Charting the behavioural state of a person using a Backpropagation Neural Network

Janet Rothwell, Zuhair Bandar, James O'Shea, David McLean

J. Rothwell, Z.Bandar, J. O'Shea, D. McLean
Department of Computing and Mathematics,
Manchester Metropolitan University,
John Dalton Building,
Chester Street,
Manchester M1 5GD,
United Kingdom
E-mail: j.d.oshea@mmu.ac.uk
Telephone: 0161-247-1546

We thank those who kindly volunteered to participate in the study.

Journal of Neural Computing and Applications. DOI 10.1007/s00521-006-0055-9.
2006. *Self-archived, final copy for typesetting before publication*

Abstract

This paper describes the application of a backpropagation Artificial Neural Network (ANN) for charting the behavioural state of previously unseen persons. In a simulated theft scenario participants stole or didn't steal some money and were interviewed about the location of the money. A video of each interview was presented to an automatic system, which collected vectors containing nonverbal behaviour data. Each vector represented a participant's nonverbal behaviour related to "deception" or "truth" for a short period of time. These vectors were used for training and testing a backpropagation ANN which was subsequently used for charting the behavioural state of previously unseen participants. Although behaviour related to "deception" or "truth" is charted the same strategy can be used to chart different psychological states over time and can be tuned to particular situations, environments and applications.

Behaviour, Deception, Nonverbal, Multichannels, Backpropagation, Chart

1 Introduction

Nonverbal behaviour consists of all the signs and signals--visual, audio, tactile and chemical--used by human beings to express themselves, except speech or manual sign language [1]. These cues hold rich information about a person's mental, behavioural and/or physical state and are, at least in part, involuntary and unintended. Physiological cues, such as breathing rate, also hold such information, but these generally require invasive contact with the subject. The work presented in this paper is limited to visible nonverbal behaviour. Data representing this behaviour can be collected from a sequence of images, is non-invasive and does not limit the movement of a subject. The work can easily be extended to include physiological measurements, other nonverbal behaviour, such as voice pitch, and verbal behaviour such as the number of self-references made when a participant speaks [2].

Although there has been much interest in nonverbal behaviour research, very limited work has been related to the use and role of machines. Typically, the collection and analysis of nonverbal data has been done manually. A human judge watches a sequence of images, one frame at a time, observing and coding a particular 'channel' such as whether a person has blinked in an individual frame. Viewing and coding is repeated a number of times so that data for many channels are collected. Coding may simply be a note of whether a particular behaviour is taking place, may be a measure of duration over a number of frames or may be a subjective opinion [1]. Once the coding is complete the multichannel codes are analysed over multiple frame groupings. This analysis is complex and time consuming, due to the high dimensionality possible and important patterns in the data may be missed altogether.

The use of Artificial Neural Networks offers a way of overcoming these problems. Pertinent to this work, researchers have used ANNs for the detection and classification of objects related to humans -- the whole body [3], head [4], face [5], gender [6, 7], identity [5, 6, 8], facial muscle movement [9], facial expression [5, 9, 10] and hands [3, 11]. Typically, concentration has been upon either tracking or the pattern classification of a stationary object. We use an ANN for tracking and pattern classification of moving objects in head and shoulder

images [12]. This allows for the automatic extraction of movement and pattern data, other data extraction such as colour information, and thus the recording of multichannel codes over multiple frame groupings.

Researchers have also used ANNs to classify emotions. The assumption has been that there is a simple relationship between an actual emotion and a particular facial expression or facial muscle movement determined by an ANN. For example posed emotions, such as happiness, sadness, surprise, fear, anger and disgust, have been linked, by a simple one-to-one correspondence, to particular facial expressions determined by an ANN [6, 10] and a posed emotion or blend of emotions has been obtained by an analysis of multiple muscle movements each determined by an ANN [9].

More complex behaviours have been classified by a single measure, a small number of measures combined simply or measures that monitor occasional, discrete events. An ANN may have determined some or all of these measures. Pavidis et al. [13] detect blood-flow under the eyes in order to determine "deception". Cohn and Katz [14] combine facial muscle movements with four vocal measures to determine "negative", "neutral" and "positive" emotion. Ekman [15] combines false smiles (ones which do not involve eye movement) and one vocal measure to determine "deception". The detection of a short-lived unexpected expression has been classified as 'deceptive' behaviour [16]. Bartlett et al [17] detect short lived, unexpected expressions and/or the presence of a false smile in order to determine "deception".

Few researchers have attempted to classify complex behaviours using ANNs to analyse many measures. For complex behaviours our hypothesis is that channel interrelationships are important rather than the individual channels. Iwano et al. [18] use a number of channels and look for closeness between a response pattern and a *typical* pattern. A computer is used to extract a number of facial distances from a roughly stationary head from the time period just prior to a person responding. These are combined with two vocal channels in an attempt to determine "agreement", "denial", "unexpected", "withhold", and "dislike". Iwano et al. calculate a typical pattern for each of the five behaviours by looking at each response from all the participants. The patterns may be quite simple, for example, a nod and raised pitch would tend to indicate "agreement". They compare each

pattern to the five typical patterns. They only tested people who had been used to create the typical patterns. For complex behaviours such as deception, there may not be a single typical pattern but a number of patterns that could represent the behaviour.

We automatically extract multichannel codes for whole responses from videos containing head and shoulder images [12]. Training a classification ANN with multichannel codes over multiple frame groupings allows the ANN to learn the many patterns that represent each complex behaviour. The extracted data, each vector representing a whole response, was used to train and test an ANN to classify whole responses related to "truth" and "deception". This is described in an unpublished manuscript entitled 'Silent Talker: An Adaptive Psychological Profiling System Using Artificial Neural Networks' (Rothwell, J., Bandar, Z, O'Shea, J., McLean, D., 2005).

The classification of whole responses in this manner raises a number of issues. First, responses can vary significantly in length. A response may be a short 'grunt' or a long protracted explanation and the patterns exhibited may be very different. Second, over a protracted collection time, such as a long response or a description of events within a normal conversation, a person's nonverbal behaviour can alter significantly. Third, although an ANN can be trained to detect the beginning and end of a response, it may be more desirable to analyse behaviour continuously because a person's nonverbal behaviour does not stop when they stop speaking. "If his lips are silent, he chatters with his fingertips; betrayal oozes out of him at every pore" [19]. In this paper we analyse continuous nonverbal behaviour. The same set of videos is used. Other videos have been recorded – the results will be presented in a future paper. Although the classification ANN is used to chart behaviour related to "truth" and "deception" the same strategy could be used to chart other complex behaviours by the use of different, appropriate, teacher signals for training and testing.

2 Materials and Methods

2.1 Participants

In this paper we just concentrate upon the *unplanned* interviews of 15 English men. This is because results from Rothwell [12] and the unpublished work showed that men, women and persons of different races might have different deceptive behaviours and that *unplanned* and *planned* interviews might be subtlety different. For the purposes of this work we define an *unplanned* interview as one in which the participant has been asked a set of questions for the first time and with no prior indication of what is to be asked.

2.1.1 Video Collection

The interview collection strategy was taken from nonverbal behaviour literature [20] where the detection of deception was the phenomena under scrutiny. In a simulated theft scenario each participant performed a task and then had an interview regarding the possible theft of some money. Some participants stole some money during the task. The interviewer asked ten ordered questions in a consistent manner. The participant lied if they had stolen the money and told the truth if they hadn't stolen the money.

Q1 I know that you've looked in the box. Please tell me what you saw:-
Q2 How much money was in the box?
Q3 What did you do with the money?
Q4 Are you sure that you didn't take the money from the box?
Q5 Have you any pockets?
Q6 When did you last use your pockets?
Q7 Did you put the money in your pockets?
Q8 Please describe the contents of your pockets:-
Q9 Are you telling the truth about the location of the money?
Q10 have you lied in this interview?

In a truthful interview all ten questions produced truthful responses. In a deceptive interview many of the questions produced deceptive responses. However Q5 produced a truthful response because all the participants had pockets and Q1 and Q2 were usually truthful responses because most participants admitted that they had looked in the box. The mean interview length, from the beginning of the response to Q1 to the end of the response to Q10 was 61 seconds (*SD*=11 seconds).

