Feature Selection of Facial Displays for Detection of Non Verbal Communication in Natural Conversation

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 human communication has previously focused on deliberately acted emotions or in structured or artificial social contexts. This makes the result hard to apply to realistic social situations. This paper describes the recording of spontaneous human communication in a specific and common social situation: conversation between two people. The clips are then annotated by multiple observers to reduce individual variations in interpretation of social signals. Temporal and static features are generated from tracking using heuristic and algorithmic methods. Optimal features for classifying examples of spontaneous communication signals are then extracted by AdaBoost. The performance of the boosted classifier is comparable to human performance for some communication signals, even on this challenging and realistic data set.

Figure 1. Example frames from natural conversation video sequences.

1. Introduction

This paper investigates which visual non-verbal signals may be used in recognition of human communication and to learn classifiers that can automatically recognise certain types of signals. Spontaneous human communication is a multi-modal and multi-directional transfer of information of which non-verbal communication is often a key factor required for understanding. Automatic understanding of human communication is a requirement for human centric computer interfaces that allow humans to interact with computers using intuitive communication. Many previous approaches have attempted to simplify the problem using deliberately posed expressions or by constraining the social situation. A potential problem with these studies is that verbal and non-verbal communication signals are social context dependent. A system trained in an artificial social situation may not transfer to other unseen situations. This paper describes an attempt to address these challenges by initially creating a data set of natural conversation which is then annotated by multiple observers. The participants are selected to have similar social relationship so the interaction is spontaneous but weakly controlled to make each conversation comparable. Multiple observers perform the annotation to reduce the effect of individual variations in interpreting social signals. Non-verbal communication signals may involve the position and motion of all parts of the body. The optimal features for detection of a specific communication need to be determined. This paper uses feature selection from both heuristic and algorithmic methods. Temporal features are also generated and included in the feature selection to determine if these can be applied to recognition.

Feature selection from training data shows specific features are consistently selected for various types of human communication. Classification performance is consistent with the difficulty of the data set, particularly the wide range of expressions that are present in natural conversation.

1.1. Background

Some expressions, described by Ekman [4], are culturally independent (fear, sadness, happiness, anger, disgust

and surprise). A great deal of automatic recognition research has been based on these categories. Although these expressions are quickly recognised by humans [6], they only represent a subset of possible human expressions and they appear relatively infrequently in natural conversation. El Kaliouby and Robinson [5] used categories intended to be more applicable to common human interactions (agreeing, concentrating, disagreeing, interested, thinking, unsure). Another branch of emotion detection research uses a low level representation of expression, such as Facial Action Coding System (FACS) [4, 8]. FACS is an expression encoding scheme based on the facial muscles that produce them. In principle, any facial expression can be described using a combination of facial action units (AUs) but for the FACS representation to be used in many applications, cultural information must be used to interpret the facial actions into a useful form.

Cowie [2] highlighted the difficulty of creating non-verbal communication databases, particularly the failure to generalise from prototypical expressions to new expressions, the effect of context and multi-modality, as well as the enormity of the task of annotation. Annotation data typically has low inter-annotator agreement due to annotator sensitivity to social and cultural factors. Reidsma *et al*. [12] found agreement is higher in certain social circumstances. Previous research has reduced the experimental difficulty and quantity of video by using deliberately acted expressions, such as those proposed by Ekman. Systems trained on deliberate expressions do not necessarily generalise to spontaneous expressions. Both humans and machines can discriminate between posed and spontaneous expressions [2]. Several databases exist that use spontaneous expressions, but are in a task-based social context (AMI Meeting Corpus [1], EmoTABOO [3]) or collected from multiple sources from variety of social contexts (EmoTV Database [3]). Staged task based social contexts or structured interviews require some experimenter intervention and therefore not necessarily natural conversation. Examples of work that address emotion recognition in spontaneous videos are Petridis and Pantic's work [11] on multi-modal discrimination between speech and laughter, and Zeng *et al*. [14] who detected positive and negative emotion in structured interviews.

