Defence Research and Development Canada
Recherche et développement pour la défense Canada

DRDC | RDDC
technologysciencetechnologie

Design and Evaluation of Biometric-enabled Interview Assisting Traveller Screening Technology

Dmitry Gorodnichy
Canada Border Services Agency

Prepared by:
Canada Border Services Agency
280–14 Colonnade Rd.
Ottawa, ON K2E 7M6
Canada

2008 Public Security Technical Programs MOU
DRDC Contact: Brian Greene, Defence Scientist, DRDC – Centre for Security Science.
Contractor's date of publication: October 2018

Defence Research and Development Canada

Contract Report
DRDC-RDDC-2018-C223
November 2018

Canada

IMPORTANT INFORMATIVE STATEMENTS

This document was reviewed for Controlled Goods by Defence Research and Development Canada using the Schedule to the *Defence Production Act*.

Disclaimer: This document is not published by the Editorial Office of Defence Research and Development Canada, an agency of the Department of National Defence of Canada but is to be catalogued in the Canadian Defence Information System (CANDIS), the national repository for Defence S&T documents. Her Majesty the Queen in Right of Canada (Department of National Defence) makes no representations or warranties, expressed or implied, of any kind whatsoever, and assumes no liability for the accuracy, reliability, completeness, currency or usefulness of any information, product, process or material included in this document. Nothing in this document should be interpreted as an endorsement for the specific use of any tool, technique or process examined in it. Any reliance on, or use of, any information, product, process or material included in this document is at the sole risk of the person so using it or relying on it. Canada does not assume any liability in respect of any damages or losses arising out of or in connection with the use of, or reliance on, any information, product, process or material included in this document.

Canada Border Services Agency
Agence des services frontaliers du Canada

Science and Engineering Directorate

Border Technology Division
Division Report 2018 – 04 (TR)
October 2018

Design and Evaluation of Biometric-enabled Interview Assisting Traveller Screening Technology

CSSP-2015-TI-2158 Study
Final Report

Dr. Dmitry O. Gorodnichy

PROTECTION • SERVICE • INTEGRITY

Canada

Version #: 1.0		
Action	Name	Date
Prepared By	Dmitry O. Gorodnichy	2018-04-03
Reviewed By	Mehdi Tabib	2018-09-14
Reviewed By	Alain Bourgon	2018-10-05
Reviewed By	Phil Lightfoot	2018-10-22
Final By	Dmitry O. Gorodnichy	2018-10-25

Abstract

This report presents the deliverables for the “Roadmap for Biometrics at the Border" (CSSP-2015-TI-2158) study conducted by the Canada Border Services Agency (CBSA) in partnership with University of Arizona (UA) and San Diego State University (SDSU) through support from the Defence Research and Development Canada, Canadian Safety and Security Program (CSSP). The main objective of this study was to generate critical knowledge related to the use of biometric-enabled Interview Assisting Traveller Screening (IATS) technology, such as AVATAR kiosks developed by UA and SDSU. The deliverables include: overview of manual behaviour screening limitations, overview of challenges related to designing biometric-enabled behaviour screening (BEBS) systems, development of a novel framework for designing and evaluating BEBS systems, conducting a mock-up experiment with the AVATAR kiosk at the CBSA, and recommendations based on the insights gained from the conducted evaluations.

Acknowledgements

This work was done under the funding from the Canadian Safety and Security Program (CSSP) managed by the Defence Research and Development Canada, Centre for Security Science (DRDC-CSS), in partnership with University of Arizona (UA) and San Diego State University (SDSU). Other members of the project included Aaron Elkins (from SDSU), Judee Burgoon, Elyse Golob, Jay Nunamaker, and Bradley Walls (from UA). Their feedback and contributions are gratefully acknowledged and are presented in separate technical papers.

The feedback and assistance from the Center for Identification Technology Research (CITeR) and, in particular, from the CITeR Director, Stephanie Schuckers, in setting and managing the project between the CBSA, UA and SDSU are also gratefully acknowledged.

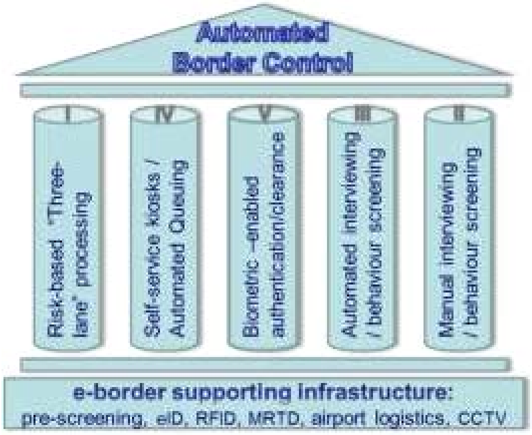
Finally, the help of former frontline officers Marie-Eve Boulanger and Moe Mohtadi, in preparing the interview questions for AVATAR, feedback from Traveller Transformation Program Advisor, Kevin Kearney, and help of many other of CBSA employees in conducting the AVATAR mock-up test in March 2016 are also gratefully acknowledged.

