Cultural Factors in the Regression of Non-verbal Communication Perception

Tim Sheerman-Chase, Eng-Jon Ong and Richard Bowden
CVSSP, University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom
t.sheerman-chase,e.ong,r.bowden@surrey.ac.uk

Abstract

Recognition of non-verbal communication (NVC) is important for understanding human communication and designing user centric user interfaces. Cultural differences affect the expression and perception of NVC but no previous automatic system considers these cultural differences. Annotation data for the LILiR TwoTalk corpus, containing dyadic (two person) conversations, was gathered using Internet crowdsourcing, with a significant quantity collected from India, Kenya and the United Kingdom (UK). Many studies have investigated cultural differences based on human observations but this has not been addressed in the context of automatic emotion or NVC recognition. Perhaps not surprisingly, testing an automatic system on data that is not culturally representative of the training data is seen to result in low performance. We address this problem by training and testing our system on a specific culture to enable better modeling of the cultural differences in NVC perception. The system uses linear predictor tracking, with features generated based on distances between pairs of trackers. The annotations indicated the strength of the NVC which enables the use of ν-SVR to perform the regression.

1. INTRODUCTION

Spontaneous non-verbal communication (NVC) is key to human understanding of natural conversation and occurs multi-directionally and multi-modally. Expression of non-verbal signals is dependent on the cultural and social context, but previous research into automatic recognition of NVC has not considered these differences. Automatic recognition of NVC is important, as it enables more user centric computer interfaces. Recognition of non-verbal signals specific to a culture should make a system more sensitive to NVC in that context. Although NVC is usually a concurrent multi-directional interaction, we can assume communication can be decomposed into a flow of single communication actions, with each action having a sender and receiver. Cultural differences in sending and receiving NVC has been the subject of research for many decades but little work has addressed it in the context of automatic recognition. This paper focuses on automatic recognition while considering the cultural dependency of the receiver's perception of NVC. The sender's NVC was recorded in a single context to enable cross culture comparison of NVC perception.

Figure 1. Example frames from recorded conversations in the LILiR TwoTalk corpus. Clockwise from the top left are examples of agree, thinking, understand and question, based on UK annotator responses.

Although many emotion and human interaction data corpuses are available, few available data sets have an appropriate social, annotation and technical specification for the investigation of culturally specific perception of NVC. Particularly, the social context of many data sets is either extremely diverse or artificially contrived for the recording session. LILiR TwoTalk corpus [22] was selected for use because it occurs in a well defined and common social situation. The corpus comprises of 527 clips of casual dyadic conversation with minimal experimental constraints. This corpus was used for training and evaluating an automatic recognition system. As part of this study, annotation of the video clips was conducted by paid and volunteer Internet

workers from a broad range of cultures, with multiple annotators answering each question. This enables cross cultural comparison of perception of NVC, as well as training an automatic recognition system that specializes in specific cultures. Although many other cross cultural surveys have been conducted, this is the first cross cultural survey that may be used as the basis for an automatic NVC recognition system.

This annotation data is used for training and testing an automatic recognition system. The system uses linear predictor tracking [17], geometric feature extraction and support vector regression (SVR) to extract facial information, and to perform supervised learning and classification. The performance of the system is shown to be low if the training data is not culturally representative of the testing data. This is a significant finding, because all previous automatic NVC and emotion recognition systems have only considered a single culture. We address this problem by training and testing our system on a multiple specific cultures to enable better modeling of the cultural differences in NVC perception. This results in better regression performance. The main contributions of this paper are the multi-culture annotation of a naturalistic corpus suitable for automatic learning, an investigation of performance of automatic NVC recognition between multiple cultures, the use of regression in NVC recognition rather than classification. The video corpus and the annotation data collection is described in Section 2, an overview of the methodology is provided in Section 3, details of feature extraction and regression is discussed in Section 4 and results are presented in Section 5. These results are discussed in Section 6 with conclusions drawn in Section 7.