The novel contribution of this paper is training on and detection of natural human communication in a specific and realistic situation. The extraction and selection from both heuristic and algorithmic features is also new, as well as the use of head pose to detect complex human communication cues. Automatic classification of non-verbal signals shows potential despite the complexity of realistic conversation. Automatic classification is compared to human performance of examples of clear strong communication signals and found to be similar for some communication signals. The recording of natural conversation is described in Section 2. The annotation of clips extracted from the video sequences is described in Section 3. A proposed automatic recognition system of human communication using visual information is described in Section 4 with performance evaluation described in Section 5. Conclusions are drawn in Section 6.

2. Data Capture of Social Interaction

To maximise applicability, it would be ideal to test and train on completely spontaneous and natural interactions. Stubbs defines natural language as occurring "*without any intervention from the linguist*" and "*is spontaneous in the sense of unplanned, and which is composed in real time in response to immediate situational demands*" [13]. Since it is considered ethical to ask for permission before recording a participant and considering the potential reactivity (the change in behaviour due to the awareness of being recorded) of participants in recording sessions, it may be impossible to record truly natural data without resorting to subterfuge. Even if secret recording is conducted, high quality video cameras and lighting are difficult to conceal. If the social situation is completely uncontrolled, an individual may turn away from the camera. A compromise is to minimise the instructions given to willing and informed participants while also satisfying the need for usable data.

A single social situation was chosen as a focus for this study: two people engaged in conversation. To record high quality video and sound, a media laboratory was the selected venue. Eight participants were recorded in one of four conversation pairs, which enables individual variations to be studied. The participants were selected by the experimenter which enabled the selection of participants of equal social seniority. 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 standard definition PAL cameras at 25 fps, positioned behind the shoulder of each participant, and a single microphone placed on the table.

The database contains 6 males and 2 females from various backgrounds, all of whom were English speakers (some native and some non-native). Figure 1 shows typical frames taken from the video sequence.

3. Multi-observer Annotation

The interpretation of non-verbal communication signals is dependent on their social context. The inherent subjectivity of signals makes the annotation of natural conversation by a single annotator very sensitive to person specific

Question for Category	Minimum Score	Maximum Score
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

Table 1. Questions used in web based annotation of video.

factors [12] (e.g. feelings to persons in a social situation, cultural norms, attitudes to the topic under discussion, the annotator's mood at the time of annotation, etc.) To overcome this, multiple annotators can review video sequences and a wider consensus is created. This consensus would then be specific to the cultural background of the annotators and less susceptible to individual variation.

In contrast to deliberately acted videos, natural conversation is relatively sparse in non-verbal communication actions. Annotation of large amounts of video data is a time consuming process. To enable efficient use of annotator effort, periods containing potential non-verbal signals are prioritised over periods which do not seem to contain any non-verbal content. Speech with constant gaze and body pose, as well as passive listening, were considered having the lowest non-verbal content. The disadvantage of splitting a longer video into clips is that the clip boundaries need to be determined within a stream of continuous human interactions. This potentially leads to loss of relevant contextual information or the inclusion of multiple or conflicting non-verbal communication signals. The categories used in this paper are thinking, understanding, agreeing and questioning due to their common occurrence in natural conversation (see Table 1). Although users were asked to rate disagreement, this was excluded from the analysis due to the rarity of examples. The length of the clips ranged from length $l = 0.6$ to 10 seconds ($\bar{l} = 4.2s$, $\sigma = 2.5s$).

The recorded conversations, described in Section 2, were manually divided into clips by a single observer which were thought to contain examples of the selected categories. The number of clips extracted were: 109 of agreement or disagreement, 140 of understanding, 93 of thinking and 65 of questioning. An additional 120 clips where selected at random to increase the variety of non-verbal signals (total 527 clips). The annotators were not told what non-verbal signals were expected to be present in each clip. The user reviews each clip and assigns their rating on an 11 point scale. [1]

3.1. Annotation Data Collected

The web based annotation had 21 participants, who rated clips presented in a random order and completed 94 clips on average. After each clip was viewed, all four categories were rated by the participant. The category "thinking" had 1981 annotator clip ratings, with an average of 3.8 ratings per clip. The other categories had a similar level of participation. Figure 2 shows a histogram of average ratings from the annotation. The questioning category has a sharper peak and is probably due to the relatively clear distinction between questioning and not questioning in human communication. The categories of thinking and understanding show an approximately even distribution throughout the possible range of ratings, indicating weak, strong and intermediate expression of these signals are relatively common.

