Improving Recognition and Identification of Facial Areas Involved in Non-verbal Communication by Feature Selection

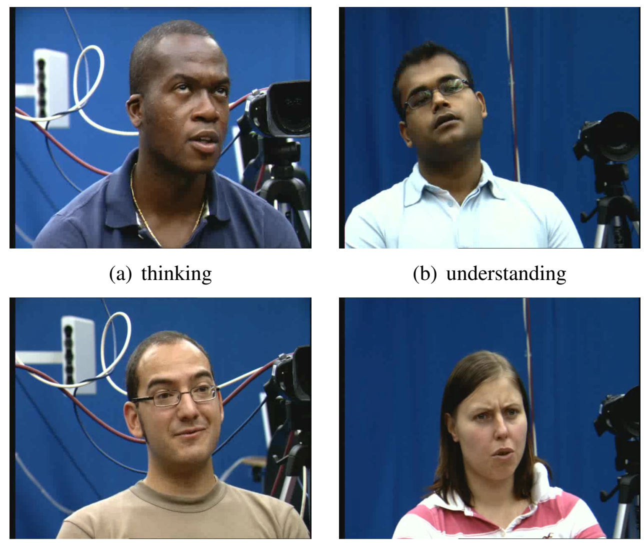
Tim Sheerman-Chase, Eng-Jon Ong, Nicolas Pugeault and Richard Bowden
CVSSP, University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom
Email: t.sheerman-chase,e.ong,n.pugeault,r.bowden@surrey.ac.uk

***Abstract*—Meaningful Non-Verbal Communication (NVC) signals can be recognised by facial deformations based on video tracking. However, the geometric features previously used contain a significant amount of redundant or irrelevant information. A feature selection method is described for selecting a subset of features that improves performance and allows for the identification and visualisation of facial areas involved in NVC. The feature selection is based on a sequential backward elimination of features to find a effective subset of components. This results in a significant improvement in recognition performance, as well as providing evidence that brow lowering is involved in questioning sentences. The improvement in performance is a step towards a more practical automatic system and the facial areas identified provide some insight into human behaviour.**

I. Introduction

Non-verbal communication signals are essential to understanding in almost all common social situations. They consist in an ensemble of wordless cues, both visual and audible, that convey information about the meaning expressed. Automatic systems are beginning to address the recognition of Non-Verbal Communication (NVC) and emotion [1]. However, the difficulty to choose, detect and track accurately facial features often leads to the generation of features that contain irrelevant or redundant information, which may compromise the performance of system. A feature selection approach can address this problem, leading to both improved performance and allowing to identify the facial areas used in the communication or emotion act [2]. Furthermore, understanding which facial areas are useful for automatic recognition may provide insight into human perception and behaviour. This paper will propose a novel feature selection approach for automatic NVC recognition based on sequential backward selection of facial shape features [3]. Moreover, a novel method for the visualization of relevant facial areas is described.

For the evaluation of our method, we selected the TwoTalk corpus [4] because it features spontaneous human NVC. The corpus comprises of manually selected clips of casual dyadic conversation with minimal experimental constraints (see Figure 1). The annotation of the video clips was conducted by paid and volunteer Internet workers from three distinct cultures. Specifically, humans NVC during natural conversation were manually annotated for the following categories: *thinking*, *understanding*, *agreeing* and *questioning*—see Figure 1 for an illustration. This corpus was used for training and evaluating an automatic recognition system.



(a) thinking (b) understanding (c) agreeing (d) questioning

Fig. 1. Some example frames taken from the TwoTalk corpus, captured from clips that were strongly labelled (according to British annotators) as, respectively: (a) thinking, (b) understanding, (c) agreeing and (d) questioning. Note that NVCs are dynamics and therefore the actual clip features convey more information than still images.

The proposed system is based on the system proposed by Sheerman-Chase et al. [5] in which facial shape features were based on geometric relations between tracked facial points. The system uses linear predictor tracking [6] to track a selected set of facial locations, and makes use of geometric relations between points to encode facial shape information. Feature selection is then used to select the subset of feature components that are relevant to a specific NVC. Because the annotation in the TwoTalk corpus is gathered from three distinct cultural groups, feature selection is separately computed for each culture and for each NVC category. After feature selection,

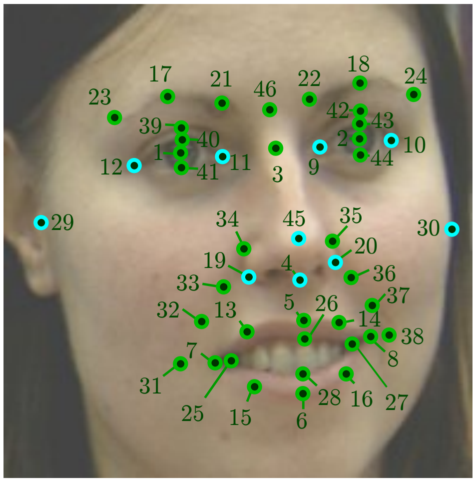


Fig. 2. Points on the face were tracked to encode the face position. The points were manual assigned to the flexible or rigid set. Flexible points are shown in green. Rigid points are shown in cyan. Humans have relatively little ability to move these rigid facial points relative to the skull.