During each interview a camera was used to record the interviewee's head and shoulders. Recording was to SVHS tape. Each video was digitised into a 15 frames-per-second Audio–Video Interleaved (AVI) file of width 384 pixels and height 288 pixels. A response was deemed to start one second before the interviewee started talking and end two seconds after they stopped talking. Using this criterion the mean response length for the *unplanned* interviews of English men was 5.8 seconds (*SD*=3.2 seconds). A response *overlap* sometimes occurred if the interviewer asked a question in less than three seconds.

From the 15 English men there were 14 useable *unplanned* interviews each containing 10 responses. One man's interview was faulty because the interviewer was partly obscuring the camera. By removing this man from the data set there were equal numbers of deceptive and truthful *unplanned* interviews. Seven participants (persons 1, 3, 6, 7, 8, 11, 13) stole some money prior to the interview. These *deceptive* interviews contained some responses that were deceptive (falsifying/concealing) and some responses that were truthful. Seven participants (persons 2, 4, 5, 9, 10, 12, 14) hadn't stolen any money prior to the interview. These interviews were *truthful* and contained only truthful responses.

2.1.2 Channels

In this work we use the word *channel* to describe behaviour at a finer level of granularity than is normally used in deception literature. Each channel is an attribute being measured, such as an eye contact event, gaze direction or body movement. In this work 37 channels are used. The left eye provides 8 channels. The right eye provides 8 channels. Head movements in the x, y and z planes provide 5, 5 and 3 channels respectively. Head angle and rotation provides 5 channels. Blushing and blanching provides 2 channels. One channel called 'slot'

relates to the time period for the current collection. When short fixed time periods are used ‘slot’ does not vary between input vectors. This channel could be removed in this circumstance. In this work it is retained - it simply acts as an extra bias to the ANN being trained.

2.1.3 Extraction of Multichannel Codes over Multiple Frame Groupings

An automatic tracking and extraction system was employed which itself uses ANNs [12]. Initially each channel of interest is allocated a data store to hold data for a particular number of frames. In the first frame an initial search area is determined. The face is located followed by the other objects of interest – for example the eyes, eyebrows, nose. The locations and state of each object along with other data (such as colour) is collected. This data is used to add values to each channel store. Figure 2 shows the tracking in process. Although the figure is shown in grey-scale, the actual program displays and uses colour data. The process is repeated for the frames of interest for which collection is taking place. Whilst this collection occurs multiple values within a channel store, representing the multiple frames, can be used to produce a single value that represents the channel over those multiple frames. Because this can happen for all the channels a vector is produced that represents the multichannels over multiple frames. These vectors, which each contain data from the 37 channels derived from the video frames, form the input of the ANN which acts as the behaviour classifier. The inputs to the ANN are real-valued in the range –1 to 1.

2.2 Implementation

There are 14 useful unplanned interviews and 139 useful unplanned responses – one response being lost due to interviewer error. For each response it is known whether the response is deception (falsification/concealment) or the truth. We make the assumption that if a whole response is a lie and contains data indicative of deceptive behaviour then some of the 1-second time periods taken from that response will also contain data indicative of deceptive behaviour. A response was deemed to start one second before the interviewee started talking and end two seconds after they stopped talking.

Three data sets were collected for training and testing purposes. Data set A contains 133 vectors from the 139 responses each representing one-second time data extracted from the second just prior to a person opening their mouth for each response. Sometimes a sequence of frames did not produce a data vector. This was because the extractor was unable to collect enough valid data due to, for example, a loss of tracking for a short period of time. Data set B contains 2040 vectors - these being overlapping one-second time periods collected at 5 frame intervals from the 139 responses. Although it is possible to collect every possible overlapping one-second time period from throughout a response this strategy would provide too many vectors close to one another and would tend to cause overfitting of the solution to the training and/or validation set. Therefore we 'thin' this data such that a vector representing one-second (15 frames) is collected every 5 frames rather than every frame. Data set C contains 137 vectors each representing a whole response. This data was collected for comparison purposes.

2.3 ANN Training and Testing

In this work we partition the data set into training (Tr) , validation (Val) and testing (Te) sets. Tr was used for ANN weight adjustments and the selection of other neural network parameters. Val was used as a stopping condition. Te was used for final testing. Neither Val nor Te was used to adjust the ANN parameters. Te is used for testing the trained ANN. We also use the n-fold cross-validation technique [21] where the data is randomly divided into n subsets. One subset is reserved for testing and the remaining n - 1 are used for training and validation. The role of the test set is rotated through the n subsets and the n results are averaged to give the overall classification accuracy. A further set of permutations in the training runs is due to the need to initialise the random connection weights, in the range –1 to 1, before starting training. The training runs will be repeated a number of times (e.g. 3) with different starting seed values. This is because some (unpredictable) combinations will result in the training of the ANN stalling at a sub-optimal level of performance [22]. Just as the classification accuracy for the validation set is used as a stopping criterion for training it is also used to choose a single ANN from a set of trained ANNs that had different seeds (different random starting weights).

In this work a number of trials were done. Each trial was the training and subsequent testing of an ANN system. A standard feedforward backpropagation (BP) ANN [22] was used which employed the delta rule and incremental updating. A number of initial exploratory trials, described below, were conducted in order to choose workable ANN training parameters, topologies and stopping conditions.

2.3.1 Data Representation and Activation Function

McLean (1995) showed that improved training times, stable learning and improved generalisation can be achieved by using each vectors element scaled to a width of 2 centred on zero thus giving values in the range –1 to 1. Assuming multiple layers the bipolar activation function has to be used in order to present data to each following layer in this format. In this work data was presented in the range –1 to 1 and a bipolar sigmoid activation was used with a steepness of value of one. The output from the ANN was bipolar in the range –1 to 1, '1' representing a deceptive response and '-1' representing a truthful response.

2.3.2 Initial weights

It is common practice to initialise the weights to small randomly chosen zero-mean values [24]. A weight range on the order of $1/\sqrt{f}$ where f is the 'fan-in' (number of neuronal connections) has been suggested [25] because the weights have zero mean and a standard deviation of unity. However regions of high sensitivity in weight space have been detected [26] where two very close initial points can lead to substantially different learning curves. The weight search is primarily governed by the initial starting conditions of the later layers of neurons and much of the search for the solution through weight space is conducted along the axis corresponding to the first layer's weights which attempt to satisfy the requirements set by the weights in the later layers [27]. Small changes in the initial weights of later layers cause a large difference in the final network state. For this data domain we compared three values for the initial weight range, these being 0 ± 1, $0 \pm 1/f$ and $0 \pm 3/\sqrt{f}$. Classification Accuracies were similar. The 0 ± 1 initialisation required the largest training time. The $0 \pm 3/\sqrt{f}$ initialisation offered the best classification overall for a single ANN but the $0 \pm 1/f$

initialization offered the best mean classification for all the tested topologies, learning rate parameter values and starting seeds. This work uses starting weights of $0 \pm 1/f$.

2.3.3 Learning Rate Parameter

Plaut et al. [28] suggested that for optimal weight changes the learning rate parameter (η) should be inversely proportional to the fan-in. Schraudolph and Sejnowski [29] also found that using different η values per layer was the key to efficient learning. It reduced both the overall training time and variance, indicating a more reliable (in terms of generalisation) optimisation process. For this data domain we initially compared five values of η these being 0.001, 0.01, 0.1, 1, 10 for each layer. As expected 0.001, 1 and 10 were unsuitable values. When η was 0.001 there was extremely slow learning. Little or no learning took place when η was 1 or 10 due to excessive oscillation and/or plateau (indicating a saturation of one or more neurons). Different values per layer were then tried for a one-hidden layer ANN. All topologies tested had one output neuron and fewer hidden layer neurons than input connections thus each hidden layer neuron had a larger fan-in than the output neuron. A complex relationship emerged. The best ANN performance was achieved when the output layer η was larger than the hidden layer η. For all topologies a good working value of η was $1/f$ for each layer. However, when there were significantly more positives than negatives or vice-versa, improved learning resulted when the ANN was forced to pay more attention (increasing η) to the positive or negative samples occurring less frequently. This was achieved by setting $\eta = 1/(2pf)$ for each vector presented (a 'dynamic' η) where p is the proportion of vectors having this vector's bipolar value and f is the fan-in of a unit. When the data set had equal proportions then η was $1/f$ for both positives and negatives. When the proportions were different, for example 0.4 positives and 0.6 negatives, then η was slightly larger for any positive ($1.25/f$) compared to any negative ($0.83/f$).

