# Intelligent Deception Detection through Machine Based Interviewing

James O'Shea[1], Keeley Crockett[1], Wasiq Khan[1], Philippos Kindynis[2], Athos Antoniades[2], Georgios Boultadakis[3]
[1]School of Computing, Mathematics and Digital Technology,
Manchester Metropolitan University, Manchester, M1 5GD, UK, K.Crockett@mmu.ac.uk
[2]Stremble Ventures LTD, 59 Christaki Kranou, 4042 Germasogeia, Limassol, Cyprus
[3]European Dynamics, Brussels

*Abstract*— **In this paper an automatic deception detection system, which analyses participant deception risk scores from non-verbal behaviour captured during an interview conducted by an Avatar, is demonstrated. The system is built on a configuration of artificial neural networks, which are used to detect facial objects and extract non-verbal behaviour in the form of micro gestures over short periods of time. A set of empirical experiments was conducted based a typical airport security scenario of packing a suitcase. Data was collected through 30 participants participating in either a truthful or deceptive scenarios being interviewed by a machine based border guard Avatar. Promising results were achieved using raw unprocessed data on un-optimized classifier neural networks. These indicate that a machine based interviewing technique can elicit non-verbal interviewee behavior, which allows an automatic system to detect deception.**

**Keywords- neural networks, avatar, deception detection**

## I. Introduction

Border control officers' tasks rely on bilateral human interaction such as interviewing an individual traveller using verbal and non-verbal communication to both provoke response and interpret the traveler's responses. Automated pre-arrival screening could greatly reduce the amount of time a participant spends at the border crossing point and may improve security control. Such a system would complement existing border control technology such as Advanced Passenger Information systems and future systems such as the new Entry/Exit System centralized border management system which will facilitate the automation of border control process (due for implementation in 2020) [1].

This paper presents initial work on an Automated Deception Detection system known as ADDS which is powered by a conversational agent avatar and is capable of quantifying the degree of deception on the part of the interviewee. ADDS forms part of the iBorderCtrl (Intelligent Portable Control System) [2] whose aim is to enable faster and more thorough border control for third country nationals crossing the land borders of EU Member States (MS) [2,3]. The final version of ADDS will utilize an advanced border control agent avatar which conducts an interview with a traveller. The avatar attitudes will be personalized to communicate with the traveler including utilizing subtle non-verbal communication cues to stimulate richer responses from them. A strong focus will be on identifying the impact on non-verbal communication expressed by the avatar on the performance of ADDS.

Nonverbal behaviour is used by humans to communicate messages, which are transmitted through visual and auditory features such as facial expressions, gaze, posture, gestures, touch and non-linguistic vocal sounds [4]. A human being continually transmits nonverbal behavior, which can be produced subconsciously, in contrast to spoken language. The majority of work on the use of non-verbal behaviour (NVB) to determine a specific cognitive state has been undertaken by human observers, who are often prone to fatigue and produce different subject opinions. Hence, an automated solution is preferable. Related, but limited work has been done in the automated extraction of NVB from a learning system [5] to detect comprehension levels and also in detection of guilt and deception [6, 7]. Both of these examples have used artificial neural networks to first detect micro gesture patterns and then perform classification successfully.

Time is also a factor, as interviewers need to interact longer with travelers to reach a conclusion on their deception intent. Such time comes at a premium in border control, resulting in short and potentially false positive results in the field. An automated system, which utilizes a few minutes of traveler time at the pre-crossing stage without increasing the amount of time they spend with a border control agent, could thus potentially increase efficacy while reducing cost. In this work deception detection in ADDS is performed by an implementation of the patented Silent Talker artificial intelligence based deception detector [6, 7].

The aim of research presented in this paper was to firstly produce a prototype trained artificial neural network (ANN) classifier to be used within the automatic deception detection system; secondly to investigate whether an avatar, machine based interviewing technique could be developed for a border security application which requires large volumes of interviews. Thus, the research question addressed in this paper can be stated as:

*Can a machine based interviewing technique elicit non-verbal behavior, which allows an automatic system to detect deception?*

This paper is organized as follows; Section II provides a description of prior work in the field of deception detection systems with emphasis on automation. The use of conversational agents is also examined in the border control context in terms of being used as avatar interviewers. Section III describes the ADDs system. Section IV presents the overall methodology of the data collection process and describes a series of border control scenarios, which are used to simulate truthful and deceptive behaviour of participants. Results and findings of a series of experiments are highlighted in Section V. Section VI presents the conclusions and future directions.