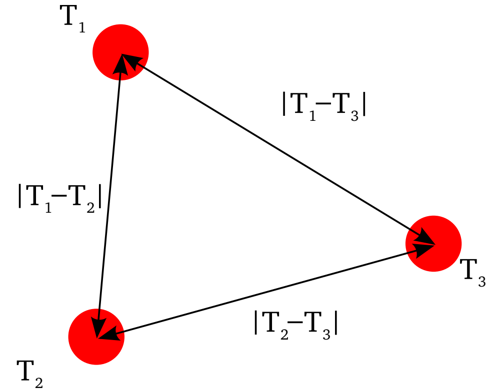


Fig. 3. Geometric features were generated, based on distances between pairs of trackers, that encode local deformation information.

the contribution of each feature component is also evaluated, resulting in a set of feature relevance weights for each NVC signal. These feature weights can be visualised to show the involvement of facial areas in the expression of NVC in an intuitive manner. This is based on segmenting a face using Voronoi tessellation around the position of trackers. Voronoi tessellation segments an image into cells based around seed positions; each point in the space is assigned to a cell based on the nearest seed position. This visualisation can either be used to check if the relevant facial areas conform to our expectation, or provide an indication as to which areas are used by the automatic system. This in turn may provide clues as to human NVC perception, although facial areas used by human perception may differ from those used by an automatic approach.

The resulting feature component subsets are shown to be more effective than the original feature vector. Moreover, the visualisation of NVC-selected facial features yields interesting insights in NVC perception: for example, the visualisation of thinking NVC confirms out expectation that it is related to gaze aversion. Also, questioning NVC appears to be related to brow movements, which is an association that is little reported outside of the sign language community.

The next section provides an overview of relevant existing research. Section II reviews the existing research. The dataset is described in Section III. Section IV describes the methodology used for tracking and feature selection. Section V contains experimental results and discussion. Conclusions are drawn in Section VII.

II. Background

There are many generic approaches to feature selection (see [7] for a review), which vary in performance, computational cost and restrictions on the type of input data. A technique can be either an embedded, filter or wrapper method. Embedded feature selection methods, such as a boosting classifier, can be used to weight a set of features based on relevance or isolate a suitable subset of components. This subset can then be used by a second, more sophisticated classifier. This approach was used by Valstar [8] to select shape features by Gentleboost, and Petridis and Pantic [9] used Adaboost to select relevant audio and visual features. However, performing feature selection in this way assumes that the optimal set of features for both methods is similar—which is not necessarily the case. Yang et al. [2] propose a feature selection method based on rough set theory on audio visual features. This avoids the discretisation of feature values required by some classifiers, such as Adaboost, and the associated loss of information.

Filter based feature selection appears to have been largely avoided in the context of emotion and NVC recognition, probably due to the relatively small number of feature components in the original feature vector (usually thousands of feature components at most) and the often significant importance of feature interaction for emotion and NVC.

Wrapper based methods include randomised feature selection approaches such as simulated annealing and genetic approaches, but these have not been popular in facial analysis. Deterministic wrapper based approaches have been applied to emotion recognition: Grimm [10] used Sequential Forward Selection (SFS) to isolate relevant audio features. This method begins with an empty set and incrementally adds features that produce the greatest performance increase, in a greedy fashion. An alternative, called Sequential Backward Elimination (SBE), is to start with a full set of features and incrementally eliminate features that result in the best performance [3]. The SBE approach was used by el Kaliouby and Robinson [11] to find the most relevant geometric features. The method described in this paper is of this type.

There are several existing papers that identify which features have been selected for emotion or NVC recognition, but it is less common to attempt to visualise which features have been selected. If features are shown, they are often visualised individually (e.g. [2]), which can make comprehension of the overall distribution difficult. In experimental psychology, gaze patterns in perception have been visualised by Jack et al. [12].

In a similar way, the present work provides a data-driven visualisation of the relative importance of facial features for NVC recognition.

This study describes a visualisation that is as intuitive to interpret as a density map of visual attention and is somewhat comparable to Jack et al. .

III. Dataset Description

This paper makes use of the LILiR TwoTalk dataset [4]. The 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). 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 dataset contains a range of spontaneous emotions, lip movements, head pose changes and occasional hand gestures that occasionally occlude the face. The colour images are converted to grey-scale using the ITU-R 601-2 luma transform.