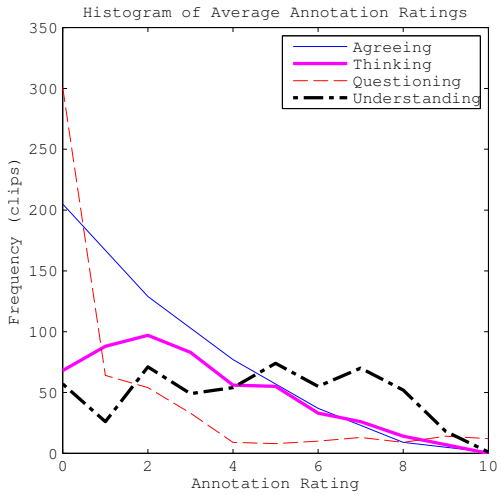

Figure 2. Histogram of Average Rating based on multi-annotators. Blue thin line is agreeing, magenta thick line is thinking, black dot-dashed line is understanding and red dashed line is questioning. Zero is a neutral score and 10 is strongly showing the communication signal. Disagreement ratings have been omitted.

To investigate the dependence of different non-verbal signals, the correlation of average scores for two categories was calculated. As can be seen in Table 2, there are weak correlations between most of categories. This implies that some of the categories are dependent. A correlation coefficient tending to 1 indicates that both communication signals occur together consistently. A correlation that tends to -1 indicates that the expression of one signal is present, the other is consistently absent. A correlation of zero indi-

[1] Demographic data for annotators and participants: http://personal.ee.surrey.ac.uk/Personal/T.Sheerman-chase/nvc_demographic.php

	Agreeing	Understanding	Thinking	Questioning
Agreeing	1			
Understanding	0.46	1		
Thinking	-0.21	-0.23	1	
Questioning	-0.18	-0.40	0.06	1

Table 2. Correlation coefficients of the average ratings for each category.

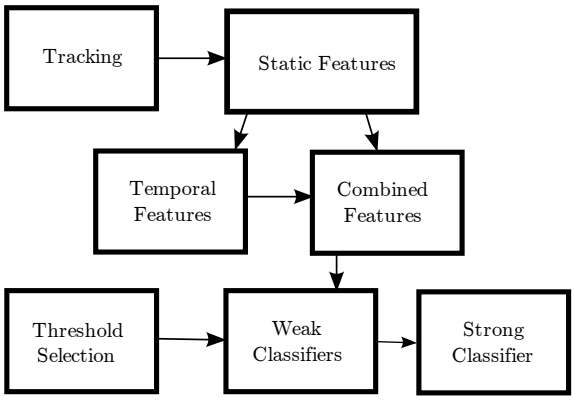

Figure 3. Overview of automatic recognition system.

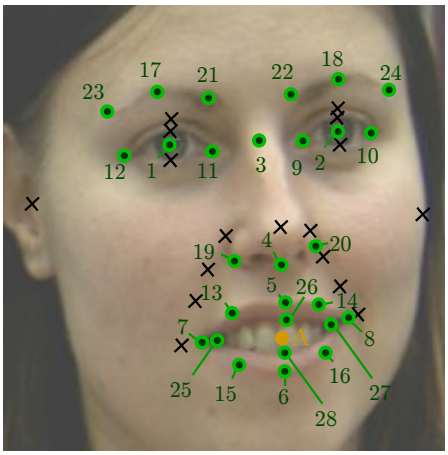

Figure 4. Facial feature positions used in tracking. Circular orange and green numbered points are trackers used in geometric feature generation. All points, including black crosses, are used for generation of the other static feature types.

cates the communication signals occur independently. The results are consistent with everyday experience of certain combinations of signals commonly occurring together (e.g. understanding and agreement) while others rarely do (e.g. understanding and questioning).