1.1. Background

This section provides an overview of the effect of context on NVC, the available data sets and the various annotation approaches that may be adopted. Expression and perception of NVC and emotion are dependent on a wide range of cultural and social factors [16]. Darwin commented that certain "gestures, which seem to us so natural that we might easily imagine that they were innate, apparently have been learned like the words of a language" [6] and described the significant cultural variations in approval and disapproval NVC. Differences in emotion expression due to culture can be detected by humans to predict a person's cultural origin [15]. Also, perception is likely to be culturally dependent, as Jack et al.[12] have discovered cultural differences in gaze patterns for people evaluating photos of facial expression. This , together with the discovery that people are more accurate in interpreting emotions from their own culture, has lead Elfenbein and Ambady [11] to argue for a dialect theory of emotion. This posits different culturally specific "rules" for both the display and perception of emotion, in contrast to Ekman's neurocultural theory that expression differences arise from display rules alone.

Annotation of emotion data is a time consuming activity and typically has low inter-annotator agreement due to annotator sensitivity to social and cultural factors [1]. Reidsma et al.[19] found agreement is higher if certain NVC signals are present. Human behavior is also affected by the participant's awareness of being recorded or being asked to act in a deliberate fashion. Cowie and McKeown [5] argued that changes in structural differences in communication, that is to say the way in which NVC signs are distributed, are more significant than deliberate vs. posed differences. Both humans [4] and machines [23] can discriminate between posed and spontaneous expressions. To make a system that is useful beyond laboratory conditions, training data of an automatic system needs to resemble spontaneous human interaction and be annotated in a particular social and cultural context.

Several databases exist that use spontaneous expressions, but are in a task-based social context (AMI Meeting Corpus [3], EmoTABOO [7]), training contexts ([2]) or collected from multiple sources from variety of social contexts (EmoTV Database [7]). Collecting video data with more constraints on the participants behavior or recording of deliberately posed expressions can reduce the experimental difficulty and quantity of video recording required. However, staging deliberate emotions, task based social contexts or structured interviews requires some experimenter intervention and may effect the genuineness and structure of the communication. Existing large data corpuses, such as the AMI meeting corpus, are not used in this work as the annotation of the desired NVC signals by multiple cultures would be unaffordable, as well as the social contexts being too broadly defined, not commonly occurring in everyday situations or requiring a large amount of experimenter intervention. Given a data set in a particular social context, training and evaluating an automatic NVC recognition system requires annotation of the video using an encoding system.

Papers that focus on emotional states have often encoded NVC into the set of culturally independent emotions described by Ekman [9] (fear, sadness, happiness, anger, disgust and surprise). El Kaliouby and Robinson [10] used categories intended to be more applicable to common human interactions (agreeing, concentrating, disagreeing, interested, thinking, unsure). Emotion and NVC recognition are usually treated as a multi class problem which avoids attempting to recognise the intensity of the expression, which would be important in most applications. Regression can be advantageous over the normal multi-class approach used by other researchers. Regression provides information on the intensity of emotion, which is often not addressed in other works.

The following section describes the video corpus and annotation used in this study.

2. VIDEO CORPUS AND ANNOTATION

To maximise applicability, it would be ideal to test and train on completely spontaneous and natural interactions. Unfortunately, due to ethical and technical restrictions, it is not usually possible to produce high quality recording of conversations with the participants being unaware of being recorded. Reactivity is the effect of a participant changing their behavior due to their knowledge of being recorded and presents a challenge to recording any naturally occurring social situation. A compromise is to minimise the instructions given to willing and informed participants while also satisfying the need for usable data.