The corpus has NVC annotation categories of *thinking*, *understanding*, *agreeing* and *questioning*. These were selected due to their common occurrence in natural conversation. The annotators were based in three cultural groups by their IP address. The three cultures that received a significant quantity of annotation were Great Britain (GBR), India (IND) and Kenya (KEN).

IV. Feature Extraction and Feature Selection

A. Tracking and Feature Generation

Features were generated by tracking a set of hand-picked facial locations over time, and the facial shape was encoded by calculating the distance between any two pairs of these points. Tracking was performed by linear predictor tracking [6]. Because the tracker requires multiple frames to be annotated for training, $\kappa = 48$ points $\{T_i\}_{i \in [1..\kappa]}$ that could be consistently located were selected for use and manually marked (see Figure 2). The system uses a single class of geometric features (distances between a pair of trackers) and exhaustively computes the frame features

$$\mathbf{F} = \{\|T_i - T_j\|\}_{i=[1..\kappa], j>i} \tag{1}$$

for every possible pair of trackers, in a similar way to Valstar et al. [8] (see Figure 3). To remove the effect of different face shapes, each feature was zero centred and whitened on a per subject basis. Therefore, for κ trackers, each frame is described by feature vector $\mathbf{F}$, the size of which is given by the triangular number $J = \frac{\kappa(\kappa+1)}{2}$ (which is the number of unique distance pairs between κ points). These features are not robust to scale changes but subsets of feature components are robust to head rotation, specifically in cases where the head rotation does not change the apparent distance of facial points.

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 likely to be present in only a subset of the frames and features. Ideally, clip features would encode relevant temporal information of the important frame features. 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 [9]) $\mathbf{C} \in \mathbb{R}^J$. For a clip that extends from frame a to b, the clip features are generated as follows:

$$\mathbf{C}_i = \frac{1}{b-a} \sum_{f=a}^{b} \mathbf{F}_i^f, i \in [1..J] \tag{2}$$

$$\mathbf{C}_{i+J} = \frac{1}{b-a} \sum_{f=a}^{b} (\mathbf{F}_i^f - \mathbf{C}_i)^2 \tag{3}$$

The training dataset is composed of M clip features and corresponding annotations $\mathbf{S} = \{(C_k, y_k)\}_{k=1..M}$

B. Feature Selection Methodology

This section describes the method in detail and the resulting performance impact. The approach used is a greedy SBE of the features [3]. A backward search (SBE) beings with a set containing every feature component and sequentially removes components from this set to maximise performance. Forward search involves beginning with an empty set and sequentially adding feature components to the set, again to maximise performance. Backward searching was thought to be preferable to forward searching because features interactions can be found and exploited. Forward search, particularly in the first few iterations, adds features without the benefit of other complementary features. In contrast, a backward search allows irrelevant features to be eliminated while retaining features that contain complementary information.

In this work, we apply feature selection within a person independent, cross validation framework. There are eight folds in cross validation, resulting in eight different partitioning of seen and unseen data sets. Feature selection is applied to the seen data of a specific cross validation fold, to determine a relevant feature subset. Support Vector Regression (SVR) is then applied to the feature subset to produce a model suitable for prediction.

The procedure for SBE is shown in Algorithm 1. The search begins with a current set $\boldsymbol{\alpha} = \{1...2|\mathbf{F}|\}$ which includes all feature components. The components to be removed from $\boldsymbol{\alpha}$ at each iteration is then determined. The current set $\boldsymbol{\alpha}$ is then updated and the process continues until the current set $\boldsymbol{\alpha}$ is empty. For the large number of components, it is too time

Algorithm 1 Algorithm **SelectFeature**, performing a single step of the feature selection algorithm. The regressor can be any suitable method, but in this study ν-SVR is used.

Require: A feature set $\boldsymbol{\alpha} \neq \emptyset$,
  a dataset $\mathbf{S} = \{(F_k, y_k)\}_{k\in[1..M]} = \bigcup_{j=1}^{N} \mathbf{s}_j$,
  an elimination rate $\eta > 0$, and a fitting function **fit**().
Ensure: A reduced feature set $\tilde{\boldsymbol{\alpha}} \subset \boldsymbol{\alpha}$.
  for $i \in \boldsymbol{\alpha}$ **do** {Assess all features in turn}
    $\boldsymbol{\beta} = \boldsymbol{\alpha} \setminus i$
    for $j \in [1..N]$ **do** {Cross validation performance}
      Regression on fold $\mathbf{s}_j$ using features $\boldsymbol{\beta}$: $\phi = \mathbf{fit}(\mathbf{s}_j, \boldsymbol{\beta})$
      $p_{i,j} = \mathbf{corr}(\phi(F_k), y_k), (F_k, y_k) \in \mathbf{S} \setminus \mathbf{s}_j$
    end for
    $\mathbf{p}_i = \frac{1}{N}\sum_j p_{i,j}$ {Total error across all folds}
  end for
  $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$
  for $1..\eta$ **do** {removes the η worst features}
    $i^* = \arg\max_{i\in\tilde{\boldsymbol{\alpha}}} \mathbf{p}_i$
    $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}} \setminus i^*$
  end for