Using these annotated clips of spontaneous conversation, we can investigate automatic recognition of non-verbal communication in more natural conditions than have previously been possible.

4. Automatic Recognition of Non-verbal Communication

Humans use a wide array of information to support spontaneous communication. The approach described in this paper focuses on visual facial information but inclusion of body position and audio features would be an obvious extension. During human communication, the relative positions and appearance of the parts of the face vary in time. The optimal features needed for the recognition of a particular non-verbal signal cannot be determined a-priori but many be deduced by supervised learning based on the multi-annotator data set described in Section 3. Humans can accurately recognise some expressions almost instantaneously while others require some time to be recognised [6]. This implies that temporal information may be necessary for automatic recognition of some types of non-verbal signals. In order to track facial features in a natural conversation, a tracker needs to be robust to large head pose variation while having to accurately track small changes due to expression. This is achieved by using a linear predictor flock tracking method that is robust to pose variation [10]. Trackers were placed on $J = 46$ salient features around the face (See Figure 4). The choice of these features was constrained by the need to consistently mark examples of training data used for offline training of the tracker. The trackers were initialised on the first frame of the sequence and used to predict the feature position 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 whenever the tracking failed.

The remainder of this section outlines a system for selection of features and automatic recognition of non-verbal signals (see Figure 3). One obstacle to recognition is motion due to head pose tends to swamping subtle expression changes. Static features are required in order to separate motion due to expression from head pose. There are many possible ways of extracting features from tracking data. To investigate which features are optimal, a few different approaches were implemented (see Section 4.1). Static features only consider information from a single frame of video based on tracking of facial feature positions. Temporal features are generated to incorporate information on the variation of the static features over time (Section 4.2). Feature selection and classification is performed by AdaBoost [7] (4.3).

4.1. Static Features

The methods for extracting static features are: tracking PCA values (Section 4.1.1), geometric features (4.1.2), Levenberg-Marquardt (LM) head pose estimation (4.1.3) and affine head pose estimation (4.1.4).

4.1.1 Tracking PCA Values

Tracking contains facial feature motion caused by both head pose and expression changes. Information needs to be extracted from tracking that separates motion due to expression from that due to head pose changes. Principal component analysis is used in a similar way to Lien *et al*. [8] to create suitable static facial features. Unlike Lien *et al*., use of all the PCA coefficients can effectively allow feature selection to determine to what extent the pose or the expression is important for classification. Tracking of J features for all c frames to form a $2J \times c$ matrix $\mathbf{C}$. Performing PCA on $\mathbf{C}$ produces $2J$ principle component vectors. The tracking for each frame was then projected into the eigenspace to give a $2J$ vector in eigenspace that represent the deformation. These eigenvalues were taken as the static features for that frame.

4.1.2 Geometric Features

Geometric features attempt to encode information about expression or head pose using distances and angles of a subset of trackers. This provides a simple way to extract static features that are independent of pose. The geometric features may be designed to focus on specific trackers that are potentially relevant. This approach was used by el Kaliouby and Robinson [5] to detect complex emotions. Their features for head yaw, head pitch, head roll, eyebrow raise, lip pull/pucker and lips part were implemented for this paper as static features. Two of their mouth features were omitted as they were appearance based. Additional features were added to investigate if gaze was useful in recognition of non-verbal signals (see Table 3 and Figure 4). The total number of geometric features was $g = 12$. Position A is the average position of the outer mouth trackers.