The LILiR TwoTalk corpus attempts to minimise experimenter interference whilst recording usable data of spontaneous dyadic conversations. Eight participants of approximately equal social seniority were recorded in a laboratory environment in one of four conversation pairs. Each participant was asked to come to the lab, be seated across a table and converse for at least 12 minutes. A seated position reduces the amount of body and head pose changes and makes further analysis easier. No other instructions were provided to the participants (e.g. no limit on the topic of conversation). The conversation was recorded by two progressive scan PAL cameras at 25 fps, positioned behind and above the shoulder of each participant, and a single microphone placed on the table. The corpus contains 6 males and 2 females from various backgrounds, all of whom were English speakers (some native and some non-native). Fig. 1 shows typical frames taken from the video sequence. 527 clips were manually extracted from the videos which were thought to contain interesting NVC signals. The length of the clips ranged from length $l = 0.6$ to 10 seconds ($\bar{l} = 4.2s$, $\sigma = 2.5s$). The duration of video clips is similar to sections of conversation analysed by Lee and Beattie [14], who approached the issue of NVC annotation from a discourse analysis perspective (the duration of their samples was $\bar{l} = 4.8s$, $\sigma = 2.5s$). The corpus has NVC annotation categories of *thinking*, *understanding*, *agreeing* and *questioning* (see Table 1). These were selected due to their common occurrence in natural conversation. The process of annotating this data corpus is described in the next section.

2.1. Multi-cultural Annotation of NVC

There are various approaches to efficiently collect annotation data from multiple cultures. Given that social context guides judgments of NVC, collecting annotation data from observers in their own environment would be ideal. Unfortunately, annotating data is a tedious process, involving viewing of many video clips and recorded the responses. To collect sufficient data from geographically diverse locations, the use of paid and volunteer Internet workers (also known as "crowdsourcing") was selected as the solution. The company Crowdflower provides access to several worker pools including Amazon Mechanical Turk and Samasource. Each pool provides a distinct demographic of worker due to their regional popularity or mandate. Workers are paid a small amount of money for answering annotation questions. This can lead to workers responding with random data in an attempt to gain an easy financial reward. For this reason, the data must be filtered to remove at least a significant proportion of these untrusted workers. Random response data may cause problems with automatic learning as some methods have difficulty with noisy ground truth data. Workers with a low correlation with the global mode were excluded from the analysis.

The TwoTalk clips were annotated by the Mechanical Turk, Samasource worker and volunteer pools.[1] The annotators were not told what non-verbal signals were expected to be present in each clip. Annotation data was collected from 32 cultures and each worker's culture, or strictly speaking their location, was determined by the annotator's Internet protocol (IP) address. This may be an imprecise method; many countries contain more than one distinct culture, people travel or emigrate to other countries and IP addresses can be misleading due to the use of Internet proxies. Particularly, due to Samasource focusing on refugees, workers located in Kenya are likely to have originated from neighboring Sudan. For this reason, the cultural groupings used in this study, while probably culturally distinct, are likely to comprise of multiple cultures. However, it is sufficient for this work that we have culturally distinct groups, rather than each group correspond to a well defined and identifiable culture.

The annotation questions and videos were presented in the same language (English) to all participants. While translation of questionnaires is often done to enable cross cultural studies, accurate translation is only feasible if the language refers to well defined concepts. This is problematic in emotion research because the concepts of emotion used in any questionnaire are language/culture specific and do not have a precise translation [24] [20]. Also, the original videos were of a specific and unalterable language, and multiple language surveys do not lend themselves to currently available crowdsourcing methods for technical reasons. For these reasons, the option of translating the questions to the annotator's native language was rejected.

2.2. Filtering Untrusted Workers and Analysis of Responses

The web based annotation involved 711 participants, who collectively provided 79130 individual ratings across

[1] Annotation data and videos available at `http://www.ee.surrey.ac.uk/Projects/LILiR/twotalk_corpus/`

Table 1. Questions used in web based annotation of the LILiR TwoTalk corpus.

Question for Category	Minimum Rating	Maximum Rating
Does this person disagree or agree with what is being said? (A score of 5 is neutral or not applicable.)	Strong disagreement	Strong agreement
Is this person thinking hard?	No indication	In deep thought
Is this person asking a question?	No indication	Definitely asking question
Is this person indicating they understand what is being said to them?	No indication or N/A	Strongly indicating understanding

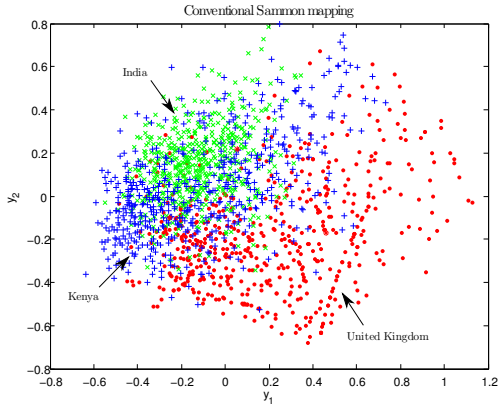

Figure 2. Sammon mapping of the filtered annotation responses. Each point represents the mean rating of a single clip within a single culture. The four NVC categories are concatenated into a 4D vector to enable distance pairs to be computed.