consuming to remove components at a rate of 1 per iteration. To accelerate the process, multiple feature components are removed nearer the start of the SBE process. As the number of components in the current set approaches zero, the rate of feature elimination returns to the standard 1 feature component per iteration. This produces a significant speed increase, but risks the removal of non-optimal components and this may result in a sub-optimal final feature set. The number of feature components removed from the current feature set at each iteration is denoted η. This depends on the number of feature components ω in the current set $\boldsymbol{\alpha}$ as follows:

$$\eta = \begin{cases} 200 & \text{if } \omega > 1000 \\ 100 & \text{if } 400 < \omega \leq 1000 \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

These thresholds were based on an intuitive expectation that only a small subset of features are required for accurate recognition.

To find an appropriate subset of features for removal from the current feature set, the contribution of each feature component needs to be assessed. An overview of this process is shown in Algorithm 1. Each feature component in the current feature set $\boldsymbol{\alpha}$ is selected as the test component and the performance impact of the removal of the component is evaluated. The features are then prioritised, with the feature components resulting in the lowest performance preferred for removal. This process becomes progressively faster as the current feature set becomes smaller.

The training data $\mathbf{S}$ is split into N cross validation folds $\{\mathbf{s}_j\}_{j\in[1..N]}$, such that $\mathbf{S} = \bigcup_{j=1}^{N} \mathbf{s}_j$. These "feature selection" folds are distinct from the "system" cross validation folds discussed earlier, so that each fold contains data from multiple human subjects.

Algorithm 2 Algorithm **FindBestFeatureSet**, calling Algorithm 1 iteratively to perform a greedy search of the best performing subset of features on the training data.

Require: A feature set $\boldsymbol{\alpha} \neq \emptyset$,
  a dataset $\mathbf{S} = \{(F_k, y_k)\}_{k\in[1..M]} = \bigcup_{j=1}^{N} \mathbf{s}_j$,
  an elimination rate $\eta > 0$, and a fitting function **fit**().
Ensure: Selects best subset of features α^{seen}.
  $r = 0$
  Initializes to the full feature set: $\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}$
  while $\boldsymbol{\alpha}^r \neq \emptyset$ **do**
    $r = r + 1$
    Call Algorithm 1: $\boldsymbol{\alpha}^r = \mathbf{SelectFeature}(\boldsymbol{\alpha}^{r-1}, \mathbf{S}, \eta)$
    for $j \in [1..N]$ **do** {Cross validation performance}
      Regression on $\mathbf{s}_j$ using $\boldsymbol{\alpha}^r$: $\phi = \mathbf{fit}(\mathbf{s}_j, \boldsymbol{\alpha}^r)$
      $p_{i,j} = \mathbf{corr}(\phi(F_k), y_k), (F_k, y_k) \in \mathbf{S} \setminus \mathbf{s}_j$
    end for
    $\mathbf{p}_r^{\text{seen}} = \frac{1}{N}\sum_j p_{r,j}$ {Total error across all folds}
  end while
  Select peak performance: $\boldsymbol{\alpha}^{\text{seen}} = \arg\max_r \mathbf{p}_r^{\text{seen}}$

Algorithm 2 calls Algorithm 1 iteratively, producing a series of sets $\{\boldsymbol{\alpha}^r\}_r$ that correspond to each stage in the progressive removal of features, and can be assessed separately on unseen data. The expectation is for performance to increase as poor features are removed. As the SBE process is nearing termination, some features that are critical to NVC regression are removed and the performance sharply declines. The performance $\mathbf{p}_r^{\text{seen}}$ of the feature subset $\boldsymbol{\alpha}^r$ at each stage is evaluated and retained for later analysis.

Because this process results in multiple sets which are used to create multiple NVC models, it is not obvious which feature set to use and how many feature components are optimal. Simply selecting the peak performance when evaluating feature sets on unseen data violates the separation of seen and unseen data. For simplicity, this method uses the feature set $\boldsymbol{\alpha}^{\text{unseen}}$, having the peak performance for unseen test data to determine the number of feature components to be used. It is likely that different NVC signals require a specific set of geometric features to be effective. Therefore, feature selection is computed for a specific NVC category and using a specific culture's annotation data. The processing of test set $\boldsymbol{\beta}$ has been parallelised in this implementation, resulting in a speed increase.

C. 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 [13]. In this system, the ν-SVR variant is used [14] with a Radial Basis Function (RBF) kernel. SVR has been shown to be an effective regressor for emotion recognition [15], and it is therefore expected to be effective in the broader area of NVC detection.