## II. Prior Work

### A) Deception Detection Systems

Human interest in detecting deception has a long history. The earliest records date back to the Hindu Dharmasastra of Gautama, (900 – 600 BC) and the Greek philosopher Diogenes (412 – 323 BC) according to Trovillo (1939). Today, the best-known method is the Polygraph [8], which was invented, by John Augustus Larson in 1921, to detect lies by measuring physiological changes related to stress. The Polygraph is a recording instrument, which displays physiological changes such as pulse rate, blood pressure, and respiration, in a form where they can be interpreted by a trained examiner as indicating truthful or deceptive behaviour. A polygraph test takes a minimum of 1.5 hours but can take up to four hours depending on the issue being tested for [8]. Individual scientific studies can be found which support [9] or deny [10] the validity of the Polygraph. A meta-study [11] conducted in 1985 found 10 studies from a pool of 250 were sufficiently rigorous to be included. From these they concluded that under very narrow conditions, the Controlled Question Test (CQT - the standard Polygraph test that could be used at border crossings) could perform significantly better than chance, but these results would still contain substantial numbers of false positive, false negative and inconclusive classifications. They also stated that many conditions needed to achieve this might be beyond the control of the examiner. Constructing a good set of control questions for this test requires substantial information about the interviewee's background, occupation, work record and criminal record to be collected before the exam. The polygraph requires physiological sensors on the traveler that would make both the set-up time and cost of an interview prohibitively expensive to apply to all travelers, thus typically if it is used, it is at a secondary stage for high-risk travelers.

Functional Magnetic Resonance Imaging (fMRI) is a technique that measures changes in activity of areas of the brain indirectly by measuring blood flow (which changes to supply more oxygen to active areas of the brain). It has been proposed that there are reliable relationships between patterns of brain activation and deception that can be measured by fMRI. It has also been reported that although fMRI is seen as overcoming some weaknesses of the Polygraph, for example by having an explanatory model based on cognitive load [12] it is highly vulnerable to countermeasures (in common with EEG-based approaches).

Voice Stress Analysis (VSA) is a technique that analyses physical properties of a speech signal as opposed to the semantic content. The technique is fundamentally based on the idea that a deceiver is under stress when telling a lie and that the pitch of the voice is affected by stress. More specifically, it claims that micro tremors, small frequency modulations in the human voice, are produced by the automatic or involuntary nervous system when an interviewee is lying. There have also been claims that the increased cognitive load of deception creates micro tremors [13]. The weight of scientific analysis is that, whatever the assumed underlying model, VSA performs no better than chance and has been described as "charlatanry" [14].

The most recent work in this area is contained in the INTERSPEECH 2016 Computational Para-linguistics Challenge: Deception, Sincerity & Native Language. Inspection of a sample of responses to the 2016 challenge shows them to be either paralinguistic, phonemic or a combination of the two, e.g. the Low Level Descriptors such as psychoacoustic spectral sharpness or phonetic features such as phonemes [15]. These techniques achieved approximately 67% using a technique called "Unweighted Average Recall" intended to take account of the fact that the Deceptive Speech Database (DSD) (dataset) from the University of Arizona was unbalanced (test set contained 24% deceptive / 76% truthful classes). We have not found evidence of a significant degree of paralinguistic research outside English.

Facial micro-expressions are short-lived, unexpected expressions. There is said to be a small "universal" set of expressions of extreme emotion: disgust, anger, fear, sadness, happiness, surprise, and contempt, meaning they are common across cultures. A formalized method of encoding micro expressions was defined by Paul Ekman, who developed commercial tools for training interviewers to recognize them [16]. One of the resources is a manual on a Facial Action Coding System for training in expression recognition. This has generated a large body of research in automating FACS for applications such as lie detection Virtually all of the findings from micro expression studies are closer to a CKT than genuine lie detection, so they do not constitute persuasive evidence for using the technique at border crossings.

### *B) Automated Deception Detection*

Silent Talker (ST) was designed to answer the criticisms of the psychology community that there are no meaningful single non-verbal indicators of deception (such as averted gaze), by combining information from many (typically 40) fine-grained nonverbal channels simultaneously and learning (through Artificial Neural Networks) to generalize about deceptive NVB from examples [6, 7]. In this respect, it does not depend on an underlying explanatory model in the same way as other lie detectors. However, it does have a conceptual model of NVB. This model assumes that certain mental states associated with deceptive behaviour will drive an interviewee's NVB when deceiving. These include Stress or Anxiety (factors in psychological Arousal), Cognitive Load, Behavioral Control and Duping Delight. Stress and Anxiety are highly related, if not identical states. The key feature of ST, as a machine learning system, is that it takes a set of candidate features as input and determines itself which interactions between them, over time, indicated lying. Thus is not susceptible to errors caused by whether particular psychologists are correct about particular NVB gestures.