Head yaw	$\frac{\overline{P_9P_{10}}}{\overline{P_{11}P_{12}}}$
Head pitch	$P_4[t] - P_4[t-1]$
Head roll	$\angle P_9P_{11}$
Eyebrow raise	$\frac{(\overline{P_{11}P_{21}}+\overline{P_1P_{17}}+\overline{P_{12}P_{23}})_t}{(\overline{P_{11}P_{21}}+\overline{P_1P_{17}}+\overline{P_{12}P_{23}})_0}$
Lip pull/pucker	$\frac{(\overline{AP_7}+\overline{AP_8})_t-(\overline{AP_7}+\overline{AP_8})_0}{(\overline{AP_7}+\overline{AP_8})_0}$
Lips part	$\overline{P_{26}P_{28}P_{25}P_{27}}$
Right eye horizontal	$\frac{P_{11}P_{12}\cdot P_{11}P_1}{\lvert P_{11}P_{12}\rvert}$
Left eye horizontal	$\frac{P_9P_{10}\cdot P_9P_2}{\lvert P_9P_{10}\rvert}$
Mean eye horizontal	$\frac{(P_{11}P_{12}\cdot P_{11}P_1)(P_9P_{10}\cdot P_9P_2)}{2\lvert P_{11}P_{12}\rvert\lvert P_9P_{10}\rvert}$
Right eye vertical	$\frac{\lvert P_{11}P_{12}\times P_{11}P_1\rvert}{\lvert P_{11}P_{12}\rvert}$
Left eye vertical	$\frac{\lvert P_{11}P_{12}\times P_{11}P_1\rvert}{\lvert P_{11}P_{12}\rvert}$
Mean eye vertical	$\frac{\lvert P_{11}P_{12}\times P_{11}P_1\rvert\lvert P_{11}P_{12}\times P_{11}P_1\rvert}{2\lvert P_{11}P_{12}\rvert\lvert P_{11}P_{12}\rvert}$

Table 3. Geometric features used to extract expression that is robust to pose.

4.1.3 LM Head Pose Estimation

LM minimization can be used to determine pose from a cloud $\mathbf{a}$ of J points when point correspondence to mesh positions $\mathbf{p}$ are known [9]. This static feature does not consider expression of the face. Each tracker is associated with a node in a generic head model. The pose is estimated by minimizing the cost function:

$$\mathcal{F}(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{J} ||\mathbf{a}_i - proj(\mathbf{R}\mathbf{p}_i + \mathbf{t})||^2 \qquad (1)$$

where $proj()$ is the projection function, $\mathbf{R}$ is the current rotation matrix and $\mathbf{t}$ is the head translation. The resulting $\mathbf{R}$ and $\mathbf{t}$ matrices represent the estimated head pose (where $\mathbf{R}$ is the rotation matrix corresponding to the Euler angles $\{R_{pitch}, R_{roll}, R_{yaw}\}$ and $\mathbf{t}$ is the head translation $\{t_x, t_y, t_z\}$). Perspective geometry is used for the projection function. The estimated pose matrices $\mathbf{R}$ and $\mathbf{t}$ are concatenated to form a static feature of size 6.

4.1.4 Affine Head Pose Estimation

Affine head pose estimates a transformation from the tracking positions of current video frame to a frame showing a frontal view of the face. For head poses near frontal view, the change in feature positions due to pose can be approximated by an affine transform. Therefore, the affine transform roughly encodes the head pose information. This method is computationally simple compared to LM minimisation as it is the pseudo-inverse of a single $3 \times J$ matrix. The affine matrix is reordered to form a static feature of size 6.

The static features from each of the above sources are concatenated to form the static features on a single frame (of size $z = (2J + g + 6 + 6)$) which is used as the basis for temporal feature generation.

4.2. Temporal Features

To create temporal features, each static feature is considered independently and within a temporally sliding window (See Figure 5). A polynomial equation is fitted to this temporal window of static features, as previously described by Petridis and Pantic [11]. The polynomial parameters are determined by regression and then used as temporal features that describe the evolution of the value of a static feature. A quadratic equation is used as this was found to be most