2.3.4 Number of Layers and Neurons

Cybenko [30] proved that at most two hidden layers are needed given enough units per layer and that a single hidden layer feedforward network employing

sigmoidal hidden unit activations, a sufficient number of hidden units and a single linear output is capable of approximating any continuous function $f:R^n \rightarrow R$ to any desired accuracy [31].

It is desirable that a network is not too small. Having too few neurons in the hidden layers provides too few weights that can be trained with the result that the network is unable to converge to a suitable solution. It is also desirable that the network is not too large. Just as using a stopping criterion based upon Val reduces the likelihood that the network will overfit to Tr so does limiting the number of weights.

Using a randomised data set collected from the same domain we trained and tested a single hidden layer ANN using the data representation, activation function, initial weights and learning rate parameters chosen above. The trial was repeated 64 times each ANN having a different number of hidden layer neurons (8) and starting seed value (8). The number of hidden layer neurons tested was 4,5,6,7,8,9,10 and 11. The best Te performance was achieved when the number of hidden layer neurons was 7 (Figure 1).

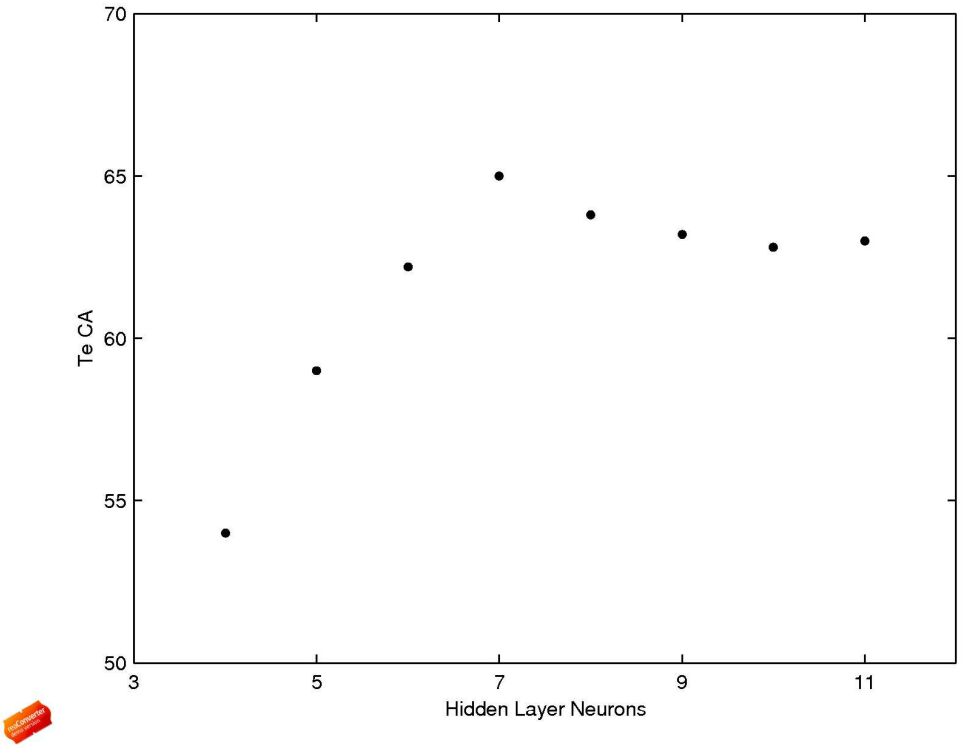

Fig. 1

Tr, Val and Te were close in vector space because the data set comprised randomised vectors. When Tr, Val and Te are further apart in vector space, for example, because Tr, Val and Te each represent a small set of completely different people, then it is even less desirable that the ANN should overfit to Tr or rely too much on the influence on Val. This can be achieved by reducing the number of neurons slightly. Trials were done using data sets from the problem domain. These data sets were partitioned in various different ways. A workable topology, training within a sensible time-frame for all methods of partitioning, was a single hidden layer ANN containing 6 neurons. Trials were also conducted using ANNs with two hidden layers which had similar numbers of weights as their single layer equivalents. There was no significant difference in performance but the two hidden layer network offered a slightly increased training speed. A topology of 38:6:1 plus a bias connection is used for the main trials in this paper.

2.3.5 Calculation of the ANN Performance

For the purposes of this paper we define DA as the 'deception accuracy', the percentage of deceptive time periods classified correctly. We define TA as the 'truth accuracy', the percentage of truthful time periods classified correctly. CA is the overall 'classification accuracy' i.e. the combined percentage of truthful and deceptive responses classified correctly in their respective categories. This is calculated as CA=(TA+DA)/2.

2.3.6 Stopping Criteria

During training CA was monitored for a Validation Set (Val). An epoch (one full pass through the training set) was designated *good* if Val CA was higher than the largest previous value. When calculating Val CA a limit value of 0.5 was used. The teacher signal representing a truthful time period is '-1'. The teacher signal representing a deceptive time period is '1'. A time period was classified correctly if the ANN gave a value in the range 0.5 to 1 for a time period representing "deception" or a value in the range –0.5 to –1 for a time period representing "truth". All other ANN responses were incorrect. Weights were saved for each *good* epoch. There were a number of *checking* epochs and a *maximum* number of epochs designated that affected the stopping point. For data sets A and C

checking epochs=1000 and *maximum* epochs = 20,000. Training was stopped if there were 1000 consecutive *checking* epochs without a *good* epoch occurring or if the total number of epochs exceeded 20,000. The values were 100 and 2000 respectively for data set B which had a larger number of training vectors. The weights finally saved were those for the last *good* epoch.

2.3.7 Summary of ANN Topology and Parameters Used

A standard BP ANN was used which employed the delta rule, incremental updating and a topology of 37:6:1 plus a bias connection. A bipolar activation function with steepness value of 1 and an input data representation of –1 to 1 was used. Output data from the ANN was in the range –1 to 1, '1' representing "deception" and '–1' representing "truth". The starting weights were randomly initialised into the range $0 \pm 1/f$ where f was the fan-in of a unit. The learning rate dynamically changed for each vector presented, and was $\eta = 1/(2pf)$ where p is the proportion of vectors having this vector's bipolar value and f is the fan-in of a unit. Tr was used for training, Val was used for validation and Te was used for testing. For data sets A, B and C the checking epochs were 1000, 100, 1000 and the maximum epochs were 20000, 2000, 20000.

3 Training and Testing Results

Half of the interviews were truthful and contained 10 truthful (tt) responses. The other interviews were deceptive and contained 7 deceptive (dd) responses, which were falsification or concealment, and 3 truthful (td) responses. In real-life situations it is highly likely that a person deceiving will tell the truth some of the time. It has been found that people telling the truth in an unplanned deceptive interview tend to exhibit a deceptive behaviour because they know that they have or are about to deceive [12]. The behaviour may be ambiguous. In these trials the classification for tt responses was always *truth* and the classification for dd responses was always *deception*. However, in some trials we treated td responses as *truthful responses* and in other trials as *deceptive responses.* A correct classification of dd as *deception* correspond to DA; a correct classification of tt as *truth* corresponds to TA; When td is treated as *deception* a correct classification

corresponds to DA. When td is treated as *truth* a correct classification corresponds to TA.

The first main trial, Trial A, uses data set A (First second only). With this trial we wish to confirm our assumption that a short one-second time period taken from a response does indeed contain data indicative of "truth" and "deception". This data set is a good starting point because Iwano et al. [18] showed that collection immediately prior to a response was important. Trial B uses data set B (overlapping one-second time periods) in order to chart responses and whole interviews. Trial C uses data set C (whole responses). This trial is for comparison purposes.