Evidence to date is that no individual feature can be identified as a good indicator, only ensembles of features over a time interval provide effective classification. Early experiments with ST showed classification rates of between 74% and 87% ($p<0.001$) depending on the experimental condition [6]. There are no single, simple indicators of deception; ST uses complex interactions between multiple channels of microgestures over time to determine whether the behaviour is truthful or deceptive. A microgesture is a very fine-grained non-verbal gesture, such as the movement one eye from fully-open to half-open. This gesture could be combined with the same eye moving from half-open to closed indicating a wink or blink. Over a time interval, e.g. 3 seconds, complex combinations of microgestures can be mined from the interviewee's behaviour. Microgestures are significantly different from micro-expressions (proposed in other systems), because they much more fine-grained and require no functional psychological model of why the behaviour has taken place [6].

### *C) Conversational Agents in the Border Control context.*

A Conversational Agent (CA) is an AI system that engages a human user in conversation to achieve some practical goal, usually a task perceived as challenging by the user. Embodied CAs offer the opportunity of more sophisticated communication through gesture and supplementing the dialogue with non-verbal communication [17]. The persona of an embodied CA is referred to as an Avatar and there is (limited) evidence supporting the use of an Avatar interviewer for automated border crossing control. Nunamaker [18] reported a group of experiments, culminating in an attempt to smuggle a concealed bomb past an avatar interviewer. These, collectively, suggest that an avatar can simulate affective signals during dialogue, can have a definable persona (gender, appearance) and can elicit cues to deception. In practice, such systems tend to rely on vocal features [18] or electrodermal activity and measure arousal as a proxy for deception. Hooi & Cho [19] have reported that perceived similarity of appearance between the avatar and interviewee reduces deceptive behaviour. Furthermore, Strofer et al. [20] observed that when interviewees believe that the avatar they are interacting with is controlled by a human, they produce more physiological responses (electrodermal), e believed to indicate deception. In a cognitive neuroscience review, de Borts and deGelder [21] reported that human-like avatars that move realistically are more likeable and perceived as similar to real humans. This prior work suggests a strong potential for the use of avatars in border control interviews and the need for substantial research into the influential factors. The state of the art of this combination of technologies suggest that Avatars will be suitable for detecting deception in border crossing interviews, as they are effective extractors of information from humans [22] and therefore can applied to deception detection tasks. Secondly, they can provide dynamic responses to user inputs and can simulate affective signals [23].

## III. AUTOMATIC DECEPTION DETECTION SYSTEM

Figure 1 presents an abstract of the ADDS architecture as seen from within the final iBorderCtrl system. Each traveler who engages with the function of pre-traveller registration will be required (subject to providing informed consent) to undertake an interview with an avatar. In the final system the Avatar will adapt its attitude based upon the level of deception detected by ADDS on a question by question basis. For the purpose of training the neural network classifiers within ADDS in this paper a still image was used for the Avatar.

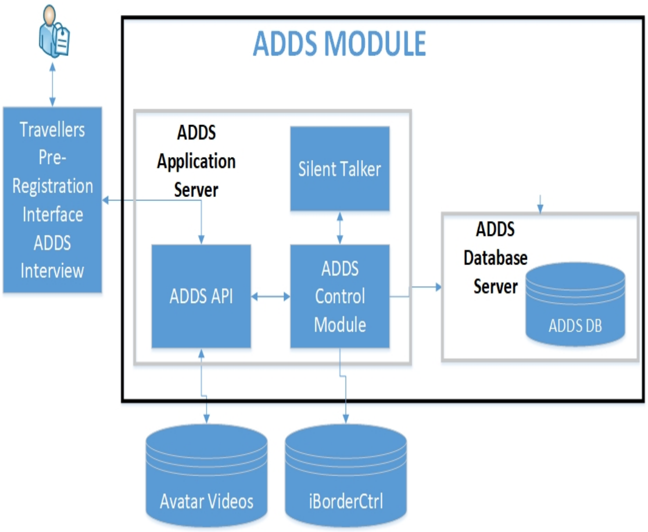

Fig.1. Automated Deception Detection System Architecture

Also, for the purpose of the research conducted in this paper, the traveler information for the simulated border crossing interview will be captured through a series of scenarios (for deceptive participants), and a post experiment questionnaire. This information will then be used to populate a local database. In practice, the ADDS API will receive encrypted information about a specific traveller from the iBorderCtrl control system database and populate an instance of a trip/traveller in the ADDS back end database server.

Classification was performed by the Silent Talker component of ADDS using an empirically determined risk level. The silent Talker component outputs the score for each of the questions and associated classification, the whole interview (score and classification) and the confirmation radio button responses. This updated the ADDS back end database server. In the final system, the ADDS Control module will use the risk scores to change the avatar attitude when the next question is asked to the traveller. In this work, the risk scores and classifications were simply stored in the ADDS local database for training, testing and validating the neural network classifiers.

*A) Silent Talker*

This work is specifically focused on the application specific development of the Silent Talker (ST) system (Fig.2.). ADDS utilizes 38 input channels to the deception network. They fall into 4 categories: eye data, face data, face angle data and 'other'.