Table 2. Inter-Culture Correlation of Various Mean Filtered Culture Responses.

	India	Kenya	UK
India	1	0.56	0.55
Kenya		1	0.64
UK			1

Table 3. Number of annotators and votes for cultures included in the unfiltered and filtered data set.

Annotator Country	Num. Annotators		Num. Ratings	
	Unfilter	Filter	Unfilter	Filter
India	304	167	37147	22754
Kenya	196	195	15452	15420
UK	36	26	8211	8167

Table 4. Mean annotator correlation within various cultures with their respective culture mean or the global Mean (taken as the combined India, Kenya and UK ratings). Annotators correlate better on average with their own culture consensus than the global consensus.

	India	Kenya	UK
Own Culture Mean	0.67	0.78	0.77
Global Mean	0.64	0.74	0.68

all categories and clips. After each clip was viewed, each of the four NVC categories were independently rated by the participant on a discrete scale. Given the 527 clips in the corpus and 4 NVC categories, there are 2108 questions. Each question received one or more rating from each culture. All but one of the annotators rated a subset of all video clips, meaning the annotations are sparse. Some annotators were uncooperative and simply answered questions at random or attempt to circumvent any trivial questions intended to spot this behavior. These uncooperative annotators must be removed before the results can be used for supervised learning. Removing random noise from NVC ratings will tend to reduce the variance of annotations. Filtering is performed by comparing individual annotator ratings with the consensus score of annotators r' within the respective culture to find the annotator correlation coefficient z. It is assumed that all cooperating annotators will correlate, at least weakly, with some robust consensus scoring. The consensus score r' for each question is taken to be the mode ($r' = mode(R)$) of the individual ratings for that question ($R = r_1, r_2 ... r_m$) within a specific culture, since the mode is somewhat robust to random noise. m is the number of annotations for a single question. For each annotator, having annotated p questions, their scores $[r^1, r^2 ... r^p]$ are compared to the consensus score of their respective culture $[r'^1, r'^2 ... r'^p]$ to find a annotator correlation z. A correlation threshold of $\alpha = 0.2$ is used to form a set of cooperative annotator's ratings D from the data. An annotator's scores is in set D if $z > \alpha$. For each question, $D \subset R$. This would result in 94% of uncooperative annotators being removed, assuming each annotator answered 48 questions, this being the median of the actual number of annotator ratings. Some uncooperative annotators would correlate well enough to the culture consensus by chance and the resulting filtered data therefore still contains noisy survey responses. Although there may be patterns within the annotation data corresponding to culture, gender or other factors, these annotations will be broadly in agreement with the consensus scores and not be removed by this filter. Only cultures that provided ratings for every video clip were included in the filtered data set (India, Kenya and the UK). The number of annotators and question ratings after filtering are shown in Table 3. Because we are interested in cultural differences, the scores for individual annotators within a culture are meaned to form a culture average clip rating $d = \bar{D}$.

Since D has been filtered, the mean is preferred to the mode rating because sensitivity to annotator ratings is more important than robustness. The remainder of this paper only considers the filtered responses d, which are treated as continuous variables.