3.1 Trial A: First second from responses only

Fourteen Tr:Val:Te sets were constructed such that Tr:Val contained the randomised vectors of 13 English men, Te contained the vectors from the remaining English man. The trial was repeated 84 times each ANN having a different Tr:Val:Te set (14), seed (3) and td classification (2). Training and testing was such that vectors related to truthful responses from a deceptive interview, *td* responses, were classified as *deception* in 42 trials and *truth* in 42 trials. When td=*deception* there were 64 vectors representing truth and 69 representing deception. When td=*truth* there were 85 and 48 vectors respectively.

Table 1 compares different methods for classifying responses. For this data set the results for classifying responses are those for the responses' first second. Row A1 relates to td=*truth* and row A2 relates to td=*deception*. The results are above chance levels. The td responses/vectors are classified more easily when it is assumed that the vector offers a "deceptive demeanour". An ANN trained upon vectors where td is assumed to be "truthful" has a strong truthful bias (row A1). This may be because there are more vectors representing "truth". However η was changed dynamically such that the ANN was forced to pay more attention to vectors representing "deception". This technique had previously caused a 'balancing' of the proportions of positives and negatives.

The trial has confirmed our assumption that a short one-second time period taken from a response can contain data indicative of "truth" and "deception". One-second time periods taken from throughout a response, rather than the

beginning of a response, are likely to have a wider and more noisy pattern in vector space. These would be more suitable for training an ANN for charting responses because the ANN is developed using a truly representative data set.

Trial	Training (Tr) Variation		Response Testing (Te)(%)				
		td	dd	td		tt	
			DA	DA	TA	TA	CA
A1	First Second of each response	"truthful"	44		52	83	65
B1	1-sec every 5 frames of each response	"truthful"	47		52	87	68
C1	Whole Responses	"truthful"	58		43	91	72
A2	First Second of each response	"deceptive"	67	62		67	66
B2	1-sec every 5 frames of each response	"deceptive"	63	62		83	73
C2	Whole Responses	"deceptive"	71	72		82	77
		weight	0.7	0.3	0.3	1	2

Table 1

3.2 Trial B: One-second Time Periods from Responses

Trial B was a repetition of Trial A except that Data set B was used. When td=*deception* there are 985 vectors representing truth and 1055 representing deception. When td=*truth* there are 1271 vectors representing truth and 769

representing deception. When td=*truth* then dd = 44%, td = 51%, tt = 75%, CA=60%. When td=*deception* then dd = 58%, td = 57%, tt = 70%, CA=63%. These are above chance but what is interesting is how these ANNs classify the individual responses and how the ANNs behave over a full interview – not only during the responses but also in the gaps between responses. Trial A indicated that the period just prior to a person opening their mouth does contain nonverbal data. It is reasonable to assume that the gaps between responses also contain some data.

Whether a response is designated as truthful or deceptive can be based upon the summation of the ANN output for the overlapping one-second time periods for a response – either collected every 5 frames or every 1 frame. If the summation is greater than zero then, overall, the ANN has indicated "deception". If the summation is less than zero then, overall, the ANN has indicated "truth". Using this criterion (and collection every 5-frames) offers the classification of responses shown in Table 1 rows B1 (td=*truth*) and B2 (td=*deception*). Again, using td=*truth* makes the ANN have a bias towards classifying truthful responses.

3.3 Trial C: Whole Responses

Trial C was a repetition of Trial A except that Data set C was used. When td=*deception* there were 69 vectors representing truth and 69 representing deception. When td=*truth* there were 89 and 48 vectors respectively. The results are shown in Table 1. Row C1 relates to td=*truth* and row C2 relates to td=*deception*. As was found previously, the td responses are classified more easily when it is assumed that the response offers a "deceptive demeanour". The lower classification accuracies for the short time periods may be because certain channels are having less of an effect. For example an eye blink may not occur at all within a one second time period.