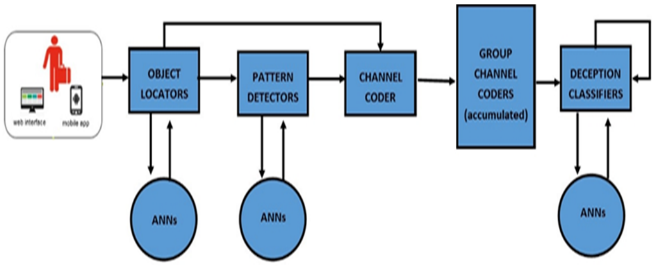

Fig.2. Silent Talker component in ADDS

ST uses features extracted from the non-verbal behaviour (NVB) of interviewees to determine whether they are lying or telling the truth. In this study, the application specific ST first receives a video stream (mobile app or web client) being received for classification  The video arrives as a sequence of frames, each frame is processed in sequence and information from the frames is compiled  or accumulated for the purpose of  classification. The deception classifiers used in this paper are multi-layer perceptrons producing a continuous output in the range -1 to +1. Empirically determined thresholding of this output was used for the truthful and deceptive classifications. The consequence of this is that some frame sequences will be labelled as “unclassifiable.”  If a single decision boundary were used, these would have outputs too close to the decision boundary to justify confidence in them. A simplified description of the components in Figure 2 now follows:

***Object Locators***: Each object locator finds the position of a particular object, (e.g. the head, the left eye, the right eye etc.) in the current video frame.  A typical object locator would consist of a back propagation Artificial Neural network (ANN) trained with samples extracted from video collected during a training experiment.

***Pattern Detectors:*** Pattern detectors detect particular states of objects located by the object locators. For example, for the left eye: left eye closed is true when the left eye is closed (1), otherwise false (0), left eye half closed is true when the left eye is half closed (1), otherwise false (0), the left eye may be considered open if neither of these pattern detectors is true.

***Channel Coder:*** The variations in the state of an object determined by specific pattern detector are referred to as a “channel”. Channel coding is the process of collecting these variations over a specific time period (i.e. over a specific number of video frames).

***Group Channel Coders:*** Group channel coders refers to the process of amalgamating and statistically summarizing the information from the individual channel coders to form a summary vector, which can be input to the final deception classifier.

***Deception Classifiers:*** Typically, the deception classifier is a single ANN trained to classify input vectors from the group channel coders as either truthful or deceptive. It is also possible to add other classifiers (for example to detect feeling of guilt) and combine these to obtain higher deception detection accuracy.

*B) Avatar*

The final ADDS system will use animation to pose each question which will be personalized for each border guard avatar in accordance with the travelers non-verbal state. A sample of a border guard avatar posing a question can be found here: http://stremble.com/iBorderCtrl/1/1/1/1.mp4). However as the development of ADDS, as a system was happening in parallel to the training and validation of the deception detection element, a still image of the male avatar developed by Stremble [24] was used within this work (Figure 3).  The avatar is presented in a uniform to convey an air of authority. In this experiment, the avatar is shown as a still image and the speech is synthesized. One reason for this was to see if any emotional states were conferred on the (actually neutral) avatar by participants.

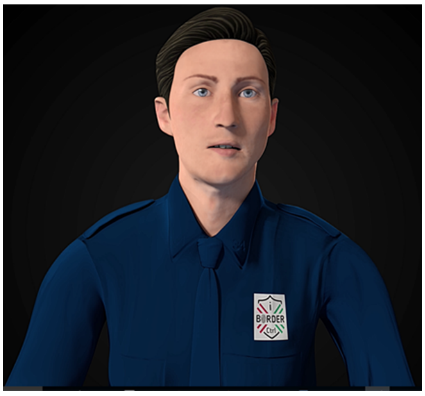

Fig.3. Male Avatar

## IV. METHODOLOGY

This section describes the methodology to conduct a quantitative empirical study of non-verbal behaviour with samples of volunteer participants concerning truthful and deceptive conditions. The hypothesis tested was:

*H0: A machine based interviewing technique cannot be used to detect deception from non-verbal behaviour*
*H1: A machine based interviewing technique can be used to detect deception from non-verbal behaviour at levels statistically greater than chance.*

### *A) Overview of data collection process*

Following an approved ethical process, data was collected through participants first performing a simulated deception task using different levels of arousal by varying the task. This was followed by a recorded interview by the Avatar Border Guard to collect video data for analysis of non-verbal behaviour. On completion of the interview, participants completed a questionnaire containing simple subjective self-ratings of internal mental states. The experiment utilised a number of scenarios in order to test the hypothesis. The two major categories are Scenario Truthful and Scenario Deceptive.