To determine if there are cultural differences in annotation, each culture's mean ratings for each video clip d was clustered using Sammon's mapping. This algorithm attempts to map a high dimensional space into a lower dimensional space while preserving distinct clusters. Each clip has four NVC category mean scores and these are concatenated to form a vector of length 4: $[d_{agree}, d_{question}, d_{think}, d_{understand}]$. The result of Sammon's mapping is shown in Figure 2. Each point represents the average ratings of a clip within a specific culture. If there were no significant differences between cultures, the figure would show each culture having the same distribution. However, it can be seen that each culture has a distinct distribution with contrasting densities in different regions. This shows that there are differences in each culture's annotator responses. The responses between different cultures do have some commonality, as there are areas of overlap. Also, we can see in Table 2 that the mean ratings of annotation for each culture is moderately correlated with each of the other cultures. This paper uses Pearson's correlation and average error as the means of comparing agreement, which is more natural to apply to continuous data. Conversely, Cronbach's alpha measure is often used on discrete data, particularly in cases with a limited number of scoring classes (e.g. yes/no) but is inappropriate for use in this paper because we use continuous variables. The cause of the observed annotation differences could be differences in cognition of NVC, different interpretations of the language used in the questions, differing familiarity and usage of the annotation equipment or some combination of these factors. UK and Kenya annotations seem to have a greater overlap than other pairs of cultures, based on Sammon's mapping and the correlation coefficients.

The common practice in automatic recognition of NVC is to ignore cultural differences. This assumes that each individual's perception of NVC may be approximated by a single consensus. The alternative is to group annotations into distinct clusters which better reflect the annotator's perception of NVC. As can be seen in Table 4, each annotator's ratings correlate better with the cluster mean consensus than with the global mean consensus. Based on the well documented cultural differences in NVC (described in Section 1.1) and the patterns in survey responses, we can say it is more appropriate to cluster annotation responses into distinct cultures, rather than taking the overall consensus. The following section provides an overview of the automatic NVC system.

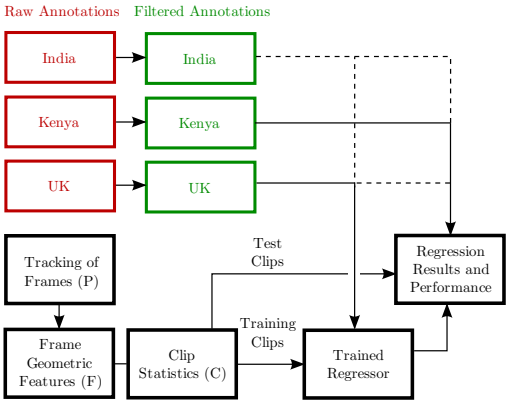

Figure 3. Overview of automatic system, showing filtering of annotations followed by training and testing on separate cultures.

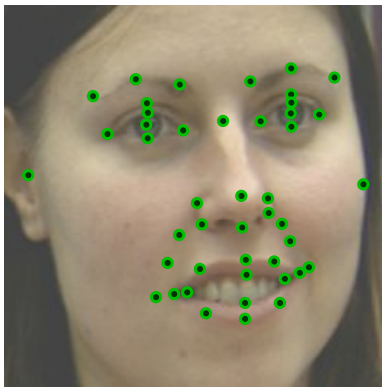
Figure 4. Facial feature positions used in tracking.

3. OVERVIEW OF AUTOMATIC REGRESSION OF NVC

The overview of the system is shown in Figure 3. During human communication, the relative positions and appearance of face regions vary over time. The optimal features needed for the recognition of a particular non-verbal signal cannot be determined a-priori but may be deduced by supervised learning based on the multi-annotator data set described in Section 2.1. In order to track facial features in a natural conversation, a tracker needs to be robust to large head pose variation while having the accuracy to track small changes due to expression. This is achieved by using a linear predictor flock tracking method that is robust to pose variation [17]. Trackers were placed on $J = 46$ salient features around the face (See Fig. 4). The choice of these features was constrained by the need to consistently mark examples of training data used for training the tracker. The trackers were initialised on the first frame of the sequence and used to predict the feature position P on all subsequent frames. The tracking occasionally suffered a complete failure due to occlusion or extreme head pose. This was overcome by manually reinitialising the tracker positions when-

ever the tracking failed.

Feature extraction was then performed on each frame to extract pose independent frame features. For each video clip, the frame features are processed to extract a clip feature that encodes feature positions and their temporal variations. This enables clips of different lengths to be compared. These clip features are taken with a culture's annotation d to train a regressor. The trained regressor can be then assessed by predicting NVC scores on unseen video clips and compared to human annotator responses.