Dedication

This report is dedicated to Diane Keller, the first Director General of the CBSA's Science and Engineering Directorate (2006-2017), whose dedication to science and scientists made it possible for the Agency to become an internationally recognized player in applying Biometric and Video Analytics research to border applications.

Diane Keller with AVATAR during the AVATAR test at the CBSA in March 2016.

1 Introduction


Automated
Border Control
I
IV
V
III
II
Risk-based "Three-lane" processing
Self-service kiosks / Automated Queuing
Biometric –enabled authentication/clearance
Automated interviewing / behaviour screening
Manual interviewing / behaviour screening
e-border supporting infrastructure:
pre-screening, eID, RFID, MRTD, airport logistics, CCTV

The precursor "ART in ABC" study (CSSP-2013-CP-1020) [1] established *five* key border modernization traveller screening components, *Pillars*, that will define automated border control (ABC) of the future:

I. "three-level" risk processing,
II. manual interviewing and behavior screening,
III. automated interview-assisting behaviour screening,
IV. automated queuing / self-service kiosks,
V. biometric-enabled authentication.

In relationship to Pillar III (automated interview-assisting behaviour screening), it has been shown that even very well trained officers, as in the US SPOT program [4], have difficulty correctly identifying "suspicious" behaviors. There is enough evidence to suggest that humans, knowingly or unknowingly, will always be vulnerable to "human bias" in making their decisions, due to their cultural, religious, gender, and other differences. It appears therefore that, with existing practices, manual interviewing and behavior screening has reached the plateau limit in its efficiency.

To address this limitation of human performance in behavior screening, the researchers from University of Arizona, started to develop AVATAR (short for Automated Virtual Agent for Truth Assessments in Real-Time) – the kiosk aimed at automating behaviour screening [7-10]. The AVATAR and other *interview-assisting traveller screening (IATS)* technologies are seen as a way to scale and further improve the interviewing and behavior screening of incoming travellers.

Following the recommendations from the "ART in ABC" study, the CBSA took the lead to further explore the readiness of IATS technologies for ABC applications. Through support from the Defence Research and Development Canada (DRDC) Canadian Safety and Security Program (CSSP), it established a new project with the objective to generate the critical knowledge related to IATS technologies and to develop practical recommendations related to further testing and deployment of these technologies in Canada. For the purpose of achieving the objectives, a new partnership with the researchers from the University of Arizona was established.

The project ran from September 2015 to December 2017. The key outcomes of this project are presented in this report, while more detailed results can be found in separate technical paper [2]. The report is organized as follows.

Section 2 addresses the limitations of manual screening of travellers. Section 3 provides background behind automated detection of stress and deceit. Section 4 describes the AVATAR technology and the mock-up experiment conducted at the CBSA in order to demonstrate and test the technology. Section 5 presents the methodology that was developed for designing and evaluating AVATAR-like IATS technologies. Results from testing the AVATAR kiosk are presented in Section 6. The concluding section summarizes the limitations and advantages of IATS technologies and presents recommendations related to testing of these technologies in the field.

2 Manual behaviour screening

2.1 Two parts of the challenge

With over a million travellers entering the country daily and over hundreds of border officers who have to assess the risks associated with each traveller, the risk of not being able to efficiently validate person's credibility can be high. This section aims at providing the information that may allow the agency to minimize this risk.

Behaviour screening needs to be understood as a *two-sided problem*. The first side of the problem relates to the limitation of human abilities in detecting lies. According to the UA scientists, who have over 40 years of research experience in deceit detection and behavior screening, poor performance affects both novices and professionals, with accuracy of detection ranging from 45% (novices) to 65% (professionals). Furthermore, research evidently shows that confidence in judgment is not correlated with accuracy (Correlation coefficient < 0.05). The summary of human limitations and typical biases and misconceptions related to lie detection is provided in Table 2-1.

The other side of the problem relates to the cultural and mental diversity of humans. As highlighted in the precursor study, manual behavior screening may lead to wrong decisions due to human error with respect to individuals who have anxiety or other mental health conditions, the percentage of whom is estimated as 20% of the household population (according to the Canadian Mental Health Association). Furthermore, travellers are commonly already under stress due to travel-related challenges and frequently come from different, possibly unknown, cultural backgrounds, which makes them even more vulnerable to wrong decisions with respect to their behaviour.

Table 2-1 Typical human biases, limitations, and misconceptions related to lie detection.

Typical human biases	Human limitations	Typical misconceptions
• Truth bias - Tendency to assume all tell truth - Common among lay-people • Othello (lie) bias - Tendency to assume all are lying - Common among law enforcement personnel • Cultural / religious bias	• Limited ability to view all signals (e.g., pupil dilation, heart beat) • Limited capacity to analyze multiple cues at a time (normally, fewer than five) • Attention required for other tasks (watching people, luggage, computer screens, etc.) • Overconfidence, which does not correlate with quality • Prone to misconceptions	– Gaze aversion – Nervous gesturing – Preening

3 Automated behavior screening

3.1 Deception signals and four benefits of computerizing their detection.

Table 3-1 shows the signals that are believed to be related to lying. Computerized recognition of these signals is seen as a way to address challenges of manual screening described above. In particular, the following four benefits are expected from using automated behavior screening for automated border control:

- Improving accuracy (less false hits and less misses),
- Alleviating the "human bias",
- Allowing scalability of screening solutions,
- Making credibility assessment auditable.