**S1: Scenario Truthful**

- All participants will use their true identities as recorded in their identification documents.
- All participants will answer questions about a real relative or friend who is an EU / UK citizen (equivalent of a Sponsor in border questions asked by EU border guards).
- All participants will pack a suitcase with harmless items typical of going on a holiday
- Participants will answer questions about identity, sponsor and suitcase contents.
- All answers to questions can be answered truthfully.

**Scenarios Deceptive**

- All participant given fake identities (male / female) and short life history.
- All participants are given a short description of a fake relative from the EU.
- Participants are stratified into 1 of 4 blocks. These blocks are intended to have different degrees of emotional intensity / deceptive stakes (in decreasing order of intensity).
  - ***S2: Simulated biohazard infectious disease in test tube with informational video about weaponization***
  - ***S3: Simulated biohazard infectious disease in test tube without informational video***
  - ***S4: Simulated Drug package (soap powder in clear packet)***
  - ***S5: Simulated Forbidden agriculture/ food product i.e. seeds.***
- Participants will answer questions about identity, sponsor and suitcase contents.

All participants engaged in truthful or deceptive scenarios were required to pack a suitcase. The standard items relevant to both deceptive and truthful scenarios that are packed are: a pair of unisex trousers, a T-shirt, shower gel, a tube of toothpaste, a hand towel and a bottle of perfume in a box. If a participant took part in a deceptive scenario, then an intervention by a confederate took place that involved the confederate modifying the contents of the perfume box with a prohibited item depending on the scenario (S2.S5). The participant was also shown typical posters of prohibited items from airport baggage handling areas.

The experimental methodology comprised a pre-interview task which sets up the scenario for truth-telling or deception, the interview itself and a debriefing stage which will include certain ethics aspects (confirmation of consent, permissions to use materials etc.) and some subjective ratings of feelings during the interview (e.g. subjective guilt etc.). Each participant was taken in to the debriefing room and asked to read the participant information sheet, invited to ask any questions and then sign the first part of the participant informed consent document. In the debriefing session, participants completed the second part of the informed consent form to confirm they still consented to their data to be used in the study.

### *B) Questions for scenario*

Table I shows the questions that all participants answered during the experiment. Some of the questions come from the set of questions actually asked by border guards at the border crossing point. However, many of these questions are not practical to ask in the experimental scenario. Therefore, a methodology was devised to substitute a minimum-sized set of proxy questions, which cover the same psychological / cognitive properties. This was found by analyzing a set of questions that were provided by experts from the Hungarian National Police Polish Border Guards, State Border Guard of the Republic of Latvia and TRAINOSE (Greece).

Table I: Experiment Interview Questions

| Question Number | Question |
|---|---|
| 1 | What is your family name? |
| 2 | What is in your case? |
| 3 | Have you seen any posters of prohibited items? |
| 4 | Are there any items from the lists of prohibited items in your case? |
| 5 | How many items are in the case? |
| 6 | If you open the case and show me what is inside, will it confirm that your answers were true? |
| 7 | What is your first name? |
| 8 | When were you born? |
| 9 | Where were you born? |
| 10 | What is your current citizenship? |
| 11 | Please tell me the name of a friend or family member who can confirm your identity? |
| 12 | What is the relationship of this person to you? |
| 13 | Where does this person live? |

### *C) Interview conducted using Wizard of Oz Methodology*

Collection of data to train the deception detection component of ADDS uses the established "Wizard of Oz" methodology. In this method (figure 4) a human, called the "Wizard" manually controls a simulated Avatar to create an experiment which is experienced (as closely as possible) by the participants as if they were being interviewed by a real Avatar. In this experiment, the Wizard operated a web app via Wi-Fi, which controlled the display on the participant's screen. The Wizard has access to a GUI allowing the selection of questions that are played in a window on the participant's screen. During the experiment:

- The participant aligned their face with the camera using on screen instructions.
- The simulated Avatar maintained a neutral expression
- The questions were delivered verbally to the participant by the static avatar through text to speech recordings.
- Video of the participant was captured on a question-by-question basis and stored for the purposes of training and testing.

The start time of the question is recorded as when the Avatar starts speaking. Once the participant has finished answering the question, the Wizard clicks to progress to the next question. The time of this click was recorded as the end of a question. The wizard also had the option of repeating a question if necessary.