V. Experimental Results

A typical plot of performance against the number of feature components in the subset is shown in Figure 4. As expected, the performance of predicting unseen test data increases at first as features are removed, until performance suffers a sharp decline. The far left starting point of the lower curve corresponds to the performance of the system discussed in the previous chapter (i.e. without feature selection). In this example, feature selection results in a significant increase in performance. Feature sets containing between 10 to 275 features deliver the highest performance, with the peak performance requiring only 10 features. However, it is unlikely that this feature set will be effective for regressing NVCs other than *thinking*. This can be a disadvantage because the SBE method is NVC category specific and a great deal of computation is required to retrain the system for a different NVC signal.

The feature selection curves are relatively linear until approximately 400 features remain. This corresponds to the threshold in Equation 4. The change in the curve behaviour at this point suggests that a different set of thresholds might result in a higher peak, although this was not investigated.

This pattern is repeated for most other NVC categories and in different cultures. While almost every test fold subject benefits from the feature selection process, not all system cross validation folds yield the same level of performance increase. The left plot of Figure 4 shows an instance in which feature selection was not effective. The performance is low before feature selection begins, which might indicate a problem with the approach in recognising this subject performing *question* NVC signals. The centre and right curves show typical feature selection behaviour in a different cultures and NVCs. A typical gradual improvement in performance can be seen, as features are removed before a sharp decline.

The optimal number of features is not known before feature selection begins. The peak of unseen performance is 10 features, while the peak for seen performance is at approximately 125 features (see Figure 4). A simple approach to determine the optimal number of features is to use the peak performance of unseen data. The performance for this method is shown in Table I. However, this method violates the separation of training and unseen test data. The table also shows the performance with the ideal termination of feature selection. This table implies that if terminated at an appropriate point, SBE can result in a significant performance gain.

The number of features for termination of the feature selection process should be determined based on seen training data. This restriction represents a system which is less reliant on manual tuning of parameters. The peak training data performance can be used to determine when to terminate the feature selection process. This is likely to select a non-optimal number of features, but this approach respects seen and unseen data separation. The results may be compared to the regression system in Sheerman-Chase et al. [5]. The performance of this method is shown in the highlighted column of Table I. Feature selection produces a large increase in performance over the existing method. Therefore, feature selection is beneficial for geometric features because it removes irrelevant features and results in a feature subset that is more suited for the specific NVC. The next section describes the visualisation of these component subsets.

TABLE I
Comparison of various approaches of termination of the feature selection process, along with the performance without feature selection from Sheerman-Chase et al. . [5]

Area	NVC Category	Terminate By Unseen Peak	Terminate By Seen Peak	Without Feature Selection[5]
GBR	Agree	0.588	0.523	0.340
GBR	Question	0.453	0.385	0.188
GBR	Thinking	0.617	0.556	0.440
GBR	Understand	0.640	0.605	0.389
IND	Agree	0.637	0.600	0.400
IND	Question	0.534	0.458	0.236
IND	Thinking	0.638	0.588	0.363
IND	Understand	0.547	0.498	0.257
KEN	Agree	0.648	0.604	0.462
KEN	Question	0.453	0.358	0.162
KEN	Thinking	0.654	0.600	0.363
KEN	Understand	0.636	0.595	0.431
All	Average	0.586	0.531	0.336

VI. Analysis: Visualising Selected Feature Subsets

Each feature component in the feature selection subset corresponds to a pair of trackers. This provides information about which facial regions are used by the regressor for NVC recognition. It is useful to know which areas of the face are involved in NVC expression: to assist understanding of human behaviour and to develop effective feature extraction methods. In order to visualise areas of the face relevant to NVC expression, each feature component of the geometric feature is assigned a weight based on the contribution that the feature component makes to the performance. As feature component i is removed at SBE iteration r, an increase $\mathbf{o}_r$ in performance from $\mathbf{p}_{r-1}$ to $\mathbf{p}_r$ where $(\mathbf{p}_r > \mathbf{p}_{r-1})$ indicates the component was detrimental and is ignored. Conversely, if the performance $\mathbf{p}_r$ drops when a component i is removed, this indicates the component was relevant.

$$\mathbf{o}_r = \begin{cases} |\mathbf{p}_r - \mathbf{p}_{r-1}| & \text{if } \mathbf{p}_r - \mathbf{p}_{r-1} > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

The modulus of the performance drop $\mathbf{o}_r$ is added to the weight of the two trackers $\mathbf{w}^a_r$ and $\mathbf{w}^b_r$ that correspond to the component i.

$$\mathbf{w}^a_{r-1} = \mathbf{w}^a_r + \mathbf{o}_r \quad (6)$$
$$\mathbf{w}^b_{r-1} = \mathbf{w}^b_r + \mathbf{o}_r \quad (7)$$
$$(8)$$

After the SBE process is run to completion, the tracker weights $\mathbf{w}^x_0$ are normalised to form normalised weight $\hat{\mathbf{w}}^x$ which makes the tracker maximum weight equal to one