3.4 Charting Responses

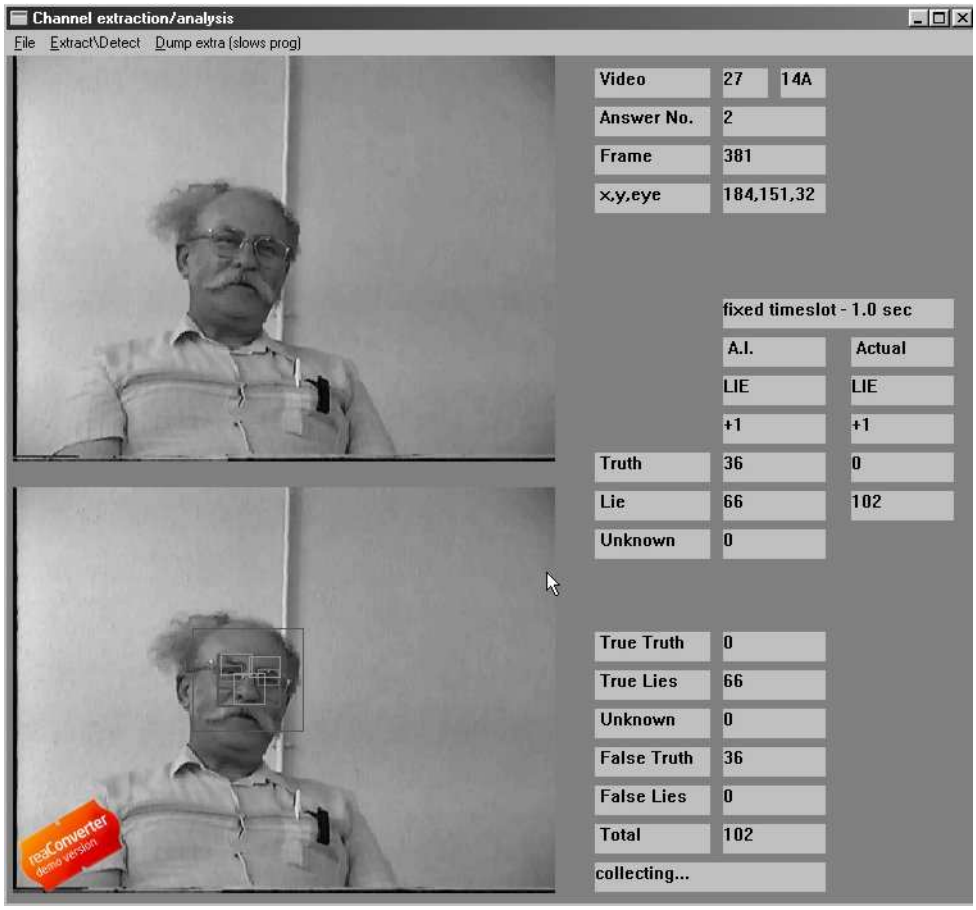

Fig. 2

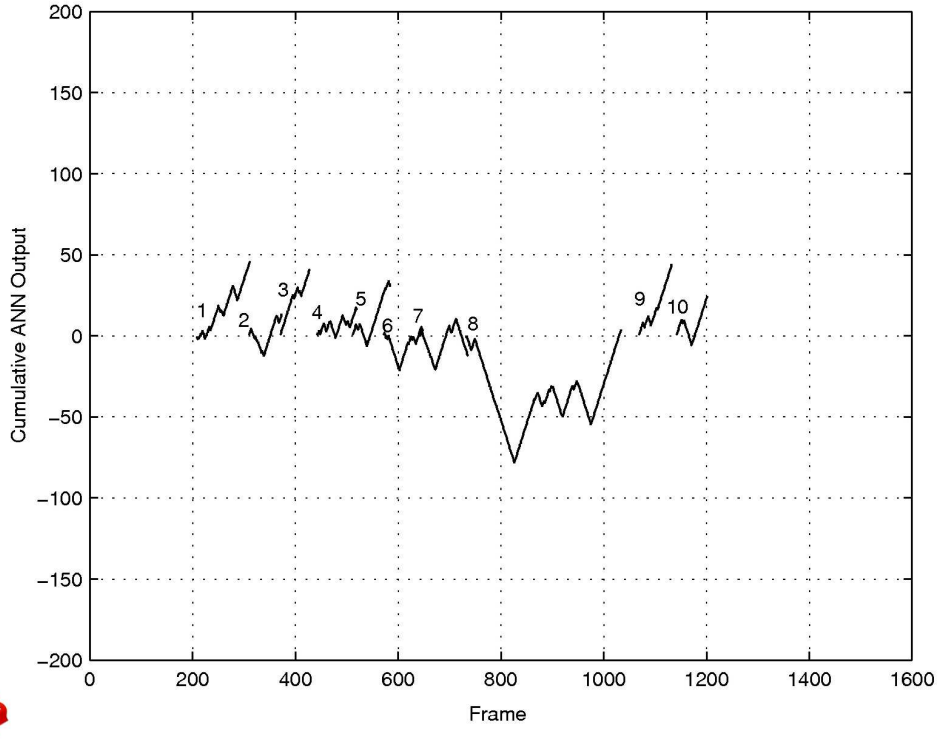
Cumulative ANN Output
Frame
200
150
100
50
0
−50
−100
−150
−200
0
200
400
600
800
1000
1200
1400
1600
1
2
3
4
5
6
7
8
9
10

Fig. 3a

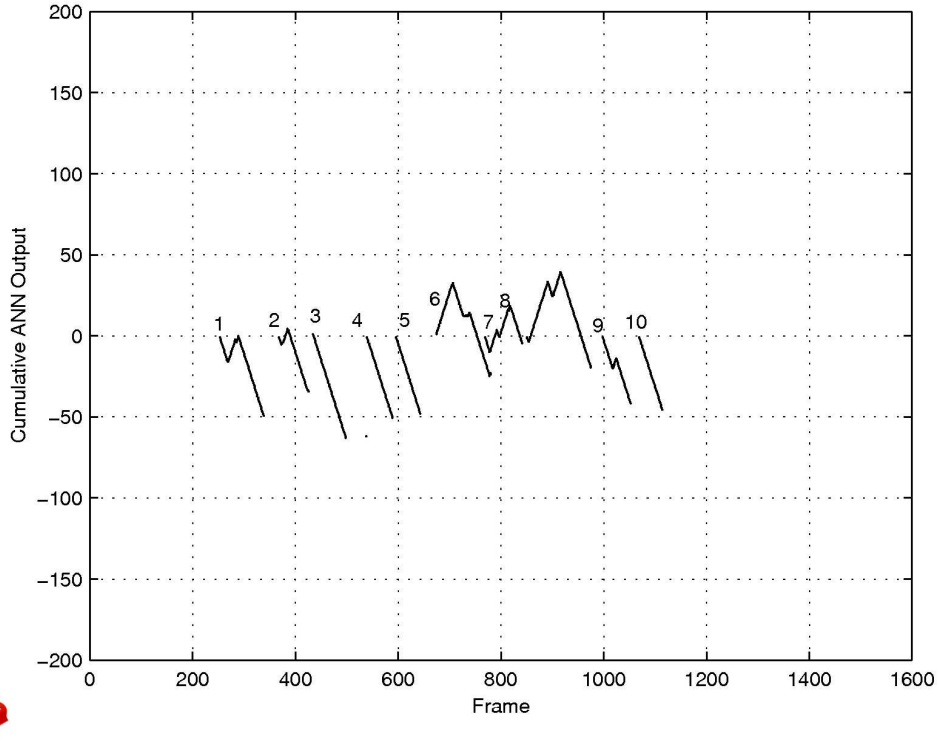
Cumulative ANN Output
Frame
200
150
100
50
0
−50
−100
−150
−200
0
200
400
600
800
1000
1200
1400
1600
1
2
3
4
5
6
7
8
9
10

Fig. 3b

Figure 2 shows the automatic extractor and one-second classification ANN developed in Trial B in use on overlapping (each frame) one-second time periods for a previously unseen man (person 6). Graphical results show a trend in behaviour. The one-second ANN can be used to look at the 'shape' of each response. Responses for Persons 1 (**a**, *deceptive*) and 2 (**b**, *truthful*) are shown in Figure 3. Choice of these two persons was simply based upon them being the first "deceptive" and first "truthful" English men in the data set.

The ANN gives an output in the range –1 to 1 for each one-second time period. When the ANN gives a value between 0 and 1 (deception) for each one-second time period, overlapping at single frame rate, then the plotted cumulative sum [32] has a positive gradient. Similarly a negative gradient is produced when the ANN is continuously giving an output between 0 and –1 (truth). The shape of a chart line gives an indication of truthful and deceptive behaviour over a period of time. When the ANN is continuously providing a value of 1 then the gradient is '1' and this is shown as a 45 degree upwards trend when the axes are drawn to a one-to-one ratio. Figures 3 and 4 have the axes at a 1:1 ratio.Figure 2 does not have the axes at a 1:1 ratio – the y-axis has been stretched to more clearly show the line shapes. Similarly a 45 degree downward trend is produced by an ANN continuously outputting a '-1'. Gradients between 1 and –1 are caused by the ANN outputs being between –1 and 1, and/or noise and/or a change of behaviour occurring.

Person 1's responses tend to show an upward trend (deceptive), person 2's responses show a downwards trend (truthful). The simple summation method described in the previous section is equivalent to a response line ending above or below the x-axis. The graphical method shows the overall trend but also a line 'shape' which may contain other useful information.

3.5 Classification of Interviews

A simple criterion can be used – for example an interview can be deemed deceptive if three or more responses out of ten are classified as being deceptive. When it is specified that only a very small number of deceptive responses have to be detected for classification of a "deceptive" interview then, when classifying

many interviews, there will tend be an increase in false positives (truthful interviews incorrectly classified as deceptive). When it is specified that a large number of deceptive responses have to be detected there will be an increase in false negatives (deceptive interviews incorrectly classified as truthful). A suitable criterion would need to be established by the analysis of many interviews and the situation being examined. In a real-life deception scenario a bias towards false negatives may be preferable so as to reduce the likelihood of incorrectly accusing an innocent person.

Using the results for responses from tasks A, B and C, and the criterion that three or more deceptive response have to be detected, then when td = *deception*, 12, 9 and 10 out of 14 interviews are classified correctly with 2, 5 and 2 false positives and 0, 0 and 2 false negatives. When td=*truth* 13, 10 and 13 out of 14 interviews are classified correctly with 1, 2 and 1 false positives and 0, 2, 0 false negatives. The particular criterion seems to be better when td=*truth*.

Altering the criterion such that more than half need to be deceptive then, when td = *deception*, 11, 12 and 10 out of 14 interviews are classified correctly with 1, 0 and 1 false positives and 2, 2 and 3 false negatives. When td= *truth* 10, 10 and 9 out of 14 interviews are classified correctly with 0, 0 and 1 false positives and 4, 4, 4 false negatives. The particular criterion seems to be better when td=*deception*.

A slight change in the criterion affects the overall interview results. A sensible strategy would be to specify the starting criterion as requiring the detection of more than half of the 'expected' deceptive data. So, when td=*deception* there are 10 possible responses portraying a deceptive demeanour, so more than 5 need to be classified as deceptive. However when td = *truth* there are a possible 7 responses portraying a deceptive content so more than 3.5 (4 or more) need to be classified as deceptive. The criterion could then be adjusted to require the detection of extra deceptive responses so as to reduce the likelihood of incorrectly accusing an innocent person for example.

To provide robustness – results from various measures could be combined – for example the classification results from data sets A, B and C weighted appropriately as determined by large volume experiments. When td = *deception*, more than 50% of responses are required to be deceptive and the ANNs created

using A, B and C are deemed to be as important as one another, then 12 out of 14 interviews are classified correctly (2 false negatives). When td = truth, and 40% or more are required to be deceptive and the ANNs created using A, B and C are deemed to be as important as one another, then 12 out of 14 are classified correctly (1 false positive and 1 false negative).