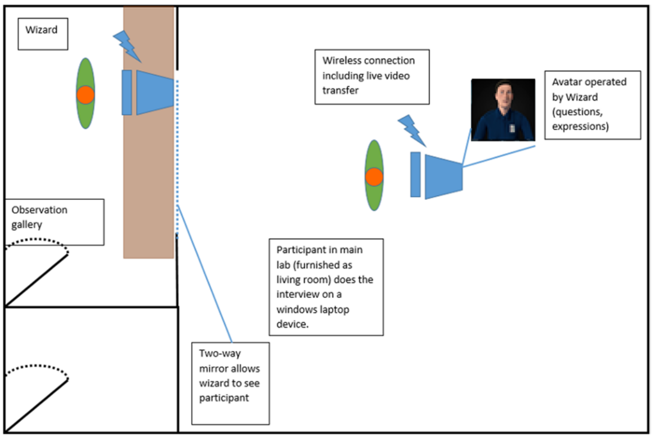
Wizard
Wireless connection including live video transfer
Avatar operated by Wizard (questions, expressions)
Observation gallery
Participant in main lab (furnished as living room) does the interview on a windows laptop device.
Two-way mirror allows wizard to see participant

Fig. 4. The Wizard-of-Oz (WoZ) experiment

### *D) Group design and stratification*

The video data of participants was recorded using the questions presented in Table I and are automatically cropped/segmented into question-by-question video files. Table II shows the dataset for truthful and deceptive participants. The data is captured using the built-in web-cam with the default video resolution of 640*480 and 30 frames per second (fps). The channel data is extracted from each question using a fixed sliding window (slot) of 1 second (i.e. 30 frames) to hold sufficient information of the channel states. Each slot is considered a single vector encoding the information/states for 38 channels. A sample size of 32 participants is used similar to other comparable studies[6].

Furthermore, the vector is considered only if it is extracted from a valid slot. A valid slot always consists of the channel information for face and both eyes. Detailed explanation of slot validity is explained in previous work [6].

Table II: Experimental Dataset

| | |
|---|---|
| No. of Question per Interview | 13 |
| Total Participants | 32 (17 Deceptive, 15 Truthful) |
| Total number of video files | 448 |
| Deceptive participants | Male (10)<br>Female (7),<br>Asian/Arabic (4)<br>EU White (13) |
| Truthful Participants | Male (12),<br>Female (3)<br>Asian/Arabic (6),<br>EU White (9) |
| No of Channel Analyzed | 38 |
| Total number of truthful vectors in dataset | 43051 |
| Total number of deceptive vectors in the dataset | 43535 |

Two strategies were employed for training, testing and validation: Bootstrapping and Leave one out which are described version in section V. These were used to develop suitable Artificial Neural Networks for the channels and stages shown in figure 2. The final classifying ANN (i.e. deception classifiers) produces an activation, which is thresholded to determine truthfulness or deceptiveness. This was not dynamic (i.e. continuously adaptive) but was optimized manually, based on the suitcase experiment data. Further manual optimization will be carried out after initial runs of the pilot. Deception risk scores obtained from the deception classifiers are bipolar in the range [-1, 1], where deceptive is +1 and truthful is -1. The deception score for a question is calculated as:

$$D_q = \frac{\sum_{s=1}^{n} d_s}{n} \tag{1}$$

Where, $D_q$ is the deception risk score of current question, $d_s$ is the deception score of slot *s* and *n* is the total number of valid slots for the current question. Then the thresholding is applied as stated below:

```
IF Question_risk ($D_q$) <= x THEN
        Indicates truthful
ELSE IF Question_risk ($D_q$) >= y THEN
        Indicates deception
ELSE
        Indicates unknown
END IF
```

Where x and y are determined empirically. Initial values were set too x = -0.05 and y = +0.05. Values between x and y are unknown and a classification cannot be obtained. This indicates that the final decision for the question classification (i.e. truthful/deceptive) is made regardless of the quantity of valid slots. In other words, a question is classified as unknown when there is no single valid slot that exists (i.e. n = 0) or the $D_q$ value lies in between x and y.

## V. Results and Findings

### *A) Initial Results*

The dataset obtained from the group channel coders is fed into the deception classifier that firstly used the 10-fold cross-validation strategy for train/validate/test the networks. The percentage split of the entire input data for training, validation and testing was 70:20:10 percent respectively. There are 38 inputs to the network with one hidden layer and single output. Networks are trained on varying number of neurons (i.e. 11-20 in these experiments) in the hidden layer to observe the impact on the performance. A bipolar sigmoid transfer function is used while training the networks. Maximum number of epochs are set to 10,000. The aim of this initial work is to establish whether a machine based interviewing technique can be used to detect deception from non-verbal behaviour, and there was no tuning of the classifiers attempted. With the exception of removing redundant duplicated vectors, the initial results presented in Table III and Table IV are derived from raw unprocessed data.

Table III shows the percentage accuracy of the deception classifiers obtained using a varying number of neurons in the hidden layer and the aforementioned parameter settings. It is observed that the network performance is gradually increased while increasing the number of neurons. The highest training/validation/test average accuracy indicated 96.55% and 96.78% in terms of truthful and deceptive classification respectively with 20 neurons in the hidden layer. The trained networks with the optimum classification accuracy are then further used for the testing on unseen dataset.