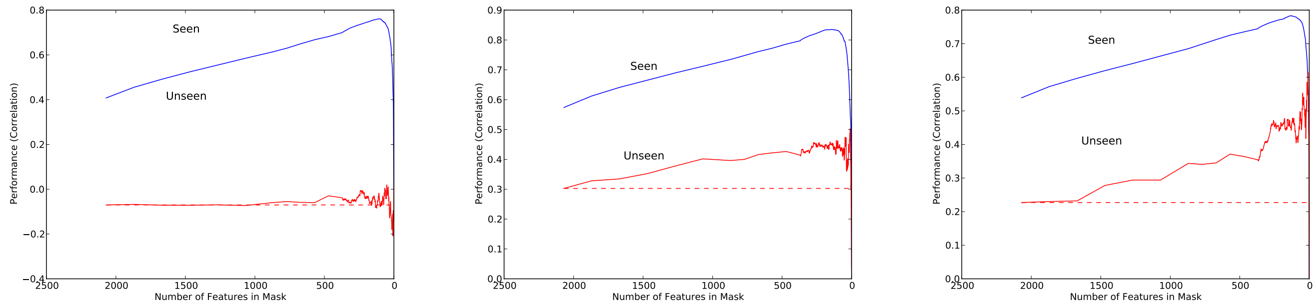


Fig. 4. The performance of the system progressively improves as backward feature selection eliminates poor features. The upper line shows the seen data, which is used in the feature selection algorithm. The lower line shows the performance of the unseen data. The left plot shows GBR *question* performance (subject 1011). The centre plot shows KEN *agree* performance (subject 1011). The right plot shows GBR *thinking* (subject 3008).

$$\hat{\mathbf{w}}^x = \frac{\mathbf{w}_0^x}{\max_x(\mathbf{w}_0^x)}, \tag{9}$$

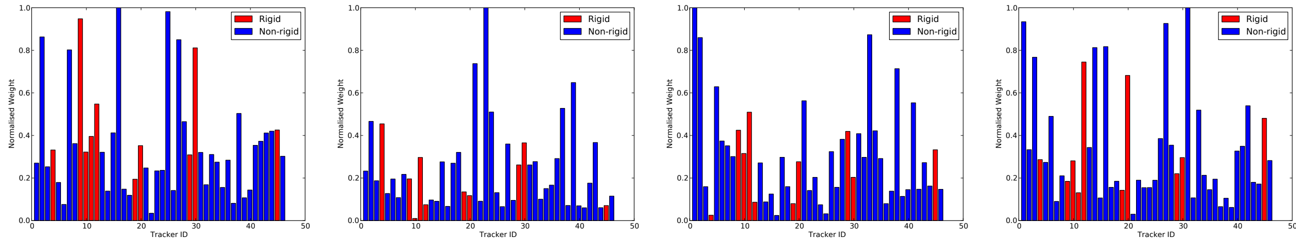
for $x \in [1..\kappa]$.

To investigate the relative importance of head pose when compared to the role expression, the trackers have been manually divided into rigid and non-rigid facial points. The manual division of trackers is shown in Figure 2. However, note that it would also be possible to automatically separate points into rigid and flexible sets, as described by Del Bue et al. [16]. The normalised tracker weights for each of the four NVC categories are shown in Figure 5. All NVC categories have significant weights assigned to trackers on flexible parts of the face, which implies expression is significant for NVC recognition. The weights assigned to rigid trackers are relatively low for *question* NVC and to some extent in *thinking*. This suggests that these NVC signals are largely conveyed by expression, with head pose having little importance. In contrast, the rigid tracker weights have higher weights in *agree*, which suggests that head pose has a role in the automatic recognition process. This confirms our expectation that agreement is often expressed by the nodding of the head. The weightings also shows that some trackers that have low weights for all of the studied NVC signals. The lowest weighted tracker overall was number 22, which corresponds to a part of the eyebrow. This may indicate either a problem with this tracker or that this area is redundant for recognising the considered NVC signals—but may be useful for others.

Although each tracker weight corresponds to a specific area of the face, it is difficult to form an overall impression of which areas of the face are involved, based only on these bar charts. A better approach is to visualise the relevant areas in relation to an actual face. However, the visualisation process is complicated by the head pose. Head pose changes are not localised to an specific area of the face and should be discarded. The head pose is generally encoded by the distance between two rigid points on the face. Facial deformations can either be encoded by distances which are either between rigid to flexible facial points or between flexible to flexible facial points. The remaining non-rigid points correspond to the flexible regions of the face and are responsible for facial deformations. The facial areas are based on a Voronoi tessellation of the face [17], based on tracker positions on a manually selected frontal view of the face. The normalised weights of each tracked point are used to control the saturation of the local area in the image. Relevant areas are shown as normal saturation. Irrelevant areas are shown as desaturated, which makes the colour tend to pure white for low weights. This enables an intuitive way to visualise relevant areas for NVC expression around the face.