3.6 Charting of Interviews

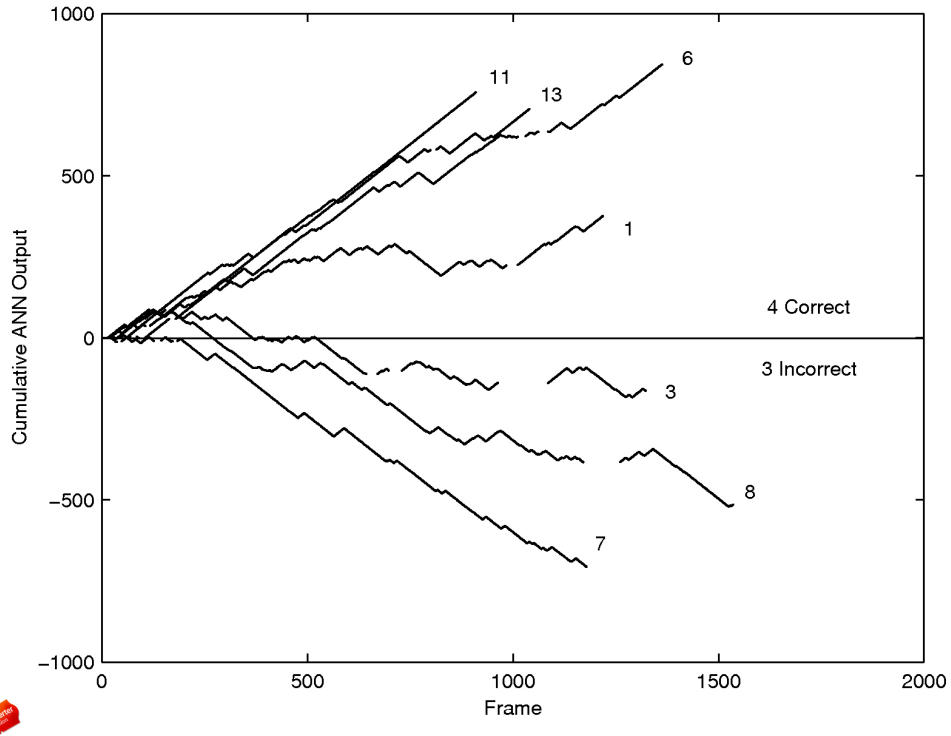

Fig. 4a

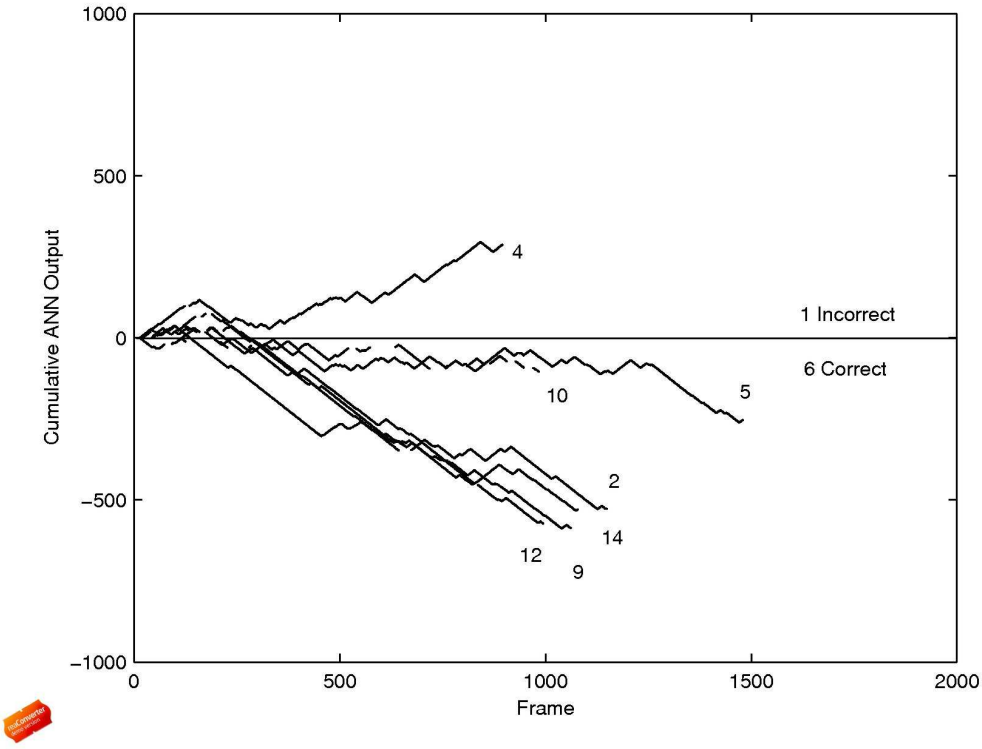
1000
500
0
−500
−1000
0
500
1000
1500
2000
Frame
Cumulative ANN Output
4
1 Incorrect
6 Correct
10
5
2
14
12
9

Fig. 4b

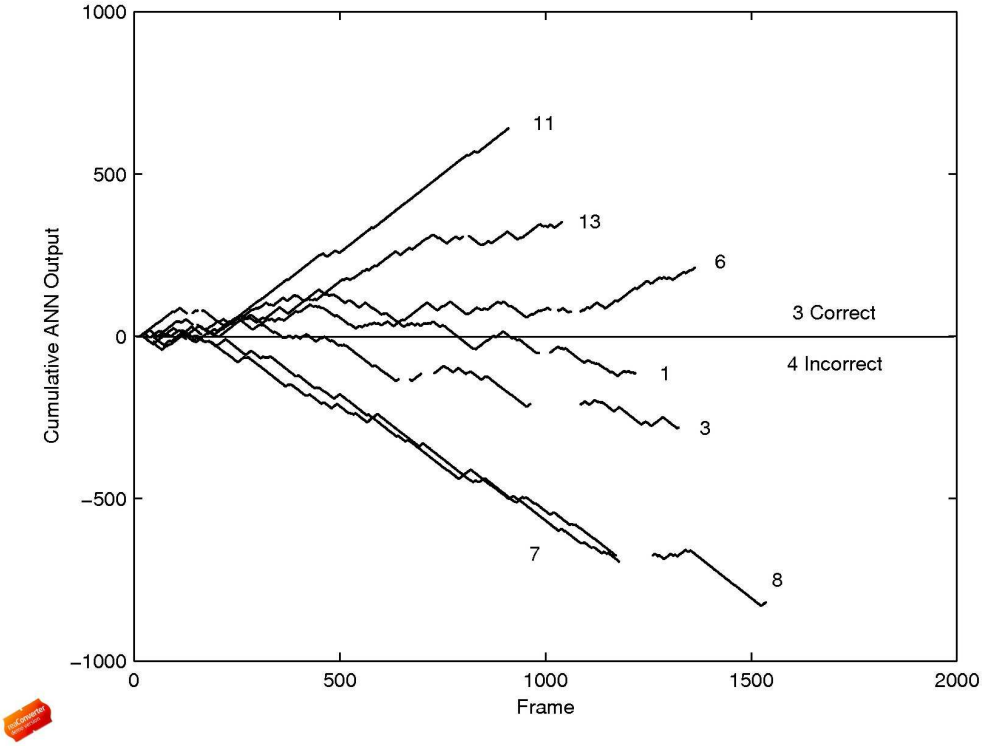
1000
500
0
−500
−1000
0
500
1000
1500
2000
Frame
Cumulative ANN Output
11
13
6
3 Correct
4 Incorrect
1
3
7
8