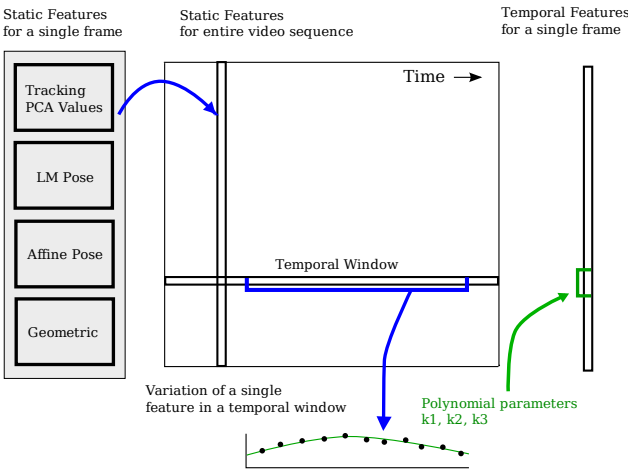

Figure 5. Temporal features are created by a quadratic curve fitted to a single static feature component within a temporal window.

effective in laughter detection by Petridis and Pantic:

$$f_k = f_{k1}t^2 + f_{k2}t + f_{k3}, k \in [1...K] \quad (2)$$

where $k1$, $k2$ and $k3$ are the quadratic parameters to be used as temporal features. As temporal features are calculated for every static feature, the resultant temporal feature size is $3z$. Because the optimal temporal window size is unknown, b sliding window sizes are used to create temporal features at different scales. The multi scale temporal features and the static features are concatenated to form a combined feature vector of size $(3b + 1)z$.

4.3. Feature Selection and Recognition using AdaBoost

Given positive and negative examples from human annotation and data from weak classifiers, AdaBoost is a popular machine learning method that incrementally adds a weighted weak classifier to a set that forms a final strong classifier. This approach provides both a classifier that may be tested on unseen examples and feature selection of relevant temporal and static features. The optimal weak binary classifier threshold is unknown, so AdaBoost operates on multiple weak classifiers for each feature, each having a different threshold. The thresholds are chosen based on the mean and standard deviation values of the features within the positive and negative classes. The features and weights are then combined to form a strong classifier for non-verbal communication signals.

4.3.1 Choice of Weak Classifier Thresholds

Binary classification is used as a weak classifier but the optimal thresholds need to be determined. Multiple weak classifiers with varying thresholds are used to classify each component from the combined feature vector. Sensible placement of thresholds avoids having an excessive number of weak classifiers. The thresholds are placed in relation to the average value for the positive and negative classes. The positive and negative examples are strongly rated examples from human annotation data. The lowest and highest thresholds are determined as follows:

$$T_0 = E_{min} - \frac{S_p + S_n}{2} \quad (3)$$

$$T_V = E_{max} + \frac{S_p + S_n}{2} \quad (4)$$

where T_0 is the lower threshold, T_V is there upper threshold, E_{min} is the class mean with the lower value, E_{max} is the class mean with the higher value, S_p and S_n are the standard deviations of the positive and negative classes. An additional $V - 1$ thresholds T_v are equally spaced between these upper and lower threshold:

$$T_v = \frac{v}{V}(E_{max} - E_{min} + S_p + S_n) + T_0 \quad (5)$$

where $v \in 0..V$. This ensures the thresholds are distributed throughout the observed range of values. The number of weak classifier results on which boosting is performed is therefore $(V + 1)((3b + 1)z)$.

5. Results

The automatic feature selection and classification system was tested on the four categories that were previously annotated by multiple human observers. Classification occurs on a frame by frame basis for each frame in an example clip. Only strongly rated examples from the human annotation were used in training and testing. The 25 highest and 25 lowest rated sequences were taken as the positive and negative categories for thinking, understanding and questioning. For agreeing, the 25 negative examples were selected as the best neutral score (at a survey rating of 5). The clips were split equally into two sets for cross validation testing. The temporal window lengths were 80ms, 160ms, 320ms and 640ms ($b = 4$). Four additional thresholds were added in the weak threshold selection ($V = 5$). The AdaBoost process was iterated until convergence of the testing performance.