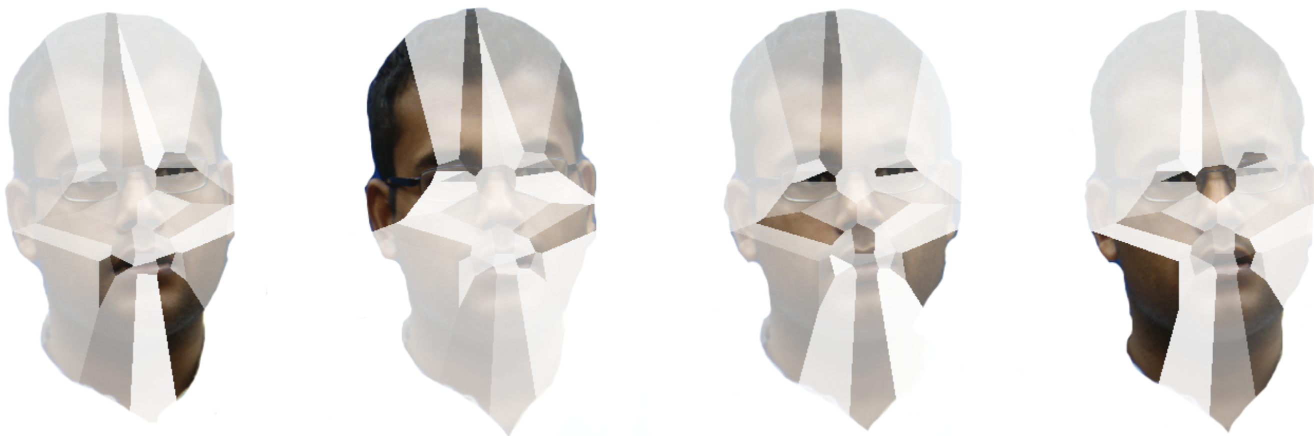
The results of the visualisation are shown in Figure 6. The clearest example of facial areas corresponding to our expectation is for *thinking*. The eyes are prominently selected and gaze is already known to play a role in *thinking* NVC. The other features provide evidence for less well understood NVC. The brow region seems significant in *question* NVC. When intense examples of *question* are viewed, there is generally consistent brow lowering, lasting for less than a second, which occurs at the end of a question sentence. The feature selection indicates this behaviour is used as the basis for recognition. This connection between verbal questioning and brow lowering has not been previously reported in published research, although Ekman mentions unpublished experiments which found this association [18]. Brow raising and lowering has also been documented in sign language but in this context, the direction of raising or lowering has a distinct semantic meaning, depending on the type of question that is being asked [19]. For *agree* and *understand*, the areas selected are less specific but generally indicate that the eyes and mouth are involved and the brow area is not used. While the visualisation shows areas that are involved in NVC recognition by machine learning, it does not necessarily imply that humans use these areas for recognition, but shows that information is present in these areas. However, there is a strong possibility that humans also use this information during NVC perception. This approach could also be improved by using additional trackers, which would increase the spatial resolution of the visualisation.

The visualisation of the feature selection subsets used annotation data from a single culture. It may be possible to investigate if other cultures use different areas of the face for



(a) agreeing (b) questioning (c) thinking (d) understanding

Fig. 5. Bar charts showing the normalised weights of tracking features for the four NVC categories. Rigid and non-rigid trackers are shown as different colours, which indicate the relative importance of expression vs. head pose in recognition. The tracker ID numbers correspond to the numbering in Figure 2. Results are from GBR culture.



Agree Question Thinking Understand

Fig. 6. Visualising the areas of face used for feature generation. The face is segmented based on Voronoi tessellation. More saturated areas indicate the importance of an area, less saturated areas are not relevant for a particular NVC. Results are from GBR culture. The visualisation areas have been averaged across test folds.

NVC perception, based on feature selection. Gaze patterns are culturally dependent for emotion recognition [12]. However, humans may be using different areas of the face for recognition compared to an automatic system, and the current feature extraction process is not expected to be as comprehensive as human perception. Regardless, the areas used by an automatic system may provide indirect clues as to the way human perception operates. This cross cultural visualisation is not attempted in this article as this would require a larger video corpus, more comprehensive facial encoding and additional annotation data to provide a reliable result.

VII. Conclusions

This paper describes a method to select an effective subset of facial shape features for the recognition of NVC. Geometric features contain a great deal of redundant and irrelevant information. A SBE based method is used to find a subset of features that are relevant for a specific NVC signal, for a particular culture annotation group. This results in a significant performance increase. The feature subset is then visualised to show the facial areas used by the automatic system. This provides evidence of which facial areas are involved in the expression of each NVC signal. Knowing the areas of the face used for NVC can suggest feature types that better encode these local areas, avoids computation of irrelevant or redundant features, as well as improving our understanding of human behaviour.