Fig. 5a

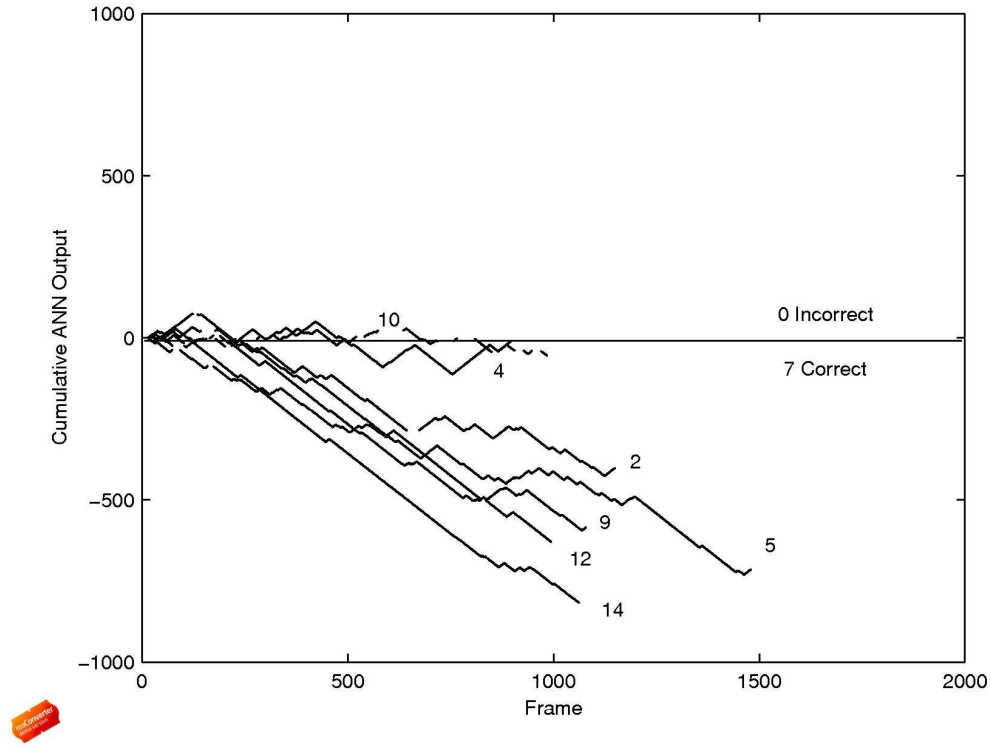

Fig. 5b

Charts for the 14 individuals are shown in figure 4 (td=*deception*) and figure 5 (td=*truth*). The top part of each figure shows results from people who lied. The lower part of each figure shows the people who told the truth. Each labelled line represents a whole interview for an individual. A positive gradient indicates deception and a negative gradient indicates truth. Gaps in the chart indicate where the tracking was lost or where not enough data was collected for presentation to the ANN. If the overall trend of each chart is used for the classification of interviews then 10 out of 14 are classified correctly. When td =*deceptio*n there is 1 false positive and 3 false negatives. When td = *truth* there are 4 false negatives. It is interesting that person 4 (the only false positive) informed us that he was using eye drops at around the time of the interview.

4 Discussion

4.1 Multichannel Approach

In this work patterns between nonverbal multichannels are detected in order to classify previously unseen people. This work shows that behaviours associated with “truth” and “deception” can be detected and that small fixed time frames can be used which allow for the charting of a person’s behaviour over time. This is useful because over a protracted collection time, such as a long response or a long description of events within a normal conversation, a person’s nonverbal behaviour can alter significantly. When short time-periods are used the start of individual responses does not need to be detected because the analysis is continuous. We believe that a multichannel approach is needed to obtain reasonable accuracy rates and a robust system. The relationship between the channels is particularly important rather than the individual channels. ANNs are a suitable way of overcoming the problems of handling and analysing high dimensional data.

Specialist multichannel pattern classifiers, such as “Unplanned Response English Female classifier” can be created fairly easily. This is useful because different scenarios, interviewer and interviewee types provide different channel patterns [33, 34, 35]. The choice of classifier can be made automatically in-use by the use of artificial intelligence. The classification accuracy compares favourably with other methods of detecting behaviours related to truth and deception [12].

When the overall trend of each chart was used for the classification of interviews then 10 out of 14 were classified correctly. When a suitable criterion is used with whole responses, classification of the whole interview appears to be slightly better (12 out of 14). However, the graphical method holds promise for a number of reasons. The start and end times of responses do not need to be determined manually or by ANN methods. This work has only presented charts created from ANNs trained upon one-second time periods. A different time period may be more suitable. The charts hold more information than an overall trend. Small twists and turns over time may hold useful data. The changes in gradient may prove to be of interest - especially over responses. Responses can be compared. For example concealment containing some truthful description (e.g.

Q8) may provide a very different chart shape to falsification (e.g. Q3 and Q10). Such data could be analysed by an additional ANN to offer improved performance.

4.2 Limitations and Future Work

This is a laboratory, low stake situation. Although we attempted to increase the motivation of the participants the behaviours being displayed are more comparable to everyday lies than real-life high stake lies. High-stakes are easier to detect [35] but currently we don't know exactly how the trained system will behave with real-life data. There are also many variables to consider - exhibited behaviour may vary depending upon the interviewer, stakes and physical setting.

The channels that we measured have been investigated by other researchers [2, 36] but as individual channels the results were contradictory and thus did not clearly indicate deception. We hypothesized that this was because the channel interrelationships were important not the individual channels. The inclusion of channels that offer more indicative behaviours, such as finger movements [20, 35], is likely to further improve the accuracy rates. This is because, as for the Polygraph machine, a subject could learn how to manipulate (or hide) a particular channel that is known to be important [37].

When filming the interviews the camera was to one side of the interviewer. Because the interviewee was looking at the interviewer the recorded interviews showed a small bias in the baseline eye gaze. An ANN trained upon such data could not be used with videos which were filmed where the interviewer was at the other side. Although each video frame can be mirrored this would not be applicable because we don't know what determines a person's gaze direction as they look away to think for example. It is possible that gaze direction is related to reading direction [38], or it could reflect hemispherical asymmetries in the brain [39]. There are various solutions. An ANN could be trained upon both types of data. The questioner could be a computer and camera directed by a human or an artificial intelligence conversational agent.

For future work, it is probable that a larger data set containing more examples of behaviours will improve the overall results. As more examples of behaviours are seen by the system the classification upon previously unseen

participants is improved. The use of specialist multichannel pattern classifiers, for example ‘English Female Classifier’, may also offer improved accuracies. If the system is to be developed for real-life situations then real-life video data will need to be collected for training and testing purposes along with the known ground truth as determined by other means. Interviewer, stakes and physical setting could be varied in order to examine exhibited behaviours in these different scenarios.

The addition of more channels (Vrij, 1994; Vrij, 2000), will very likely improve the classification accuracy. Channels could be nonverbal and auditory cues (e.g. finger/hand, foot, torso movement, voice pitch), speech content related (e.g. number of ‘self references’) or some other, such as brain wave activity.

Short time-periods offer the ability to ‘chart’ a person’s behaviour over a response, period of time or interview. Only one-second time periods were used for training and testing. Other time periods may be more suitable and an analysis of the chart shapes may prove useful.

Multiple cameras can also facilitate channel extraction from a number of people at once, so that individual behaviours and interactions between people can be analysed. In order to use the device in real-life situations, the data directly from a camera needs to be processed ‘on-the-fly’.

A typical application for non-invasively measuring deception over a short interview time for large numbers of people would be monitoring behaviour at airport check-ins or a customs declaration area. This could be made completely automatic by using a machine questioner in the training, testing and in-use phases. This would allow staff to concentrate on those passengers more likely to be of concern. Other applications may be related to medicine, education and business for example. Data is currently being analysed for a possible medical application.

5 Conclusion

A backpropagation ANN was trained upon visible nonverbal behaviour data and was subsequently used for charting the behavioural state of previously unseen English men. Although behaviour related to “deception” or “truth” was charted the same strategy can be used to chart different psychological states over time and can be tuned to particular situations, environments and applications. The system relies upon data collected non-invasively and can be used in real-time assuming

that either the start and end times of responses are not needed or that an ANN has been trained to detect responses. The system does not rely upon a small number of channels and demands no human expertise and allows for the exchange and/or addition of channels. For each application, many output classifications and situational variants can be examined. Different classifiers can be trained for different groups of people and situations in order to finely tune the system. Machines that rely upon one, two, three or a limited number of channels can easily be affected by some factor other than the one being measured or by the subject's conscious control of those channels. Because the system operates upon multiple channels of nonverbal behaviour it is likely to be a better detector of the psychology of a subject.