5.1. Recognition Using Performance

The performance on unseen test data of the classifier is shown in Figure 6. The area under the ROC curve for agreeing is 0.70, thinking 0.81, understanding 0.80 and questioning 0.73. Human performance was also estimated by taking each rating for a specific clip and comparing that to the average of the other user ratings on that clip. Only strongly rated examples were used in the analysis. The ROC of human performance is shown in Figure 7 (area under curve:

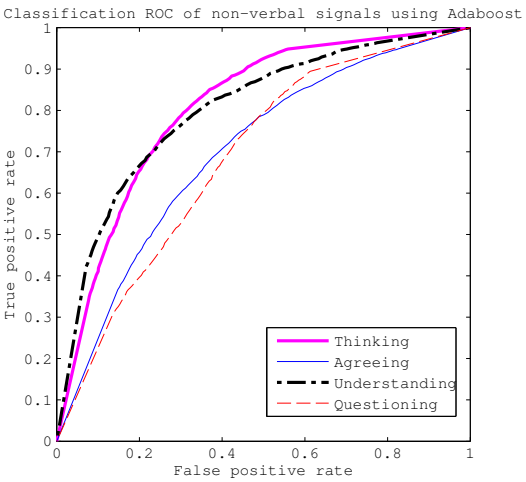

Figure 6. Classification performance for four categories. Blue thin line is agreeing, magenta thick line is thinking, black dot-dashed line is understanding and red dashed line is questioning.

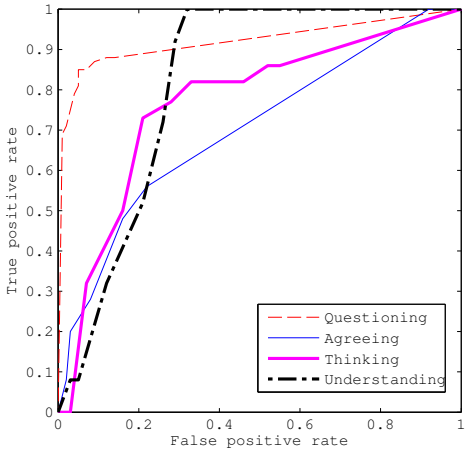

Figure 7. Human performance in classification of clips containing strongly rated examples of communication.

agreeing 0.70, thinking 0.77, understanding 0.82 and questioning 0.92).

5.2. Features Selected for Recognition

AdaBoost creates a strong classifier from a linear combination of weak classifiers. The highest weighed weak classifiers are the more significant features for discriminating between classes. The selected features for each communication category are shown in Table 4.

The feature selection always selects the constant quadratic parameter of the temporal features in the 10 highest weighted classifiers. Features were selected that utilised all the available temporal window sizes (80ms to 640ms).

5.3. Discussion

The classification performance of each natural communication category is promising and is comparable to human performance for certain communication signals. This is significant because the data is spontaneous and contains a wide range of expressions. For example, the negative data for agreeing contains a wide range of other non-agreeing emotions. Also, it is possible that there are multiple styles of human signalling agreement and that some non-verbal signals are person specific. No confusion matrix is shown in the analysis as the categories used are not independent (see Table 2) and in natural conversation, multiple expressions may occur simultaneously. This prevents any example from being assigned to any single class of expression. Each clip is also extracted from a longer video sequence. The selection of start and end times for each clip is a subjective process and tends to lead to multiple communication signals being present which is effectively noise in training and testing data.

AdaBoost feature selection uses a different set of features for each category of communication. This implies that specialised sets of features are required for recognition of each non-verbal signal. For all categories of communication, some features were consistently selected in the two training data sets (such as eyebrow raise in questioning). For some non-verbal signals, geometric features are selected as the best. These features are manually designed to be relevant for the task. It is likely that additional specialised features for each non-verbal signal would improve recognition performance. Feature selection almost exclusively selected the constant term in temporal features. Considering various lengths of temporal window were among the features, this corresponds to a static feature with different levels of smoothing applied. The absence of features using the linear or quadratic terms implies temporal features are less important than static features. Low order tracking PCA and LM pose features were often selected with a high weight. These features correspond to head rotations and translations. This suggests head pose is important in recognition for some non-verbal signals.