### *B) Testing classifiers*

The strategy used for testing the classifiers is based on leaving one pair out (i.e. one truthful and one deceptive participant) for testing while training and validating the networks on the rest of the participants' data (30 participants). Then the trained networks performance was tested using the unseen data of two participants. To examine the effect of totally unseen participants, 9 experimental runs were conducted, each involving the random selection of a pair of test participants (one truthful, one deceptive). Table IV shows the average test accuracy was measured to be 73.66% for deceptive tests and 75.55% for the truthful tests.

Table III: Results using 10 Fold Cross Validation

| No. of Hidden Layer Neurons | Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|
| | Training | | Validation | | Test | |
| | T | D | T | D | T | D |
| 11 | 94.13 | 95.04 | 93.68 | 94.41 | 94.30 | 93.69 |
| 12 | 94.45 | 95.63 | 93.75 | 94.74 | 93.62 | 94.96 |
| 13 | 94.92 | 95.77 | 94.41 | 95.14 | 94.31 | 95.15 |
| 14 | 94.85 | 96.26 | 94.29 | 95.67 | 94.23 | 95.46 |
| 15 | 96.19 | 96.19 | 95.40 | 95.50 | 95.50 | 95.41 |
| 16 | 96.16 | 96.40 | 95.58 | 95.80 | 95.45 | 95.91 |
| 17 | 96.56 | 96.98 | 95.90 | 96.32 | 95.75 | 96.22 |
| 18 | 96.81 | 97.17 | 96.14 | 96.52 | 95.91 | 96.28 |
| 19 | 97.23 | 97.11 | 96.48 | 96.48 | 96.67 | 96.45 |
| 20 | 97.28 | 97.50 | 96.53 | 96.81 | 96.55 | 96.78 |

These outcomes indicate a substantial decrease in the classification accuracy when compared with the classification outcomes presented in Table III. When using 10-fold cross-validation, the classifiers have seen *some* of the material (i.e. image vectors) from a test participant's interview in the training set (but training and test sets of vectors were mutually exclusive). Consequentially, the cross validation approach builds a model containing some of the psychological properties of the people who it classifies. In the second case (leave one pair out strategy), it sees no material of participants and relies on the commonality between their behavior and the behaviors of the participants used for training. We postulate that a large number of participants will build a larger general model, which will improve classification accuracy on previously unseen cases.

Table IV: Classification Outcomes using Unseen Participants

| Test No | Participant | | | | Accuracy (%) | |
|---|---|---|---|---|---|---|
| | Truthful | | Deceptive | | Truthful | Deceptive |
| | Gender | Ethnicity | Gender | Ethnicity | | |
| 1 | M | EU | M | A/A | 100 | 57 |
| 2 | M | A/A | F | EU | 50 | 36 |
| 3 | M | A/A | F | EU | 50 | 100 |
| 4 | M | EU | F | EU | 90 | 100 |
| 5 | M | A/A | M | EU | 100 | 10 |
| 6 | M | EU | M | EU | 72 | 100 |
| 7 | M | A/A | F | EU | 100 | 100 |
| 8 | F | EU | F | A/A | 38 | 100 |
| 9 | M | EU | M | EU | 80 | 60 |
| Overall Accuracy (%) | | | | | 75.55 | 73.66 |

It is also noted that for these initial experiments there is an insufficient amount of training data. Based on diversity of the participants (e.g. ethnicity, age, gender), a larger dataset would be more helpful to further generalize the classification networks. Despite of fair distribution of overall truthful and deceptive dataset (1.e. approximately 43000 vectors each), the unbalanced dataset in terms of ethnicity and gender might influence the deception classification network performance. For instance, the deceptive dataset consists of 4 Asian/Arabic participants compared to 13 of white EU. Likewise, in truthful scenario, there are 12 Male compared to only 3 female participants. In addition, the data used in this study was raw (apart from removal of redundant duplicated vectors), had not been preprocessed and no tuning of the ANN deception classifiers had taken place.

## VI. Conclusions and Further Work

This paper has described the first stage in development of an automated deception detection system (ADDS) which will be developed further to be utilized within the iBorderCtrl (Intelligent Portable Control System). An experiment was designed and conducted using a number of truthful and deceptive scenarios to test the hypothesis that a machine based interviewing technique could be used to detect deception from non-verbal behaviour during an interview conducted by a static avatar. The dataset collected for this experiment contained image vectors from 32 participants and contained diversity in terms of gender and ethnicity. Raw experimental participant data was used to train artificial neural network deception classifiers using two train-test strategies. The un-optimized networks gave (as expected) high results when utilizing a cross validation train-test strategy, whilst obtaining an average classification of 75% on both truthful and deceptive interviews when using a leave a pair out strategy. It was noted that given the diversity of the dataset, it might not have been large enough to train a classifier more effectively. Future work will involve capturing more data for diverse population representation and optimization of the neural network classifiers

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700626. The authors would like to thank the iBorderCtrl consortium members for their feedback in developing ADDs in this project.