The areas of the face that are used by the system either correspond to the expected areas, or for NVC signals that are less well understood, they give an indication as to the facial areas that are involved. The areas used for each NVC is different, which implies that the feature selection has isolated feature components that are specific to each NVC. Thinking is known to involve gaze aversion and this is clearly seen in that feature components that encode eye movement are retained by the feature selection process. Based on reviewing corpus videos, it was manually observed that a sentence ending with a question is often accompanied by a brief brow lowering and this is also consistent with the visualisation of questioning NVC.

The termination of the SBE process was based on the peak performance of the training data used in the optimisation. This does not select the optimal number of features but it still resulted in a significant performance increase. If a system can be manually tuned, a slightly better performance can be achieved but the optimal number of features depends on the specific NVC.

The features are only considered as simplistic temporal variations. The temporal encoding currently considers an entire clip, so cannot temporally localise relevant motion in NVC expression. However, using a more detailed temporal encoding that considers variation in a sliding window, a particular time and area of the face could be identified as important for NVC automatic regression. The feature selection framework also might provide a framework to extend the existing automatic system to other feature types. Considering many different areas of the face (or holistic facial features) over multiple time scales and temporal offsets will result in a vast number of potential features. For this reason, techniques that are suitable for spotting patterns in large data sets, such as data mining, may be relevant to facial analysis.

The feature selection method presented here is a simple but computationally intensive approach, taking several days to complete. The removal of many features during the early iterations was necessary to make the approach practical but the performance implications of this approximation are not well understood. Other feature selection methods may be investigated to reduce the computation requirements and improve performance.

Acknowledgment

This work was supported by funding from the UK Home Office and the EPSRC project EP/I011811/1.

References

[1] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775 – 1787, 2009.

[2] Y. Yang, G. Wang, and H. Kong, "Self-learning facial emotional feature selection based on rough set theory," *Mathematical Problems in Engineering*, 2009, article ID 802932.

[3] J. Kittler, *Pattern Recognition and Signal Processing*, Alphen aan den Rijn, The Netherlands: Sijthoff and Noordhoff, 1978, ch. Feature Set Search Algorithms, pp. 41–60.

[4] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, "Feature selection of facial displays for detection of non verbal communication in natural conversation," in *Proceedings of the IEEE International Workshop on Human-Computer Interaction*, Kyoto, Oct 2009.

[5] ——, "Cultural factors in the regression of non-verbal communication perception," in *Proceedings of the Workshop on Human Interaction in Computer Vision*, Barcelona, Nov 2011. [Online]. Available: http://personal.ee.surrey.ac.uk/Personal/T.Sheerman-chase/

[6] E. Ong, Y. Lan, B. Thobald, R. Harvey, and R. Bowden, "Robust facial feature tracking using multiscale biased linear predictors," in *Proceedings of the International Conference on Computer Vision*, 2009.

[7] Y. Saeys, I. n. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm344

[8] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: Automatic analysis of brow actions," in *Proceedings of the 8th International Conference on Multimodal interfaces*. New York, NY, USA: ACM, 2006, pp. 162–170.

[9] S. Petridis and M. Pantic, "Audiovisual laughter detection based on temporal features," in *Proceedings of the 10th International Conference on Multimodal Interfaces*. New York, NY, USA: ACM, 2008, pp. 37–44.

[10] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 2007, pp. 1085–1088.

[11] R. el Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, vol. 10. Washington, DC, USA: IEEE Computer Society, 2004, p. 154.

[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*, vol. 19, no. 18, pp. 1543 – 1548, 2009.

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

[14] B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, pp. 1207–1245, 2000.

[15] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," in *Proceedings of the 16th European Signal Processing Conference*, Lausanne, Switzerland, August 2008.

[16] A. Del Bue, X. Lladó, and L. Agapito, "Non-rigid face modelling using shape priors," in *Proceedings of the IEEE International Workshop on Analysis and Modelling of Faces and Gestures*, ser. Lecture Notes in Computer Science, S. G. W. Zhao and X. Tang, Eds. Springer-Verlag, 2005, vol. 3723, pp. 96–107.

[17] G. L. Dirichlet, "über die reduktion der positiven quadratischen formen mit drei unbestimmten ganzen zahle," *Journal für die Reine und Angewandte Mathematik*, vol. 40, p. 209227, 1850.

[18] P. Ekman, *Gesture, Speech, and Sign*. Oxford: Oxford University Press, 1979, ch. Emotional and conversational nonverbal signals, pp. 45–55.

[19] ——, *Gesture, Speech and Sign*, 1999, ch. Emotional And Conversational Nonverbal Signals, pp. 45–55.