References

1. Scherer KR, Ekman P (eds.) (1982) Handbook of methods in nonverbal behavior research. Cambridge University Press, Cambridge

2. Zuckerman M, Driver RE (1985) Telling lies: Verbal and nonverbal correlates of deception. In A.W. Siegman, & S. Feldstein (eds.). Multichannel integrations of nonverbal behavior. L Erlbaum Associates, Hillside, NJ

3. Gavrila DM (1999) The visual analysis of human movement: A survey. Comput Vision and Image Understanding, Academic Press 3(1):82-98.

4. Turk MA, Pentland AP (1991) Face Recognition Using Eigenfaces. IEE Comput Society Conference on computer vision and pattern recognition pp 586-591.

5. Valentin D, Abdi H, O'Toole AJ, Cottrell GW (1994) Connectionist models of face processing: A survey. Pattern Recognition 27(9):1209-1230.

6. Cottrell GW, Metcalfe J (1991) Empath: Face, emotion and gender recognition using holons. In Lippman et al. (eds.). Advances in neural processing systems, Morgan-Kaufmann Publishers Inc., San Mateo, 3, pp 573-577.

7. Golomb BA, Lawrence DT, Sejnowski, TJ (1991) SEXNET: A neural network identifies sex from human faces. Advances in Neur Information Processing Systems 3, Morgan Kaufmann Publishers Inc., San Mateo.

8. Zhang M (1996) Face recognition using artificial neural network group-based adaptive tolerance (G.A.T.) trees. IEEE Trans Neur Netw 7(3)

9. Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Measuring facial expressions by computer image analysis. Psychophysiology 36:253-263.

10. Rosenblum M, Yacoob Y, Davis LS (1996) Human expression recognition from motion using a radial bias function network architecture. IEEE Trans Neur Netw 7(5):1121-1138.

11. Fels S, Hinton GE (1998) Glove Talk 2, A neural network interface which maps gestures to parallel formant speech synthesizer controls. IEEE Trans Neur Netw 9(1):205-212.
12. Rothwell J (2002) Artificial neural networks for psychological profiling using multichannels of nonverbal behaviour. PhD Thesis, Manchester Metropolitan University
13. Pavidis I, Eberhardt NL, Levine JA (2002) Human behavior: Seeing through the face of deception [Brief communication] Nature, 415 pp 35-35.
14. Cohn JF, Katz GS (1998, September) Bimodal expression of emotion by face and voice. Proceedings of the Sixth ACM international conference on Multimedia: Workshop on Face/gesture recognition and their applications pp 41-44. NY: ACM Press.
15. Ekman P (1988) Lying and nonverbal behavior: Theoretical issues and new findings. Journal of Nonverbal Behavior 12:163-175.
16. Johnson RC (1999, April 12) Computer program recognizes facial expressions. Electronic Engineering Times, pp 51-51.
17. Bartlett MS, Donato G, Movellan JR, Hager JC, Ekman P, Sejnowski TJ (1999) Face image analysis for expression measurement and detection of deceit. Proceedings of the 6th Annual Joint Symposium on Neural Computation. San Diego, CA: The Institute For Neural Computation, University of California.
18. Iwano Y, Sukegawa H, Kasahara Y, Shirai K (1995) Extraction of speaker's feeling using facial image and speech. IEEE International Workshop on Robot and Human Communication, pp 101-106.
19. Freud S (1959) Fragments Of An analysis Of A Case Of Hysteria, Collected papers, Vol. 3, New York :Basic books.(Original work published 1905).
20. Vrij A (1994) The Impact of Information And Setting On Detection Of Deception By Police Detectives, Journal Of Nonverbal Behavior 18:117-136
21. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B 36:111-147.
22. Hassoun MH (1995) Fundamentals of neural networks. MIT Press, Cambridge, MA:
23. McLean D (1995) Improving generalisation in continuous data domains. Ph.D. Thesis, Manchester Metropolitan University
24. Rumelhart DE, McClelland J (1986) Parallel distributed processing. In PDP Research Group (eds). Explorations in the micro structure of cognition Vol 1, Bradford Books, MIT Press, Cambridge, MA
25. Wessels LFA, Barnard E (1992) Avoiding false local minima by proper initialization of connections. IEEE Trans Neur Netw 3(6):899-905.
26. Kolen JF, Pollack JB (1991) Back propagation is sensitive to initial conditions. In Lippmann et al. (eds.) Advances In Neur Information Processing Systems 3, Morgan Kaufmann, San Mateo, California, pp 860-867.
27. McLean D, Bandar Z, O'Shea J (1997) The evolution of a feed forward neural network trained under backpropagation. Proc of the 3rd International Conference on Artificial Neur Netw and Genetic Algorithms (ICANGA97); Springer-Verlag pp 518-522.

28. Plaut D, Nowlan SJ, Hinton GE (1986) Experiments on learning by back propagation. Computer Science Department, Carnegie Mellon University, Pittsburgh,PA 15213.
29. Schraudolph NN, Sejnowski TJ (1996) Tempering backpropagation networks:Not all weights are created equal. In Touretzky et al. (eds). Advances in Neur Information Processing Systems 8, MIT Press, Cambridge.
30. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Mathematical Control Signals Systems 2:303-314.
31. Cybenko G (1998) Continuous values neural networks with two hidden layers are sufficient. Technical Report, Department of Computer Science, Tufts University, Medford, MA.
32. McCormick D, Roach A (1987) Measurement, Statistics and Computation. Wiley, London
33. Bond CF (1990) Lie Detection Across Cultures. Journal of Nonverbal Behavior 14:189-204
34. Feldman RS, Rime B (1991) Fundamentals of non-verbal behavior. Cambridge University Press, Cambridge, England
35. Vrij A (2000) Detecting lies and deceit: The psychology of lying and the implications for professional practice. Wiley, NY
36. Ekman P (1985) Telling lies: Clues to deceit in the marketplace, politics and marriage. W.W. Norton and Company, NY
37. Williams D (2001) How to sting the Polygraph. Sting Publications. P.O. Box 720568, Norman, OK, U.S.A
38. Singh M, Vaid J, Sakhuja T (2000) Reading/writing vs handedness influences on line length estimation. Brain and Cognition 43:398–402.
39. Reuter-Lorenz PA, Kinsbourne M, Moscovitch M (1990) Hemispheric control of spatial attention. Brain and Cognition 12: 240–266.

Fig. 1

Classification Accuracy for a Test Set from the same problem domain for different numbers of hidden layer neurons in a single hidden layer ANN. When calculating Te CA a time period was classified correctly if the ANN gave a value in the range 0.5 to 1 for a time period representing “deception” or a value in the range –0.5 to –1 for a time period representing “truth”. All other ANN responses were incorrect

Fig. 2

The automatic multichannel multiframe extractor in use and the subsequent use of a one-second ANN to classify the behaviour of a previously unseen subject (Person 6)

Fig. 3

Charting of Individual Responses from the x-axis for Person 1 and Person 2

Fig. 4

Charting of Interviews for 14 previously unseen English men (td=*deception)*

Fig. 5

Charting of Interviews for 14 previously unseen English men (td=*truth)*

Table 1

Deception Accuracy (DA), Truth Accuracy (TA), and total Classification Accuracy (CA) in percent for the unplanned responses of 14 previously unseen English men.

Trial	Training (Tr) Variation		Response Testing (Te)(%)				
		td	dd	td		tt	
			DA	DA	TA	TA	CA
A1	First Second of each response	"truthful"	44		52	83	65
B1	1-sec every 5 frames of each response	"truthful"	47		52	87	68
C1	Whole Responses	"truthful"	58		43	91	72
A2	First Second of each response	"deceptive"	67	62		67	66
B2	1-sec every 5 frames of each response	"deceptive"	63	62		83	73
C2	Whole Responses	"deceptive"	71	72		82	77
		weight	0.7	0.3	0.3	1	2

Journal of Neural Computing and Applications. DOI 10.1007/s00521-006-0055-9.
2006. *Self-archived, final copy for typesetting before publication*