4. FEATURE EXTRACTION AND REGRESSION

Feature extraction is necessary to transform low level features, such as tracker positions or pixel intensities, into a set that encodes the relevant information while being robust to irrelevant variations, such as pose, lighting changes and identity. For machine learning based on human faces, there are two approaches to feature extraction: shape and appearance. Since NVC signals are primarily sent by shape and motion of facial features, we will focus on this class of features (although some types of NVC are certainly appearance based, e.g. blushing).

Many previous studies have examined manually engineered features, often based on facial action coding system (FACS), such as used by el Kaliouby and Robinson [10] to detect complex emotions. It is possible that better features may be found by considering a much wider range of feature generation rules than can be manually engineered. For a given set of trackers, it is an intractable problem to test every possible feature that may be extracted eg. distances between trackers, angles between trackers, ratios between distances, etc. The system uses a single class of geometric features (distances between a pair of trackers) and exhaustively computes the frame features $\mathbf{F}$ for every possible pair of trackers. To remove the effect of different face shapes, each feature was zero centred and whitened on a per subject basis. For J trackers, each frame as a feature vector $\mathbf{F}$, the size of which is the triangular number T_J (which is the number of unique distance pairs).

4.1. Clip Feature Extraction

Each clip contains the frame features from multiple video frames and these are combined to provide a single clip feature vector. The relevant NVC information is probably present in only a subset of the frame features and may also be limited to a subset of frames. Ideally, clip features would encode relevant temporal information of the important frame features. Only a simple approach is used here, which takes the mean and variance of each feature frame to produce a clip feature $\mathbf{C}$ (in a similar fashion to [18]) $\mathbf{C} \in \mathbb{R}^{2T_J}$. For a clip that extends from frame a to b, the clip features are generated as follows:

Table 5. Correlation of performance of the automatic system when training and testing on the same or different cultures.

Test Culture	Train Culture India	Kenya	UK
India	0.29	0.27	0.24
Kenya	0.27	0.35	0.33
UK	0.25	0.33	0.32

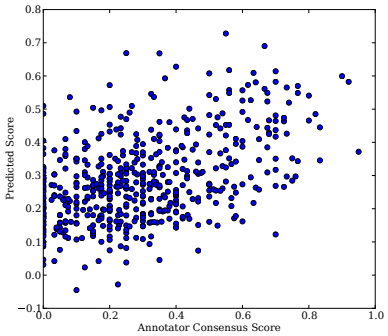

Figure 5. Scatter plot of actual and predicted thinking NVC intensities for the UK culture.

$$\mathbf{C}_i = \frac{1}{b-a}\sum_{f=a}^{b}\mathbf{F}_i^f, i \in [1...T_J] \quad (1)$$

$$\mathbf{C}_{i+T_J} = \frac{1}{b-a}\sum_{f=a}^{b}(\mathbf{F}_i^f - \mathbf{C}_i)^2 \quad (2)$$

4.2. Support Vector Regression

Support Vector Regression SVR is a supervised learning technique that takes a problem that cannot be solved by linear regression in the input space, and learns a non-linear mapping into a higher dimensional space in which the problem is suitable for linear regression [8]. In this system, the ν-SVR variant is used [21] with a radial basis function (RBF). SVR is an effective regressor for emotion recognition [13], and it is expected it may be effective in the broader area of NVC detection.

5. RESULTS

Testing was performed using 8 fold cross validation, with each cross validation test set corresponding to the clips for one of the eight recorded subjects. The ν-SVR learning parameters set as $C = 1.0, \nu = 0.5$ were found to be most effective. The training set consisted of the other seven subjects, making the testing person independent. Tables 6 and 7 shows the agreement of regression of each NVC category for the three culture groups, with the training and testing annotations taken from the same culture. Figure 5 shows a scatter plot with individual predicted and annotator consensus "thinking" scores for the UK culture. Example predic-

Table 6. Correlation of automatic system for training and testing on a single culture. Figure 5 shows the individual ratings for UK thinking NVC. The error limits are one standard deviation of the individual cross validation correlation results.

