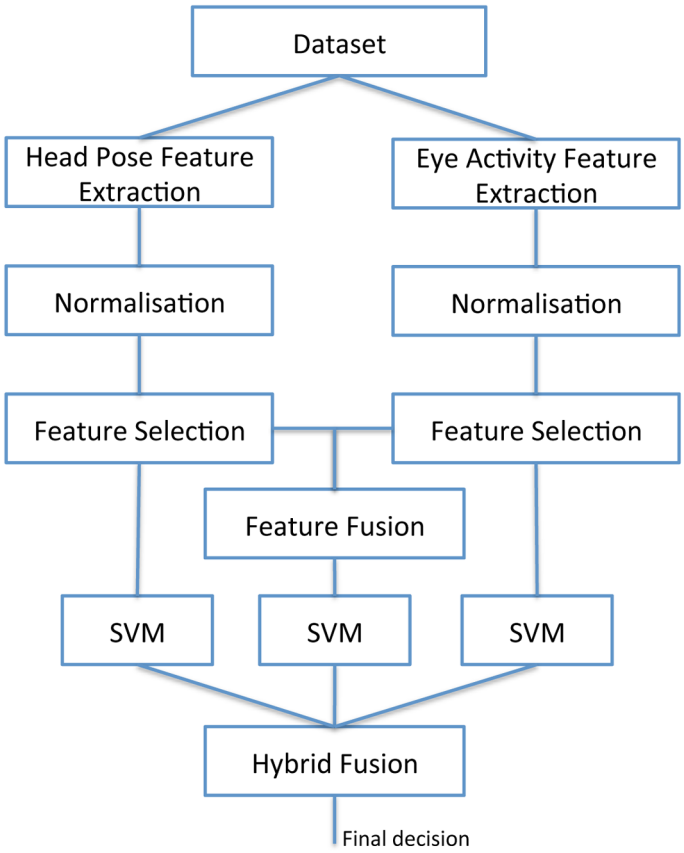

Fig. 1.
Approach to classify depressed from low-/non-depressed subjects

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

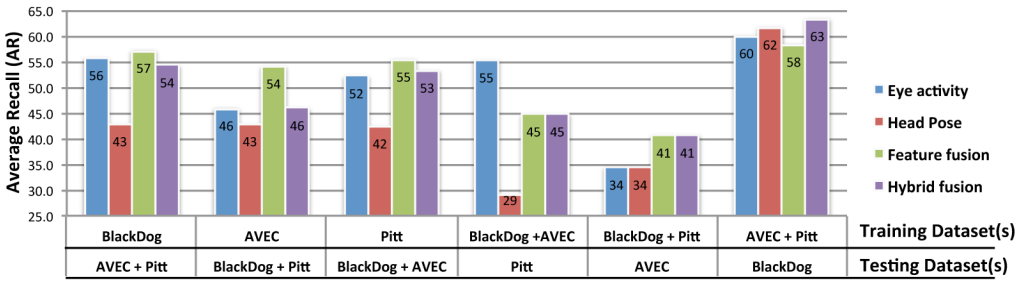

Fig. 2.

Classification results of combinations of datasets using the train-test method and fixed feature exceeding the t-statistic

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

Summary of the three datasets specification used in this research

Dataset	BlackDog	Pitt	AVEC
Language	English (Australian)	English (American)	German
Classification	Severely Depressed/Healthy Control	Severe/Low depression	Severe/Low depression
Number of subjects per class	30	19	16
Males-Females	30-30	14-24	9-23
Procedure	open ended questions interview	HRSD clinical interview	human-computer interaction experiment (story telling)
Symptom severity measure	QIDS-SR	HRSD	BDI
Mean score (range)	19 (14–26)	Severe:22.4 (17–35)/Low:2.9 (1–7)	Severe:35.9 (30–45)/Low:0.6 (0–3)
Equivalent QIDS-SR Score [33]	19 (14–26)	Severe:17 (13–26)/Low:2 (1–5)	Severe:20 (16–22)/Low:1 (0–2)
Total Duration (minutes)	509	355.9	33.2
Average duration/subject (in min)	8.4 (± 4.4)	9.4 (± 4.3)	1.0 (± 0.8)
Hardware	1 camera + 1 microphone	4 cameras + 2 microphones	1 web camera + 1 microphone
Audio sampling rate	44100 Hz	48000 Hz	44100 Hz
Video sampling rate	30 fps	30 fps	30 fps

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE II

Classification results (AR) in % and number of selected features (in parentheses) of individual datasets using leave-one-out cross-validation

Modality/Dataset	**BlackDog**		**AVEC**		**Pitt**	
Feature Selection	Variable set of features					
Eye Activity	80.0	(13)	81.3	(31)	92.1	(20)
Head Pose	73.3	(7)	68.8	(29)	86.8	(31)
Feature fusion	85.0	(20)	75.0	(60)	92.1	(51)
Hybrid fusion	85.0		75.0		92.1	
Feature Selection	Fixed set of features					
Eye Activity	73.3	(7)	71.9	(7)	89.5	(7)
Head Pose	61.7	(5)	71.9	(5)	84.2	(5)
Feature fusion	78.3	(12)	65.6	(12)	92.1	(12)
Hybrid fusion	76.7		68.8		94.7	

IEEE Int Conf Autom Face Gesture Recognit Workshops. Author manuscript; available in PMC 2016 July 21.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III

Classification results (AR) in % and number of selected features (in parentheses) of dataset combinations using leave-one-out cross-validation

Modality/Dataset	**BlackDog + AVEC**		**BlackDog + Pitt**		**AVEC + Pitt**		**All three datasets**	
Feature Selection	Variable set of features							
Eye Activity	63.0	(11)	78.6	(18)	61.4	(7)	68.5	(12)
Head Pose	62.0	(11)	63.3	(20)	65.7	(1)	51.5	(1)
Feature fusion	64.1	(22)	75.5	(38)	74.3	(8)	73.8	(13)
Hybrid fusion	64.1		76.5		72.9		71.5	
Feature Selection	Fixed set of features							
Eye Activity	69.6	(7)	73.5	(7)	78.6	(7)	70.8	(7)
Head Pose	67.4	(5)	68.4	(5)	61.4	(5)	66.2	(5)
Feature fusion	69.6	(12)	78.6	(12)	78.6	(12)	72.3	(12)
Hybrid fusion	69.6		79.6		85.7		73.1	

IEEE Int Conf Autom Face Gesture Recognit Workshops. Author manuscript; available in PMC 2016 July 21.