6. Conclusion and Future Work

This paper describes the collection of multi-annotator clips of natural conversation and investigates which visual features are useful in recognition of communication signals. These features are combined to form a strong classifier with shows some potential, consistent with human performance and challenging data. Previous research has used data that used deliberately posed examples of expression or a single annotator assigning clips into well defined classes. The approach in this paper is a step towards recognition of complex communication signals in a more realistic and chal-

Weak Classifier	Feature Selected on 1st Half	Interpretation	Feature Selected on 2nd Half	Interpretation
1st Agreeing	Tracking PCA 5th EV	Lip pull/pucker	Geometric feature 5	Lip pull/pucker
2nd Agreeing	Tracking PCA 1st EV	Horiz. head tr.	Tracking PCA 1st EV	Horiz. head tr.
3rd Agreeing	Tracking PCA 21st EV	*	Tracking PCA 30th EV	*
4th Agreeing	Geometric feature 3	Head roll	Geometric feature 1	Head yaw
1st Thinking	Tracking PCA 5th EV	Lip pull/pucker	Tracking PCA 11th EV	*
2nd Thinking	LM Pose 5th Component	Head pitch	Tracking PCA 2nd EV	Vert. head tr.
3rd Thinking	Tracking PCA 9th EV	*	Tracking PCA 5th EV	Lip pull/pucker
4th Thinking	Tracking PCA 10th EV	*	Tracking PCA 60th EV	*
1st Understanding	Tracking PCA 5th EV	Lip pull/pucker	Tracking PCA 5th EV	Lip pull/pucker
2nd Understanding	Tracking PCA 10th EV	*	Affine pose 5th component	*
3rd Understanding	Tracking PCA 15th EV	*	Tracking PCA 6th EV	*
4th Understanding	Tracking PCA 16th EV	*	Tracking PCA 3rd EV	Head roll
1st Questioning	Geometric feature 4	Eyebrow raise	Geometric feature 4	Eyebrow raise
2nd Questioning	Tracking PCA 24th EV	*	LM Pose 5th Component	Head Translation
3rd Questioning	Tracking PCA 74th EV	*	Tracking PCA 7th EV	*
4th Questioning	Tracking PCA 8th EV	Mouth open/close	Tracking PCA 8th EV	Mouth open/close

Table 4. The first 4 features selected by AdaBoost to classify various categories of communication in descending order of weight. The interpretation of the feature is possible for those targetting a particular head transformation and for low order PCA modes of variation. Interpretation labels are based on manual subjective labelling of the corresponding transformation. Features marked * are complex deformations which are difficult to interpret as simple expression changes. This is common in the higher order PCA variation modes. "Eigenvalue" has been abbreviated as EV.

lenging context. Only strongly rated examples as selected by human annotators are used in these experiments. Further investigation will be performed on weakly expressed communication. Additional features, including appearance based features to detect wrinkles may be beneficial to certain types of communication recognition. Multi-modal features (e.g. audio, speech recognition) may also be required to accurately recognise certain communication events.

7. Acknowledgements

This work has been largely supported by the EPSRC project LILiR and in part by the FP7 project DICTASIGN.

References

[1] 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.

[2] R. Cowie. Building the databases needed to understand rich, spontaneous human behaviour. In *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

[3] 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).

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

[5] 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.

[6] R. el Kaliouby, P. Robinson, and S. Keates. Temporal context and the recognition of emotion from facial expression. In *Proc. of HCI Int. Conf. 2003*, June 2003.

[7] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proc. of the 13th Int. Conf.*, pages 148–156, 1996.

[8] J. Lien, T. Kanade, J. Cohn, and C.-C. Li. Automated facial expression recognition based on FACS action units. In *Proc. 3rd IEEE Int.l Conf. on Automatic Face and Gesture Recognition*, pages 390–395, Apr 1998.

[9] Z. Liu and Z. Zhang. Robust head motion computation by taking advantage of physical properties. In *Workshop on Human Motion*, pages 73–, 2000.

[10] E.-J. Ong and R. Bowden. Robust lip-tracking using rigid flocks of selected linear predictors. In *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

[11] 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.

[12] 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, ELRA, May 2008. AMIDA publication number 99.

[13] M. Stubbs. *Discourse Analysis*. Basil Blackwell, Oxford, 1983.

[14] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006.