Culture	Agree	Question	Thinking	Understand
India	0.38±0.11	0.20±0.14	0.33±0.13	0.23±0.13
Kenya	0.43±0.15	0.15±0.19	0.39±0.20	0.43±0.17
UK	0.27±0.08	0.16±0.21	0.46±0.20	0.37±0.13

Table 7. Average prediction error of automatic system for training and testing on a single culture. Annotations scores have been normalised to the range 0 to 1.

Culture	Agree	Question	Thinking	Understand
India	0.09	0.11	0.11	0.11
Kenya	0.13	0.15	0.14	0.14
UK	0.10	0.18	0.16	0.21

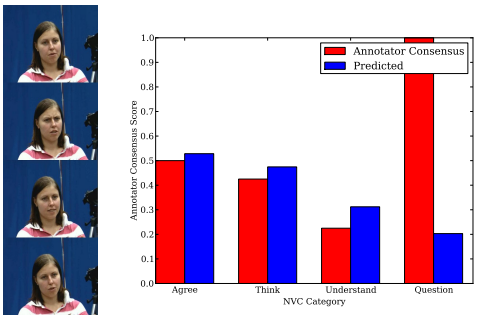

Figure 6. Example frames, annotator ratings and predicted scores for corpus clip "3dcfiL5Per".

tions of two video clips are shown in Figures 6. The performance for training and testing on the global consensus score is: agree 0.45±0.15, thinking: 0.45±0.19, understand 0.50±0.17 and question 0.19±0.17 (the global consensus is taken as the mean of the India, Kenya and UK ratings). Table 5 shows the performance of testing on one culture's mean annotation and testing on either the same or a different culture's mean annotation. The four NVC categories have been averaged to produce an overall performance measure.

6. DISCUSSION

The level of agreement between predicted and annotation scores were measured using two methods: correlation (Table 6) and average error (Table 7). Correlation is not sensitive to scaling and translation of the data but is more sensitive to outliers than the average error. As can be seen in these tables, "question" and "understand" are the NVC categories with the overall lowest agreement with the culture consensus, while "agree" and "thinking" are this most in agreement. In the case of questioning, this may be due to the NVC having a significant verbal element and not strictly non-verbal. Also, the performance of even the higher scoring NVC correlations is lower compared to the average human agreement with the consensus (Table 4) and may be due to the features not fully encoding the relevant information to perform accurate regression. Different cultures have lower or higher regression performances. This may be caused by different cultural perceptions of NVC or some methodological difference, such as varying familiarity with the language of the annotation questions and the levels of uncooperative annotators. If there is a cultural difference in the perception of NVC, it is likely that different groups of users are using different cues to make their decisions. These different sets of cues may correspond to features that are encoded by geometric features to a greater or lesser extent, leading to different levels of performance. And due to potential mislabeling of annotations for particular locations and different cultural diversities of countries, it is hard to draw firm conclusions by comparing different culture performances. The performance based on the mean global consensus is numerically higher than shown in Table 6 but this increase in performance is at the cost of the annotation ground truth labels not reflecting actual human responses as closely as the culturally specific consensus scores, as shown in Table 4.

The four NVC categories were averaged to form the diagonal for Table 5. The off diagonal values correspond to training and testing different cultures. The table shows that training and testing on a single culture produces the best performance. Therefore the system operates optimally when trained to work on a specific culture and predictions are less in agreement when compared with other cultures. This highlights the need for training data for automatic NVC recognition systems to be trained on culturally specific data that is appropriate for the intended application. Training on Kenya scores and testing on UK scores has a relatively small drop in performance and this might be due to their relatively similar culture consensus scores as seen in Table 2.