## References

[1] European Parliament. (2016). Smart Borders: EU Entry/Exit System. Brussels: European Parliament.

[2] iBorderCtrl Intelligent Portable Control System [online], Available at http://www.iborderctrl.eu/ [Accessed 12/1/2018],

[3] Crockett, KA and O'shea, J and Szekely, Z and Malamou, A and Boultadakis, G and Zoltan, S (2017) Do Europe's borders need multi-faceted biometric protection. Biometric Technology Today, 2017 (7). pp. 5-8. ISSN 0969-4765

[4] Hall, J. A. (2007) 'Nonverbal cues and communication.' In Baumeister, R. F. and Vohs,K. D. (eds.) Encyclopaedia of Social Psychology, California: SAGE Publications Inc., pp. 626-627.

[5] Holmes, M. Latham, A. Crockett, K, O'Shea, J. Near real-time comprehension classification with artificial neural networks: decoding e-Learner non-verbal behaviour, IEEE Transactions on Learning Technologies, Year: 2017, Volume: PP, Issue: 99, DOI: 10.1109/TLT.2017.2754497.

[6] Rothwell, J., Bandar, Z., O'Shea, J. and McLean, D., 2006. Silent talker: a new computer□based system for the analysis of facial cues to deception. Applied cognitive psychology, 20(6), 757-777.

[7] Silent Talker Ltd [online], Available at: https://www.silent-talker.com/ [Accessed 5 Jan. 2018]

[8] International League of Polygraph Examiners (2016), Polygraph/Lie Detector FAQs. [Online]. Available at: http://www.theilpe.com/faq_eng.html. [Accessed 16/01/2018].

[9] Mangan, D.J., Armitage, T.E. and Adams, G.C., (2008). A field study on the validity of the Quadri-Track Zone Comparison Technique. Physiology & behavior, 95(1), pp.17-23.

[10] Honts, C.R. and Reavy, R., (2015). The comparison question polygraph test: A contrast of methods and scoring. Physiology & behavior, 143, pp.15-26.

[11] Saxe, L., Dougherty, D. and Cross, T., (1985). The validity of polygraph testing: Scientific analysis and public controversy. American Psychologist, 40(3), 355.

[12] Meijer, E.H., Verschuere, B., Gamer, M., Merckelbach, H. and Ben□Shakhar, G., (2016). Deception detection with behavioural, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. Psychophysiology.

[13] Walczyk JJ, Igou FP, Dixon AP, Tcholakian T. Advancing lie detection by inducing cognitive load on liars: a review of relevant theories and techniques guided by lessons from polygraph-based approaches, Frontiers in Psychology, 4, 01 February 2013, [online] Available at: http://dx.doi.org/10.3389/fpsyg.2013.00014 [Accessed 16 Jan. 2018].

[14] Eriksson, A. and Lacerda, F., (2007). Charlatanry in forensic speech science: A problem to be taken seriously. International Journal of Speech, Language and the Law, 14(2), pp.169-193.

[15] Herms, R., (2016). Prediction of Deception and Sincerity from Speech using Automatic Phone Recognition-based Features. Interspeech 2016, pp.2036-2040.

[16] Ekman, P., (2016). Paul Ekman International Plc. [online] Available at: http://www.ekmaninternational.com/ [Accessed 18 December 2016].

[17] Cassell, J., 2001. Embodied conversational agents: representation and intelligence in user interfaces. AI magazine, 22(4), p.67.

[18] Nunamaker, J.F., DErrICk, D.C., Elkins, A.C., Burgoon, J.K. and Patton, M.W., 2011. Embodied conversational agent-based kiosk for automated interviewing. Journal of Management Information Systems, 28(1), pp.17-48.

[19] Hooi, R. and Cho, H., 2013. Deception in avatar-mediated virtual environment. Computers in Human Behavior, 29(1), pp.276-284.

[20] Ströfer, S., Ufkes, E.G., Bruijnes, M., Giebels, E. and Noordzij, M.L., 2016. Interviewing suspects with avatars: Avatars are more effective when perceived as human. Frontiers in psychology, 7

[21] de Borst, A.W. and de Gelder, B., 2015. Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. Frontiers in psychology, 6, pp.576.

[22] Derrick, D.C., Read, A., Nguyen, C., Callens, A. and De Vreede, G.J., 2013, January. Automated group facilitation for gathering wide audience end-user requirements. In System Sciences (HICSS), 2013 46th Hawaii International Conference on (pp. 195-204). IEEE.

[23] Pollina, D.A., Horvath, F., Denver, J.W., Dollins, A.B. and Brown, T.E., 2009. Development of technologies and test formats for credibility assessment. Polygraph, p.99.

[24] Stremble Ventures LTD, [online] Available at: http://stremble.com/ [Accessed 5 Jan. 2018]