7. CONCLUSION

Expression and perception of non-verbal communication is sensitive to a variety of cultural and social factors, al-

though this has not been addressed in the context of automatic emotion or NVC recognition. The multi-culture annotations were collected using Internet crowdsourced data and show cultural differences in the annotation patterns, along with some general patterns of agreement. This dataset was specifically used because it was naturalistic and suitable for training an automatic NVC recognition system. It was found that testing an automatic system on data that is not culturally representative of the training data is seen to result in low performance. We address this problem by training and testing our system on a specific culture to enable better modeling of the cultural differences in NVC perception. Although differences in NVC perception are addressed, future work is needed to address the significant differences in expression of NVC, as well as more sophisticated temporal features that are more suitable to NVC recognition.

8. ACKNOWLEDGMENTS

The work has been supported by the EPSRC project LILiR (EP/E027946/1) and the European Community Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231135 DictaSign.

References

[1] S. Abrilian, L. Devillers, and J. Martin. Annotation of Emotions in Real-Life Video Interviews: Variability between Coders. In *5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006.

[2] S. Afzal and P. Robinson. Natural Affect Data - Collection & Annotation in a Learning Context. In *Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[3] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[4] J. Cohn and K. Schmidt. The Timing of Facial Motion in Posed and Spontaneous Smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1 – 12, March 2004.

[5] R. Cowie, G. McKeown, and C. Gibney. The challenges of dealing with distributed signs of emotion: theory and empirical evidence. In *Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[6] C. Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 3rd edition, 2002.

[7] L. Devillers and J.-C. Martin. Coding Emotional Events in Audiovisual Corpora. In *Proc. of the 6th Int. Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

[8] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, pages 155–161, 1997. MIT Press.

[9] P. Ekman. Basic Emotions. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*. Wiley, Chichester, UK, 1999.

[10] R. el Kaliouby and P. Robinson. *Real-time vision for HCI*, chapter Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures, pages 181–200. Springer, 2005.

[11] H. A. Elfenbein and N. Ambady. Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science*, 12(5):159–164, 2003.

[12] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara. Cultural confusions show that facial expressions are not universal. *Current Biology*, 19(18):1543 – 1548, 2009.

[13] I. Kanluan, M. Grimm, and K. Kroschel. Audio-visual emotion recognition using an emotion space concept. In *16th European Signal Processing Conference*, Lausanne, Switzerland, August 2008.

[14] V. Lee and G. Beattie. The rhetorical organization of verbal and nonverbal behavior in emotion talk. *Semiotica*, 120(1/2):39–92, 1998.

[15] A. A. Marsh, H. A. Elfenbein, and N. Ambady. Nonverbal "accents": Cultural differences in facial expressions of emotion. *Psychological Science*, 14(4):373–376, 2003.

[16] D. Matsumoto. *The SAGE Handbook of Nonverbal Communication*, chapter Culture and Nonverbal Behavior, pages 219–236. Sage Publications, Inc, 2006.

[17] E. Ong, Y. Lan, B. Thobald, R. Harvey, and R. Bowden. Robust Facial Feature Tracking using Multiscale Biased Linear Predictors. In *International Conference on Computer Vision*, 2009.

[18] S. Petridis and M. Pantic. Audiovisual Laughter Detection Based on Temporal Features. In *Proc. of the 10th Int. Conf. on Multimodal Interfaces*, pages 37–44, New York, NY, USA, 2008. ACM.

[19] D. Reidsma, D. Heylen, and H. J. A. op den Akker. On the Contextual Analysis of Agreement Scores. In J.-C. Martin, P. Paggio, M. Kipp, and D. Heylen, editors, *Proc. of the LREC Workshop on Multimodal Corpora*, pages 52–55. ELRA, May 2008.

[20] J. A. Russell. Culture and the categorization of emotions. *Psychological Bulletin*, 110(3):426 – 450, 1991.

[21] B. Schlkopf, A. Smola, R. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[22] T. Sheerman-Chase, E.-J. Ong, and R. Bowden. Feature Selection of Facial Displays for Detection of Non Verbal Communication in Natural Conversation. In *IEEE International Workshop on Human-Computer Interaction*, Kyoto, Oct 2009.

[23] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170, New York, NY, USA, 2006. ACM.

[24] A. Wierzbicka. *Emotions across languages and cultures: diversity and universals*, chapter Why words matter, pages 24–30. Cambridge University Press, 1999.