The latter will make it possible to know on which grounds, i.e., because of which behaviour signals and actions, the credibility of a traveller is questioned. This may help to improve the service quality.

Table 3-1 Deception signals, sensors that can measure them, level of difficulty in their detection.

Deception signals categorized by biometrics modalities	**for humans to detect**	**sensors that can be used**	**for machine to capture**	**for machine to recognize**
Oculometrics:				
-Pupil size dynamics / change ↑ -Eye movement -Blink patterns	-hard -hard -medium	eye-tracker, video-camera	easy	medium
Kinesic signals:				
Liars are more tense / less expressive (fewer illustrators) -Micro-facial expressions -Body movements (head, hands, legs, torso) • Posture, Stance, proximity • Shifts & rigidity • Initial freeze response • Finger fidgeting • Hand to face adaptors • More lip presses	-medium -medium	video-camera, eye-tracker, gyro sensor (e.g., on tablet in person's hands), weight sensor	easy easy	hard hard
Physiometrics:				
-Body / face temperature ↑ -Brain activity ↑ -Heart rate ↑ -Respiratory patterns ↑	- hard, impossible	camera, IR-camera, special sensors	medium	hard
Auditory signals - Vocalics:				
Voice Quality↓ -Harmonics-to-noise ratio decreases Pitch, Tempo, Intensity↑ • Fundamental frequency ↑ • Change in pitch ↑ • Tension ↑ Response Latency↑ Disfluency↑	-easy-medium -easy-hard (depends on training)	microphones	easy	medium
Auditory signals - Linguistic:				
-What is being said (context, logic, consistency) -How something is being said (sentiment, choice of vocabulary) • Sample message features • Average sentence length • Passive voice ratio, Emotional content, Word diversity	easy-medium	microphones	hard hard	hard easy

↑ / ↓ *indicate the increase / decrease of the feature value with increase of stress.*

3.2 *Factors contributing the development of the technology*

Recognition of a person's attributes (such as facial expression, fatigue, motion patterns) from measurable data has been a popular research topic in academia for several decades with over a thousand research papers published on the topic yearly. Over the past decade however this research area has received a particular boost, due to the following three factors:

i) Sensors have become ubiquitous, affordable and diverse. For example, standard smart-phones now have at least five sensors inside: camera, microphone, GPS tracker, proximity tracker and gyroscope tracker,

ii) Many third-party open-source software packages have been developed to extract various numerical measurements from these sensors,

iii) Many open-source machine learning packages, including deep neural networks, have been developed, are available for free and are widely supported

While major advancements have been made in developing BEBS over the past decade, many challenges remained, which are summarized next.

3.3 Stress detection: two tasks of the problem and key questions

To appreciate the complexity of automated signals detection work, one needs to understand that for a computer, as opposed to a human, the “Detection” is, in fact, a two-task problem. The first task deals with being able to *measure* the signals (“Registration” problem). The second task deals with *interpreting* the measured signals (“Recognition” problem”). While the “Measuring” problem is generally easier for a computer than for a human, the “Recognition” problem may not (see Table 3-1). Therefore, the following research questions are critical for the development of BEBS technology for ABC:

For the “Measuring” problem:

- Which sensors can be used to capture deception signals?
- Which of these sensors can be used in ABC applications?

For the “Recognition” problem:

- How to convert measured raw signals into features?
- Which of features are useful for automated recognition of deceit and which are not?

These questions need to be addressed taking into account that, besides technical challenges (such as how to measure the signals under the application and how to build the recognition algorithm), there are also CELP challenges – related to Cultural, Ethical, Legal and Privacy issues, as discussed in our past projects [5].

In theory, any number of sensors can be used for BEBS. In practice however one needs to investigate which ones are efficient and which are not. Table 3-1 lists sensors that have been mentioned in literature for the purpose of measuring various biometric signals. Some of these sensors have been used for many years in interrogations using polygraphs (where detected signals are analyzed by a qualified interrogator). Some others have not been used with general public yet and therefore require more intense CELP-related effort to allow the use in the field.

Additionally, it should be realized that, while in laboratory settings some signals are easier to measure than others, in the operational settings (because of the application constraints), these signals may not be well measured, i.e. with precision required for their recognition.

4 AVATAR kiosk at the CBSA

4.1 Four advantages of AVATAR kiosk

The AVATAR kiosk is one of the most known BEBS implementations. Figure 4-1 shows its main components. The main competitive advantages of AVATAR kiosks compared to other work done in automated emotion sensing are the following:

- Realistic animated recreation of the virtual interviewer (see the figure) capable of imitating normal human facial reactions such as blinking and grimacing, which is very important in creating the much desired impression of very powerful and very intelligent machine on a traveller;
- Effective and robust design that automatically adjusts cameras to eye level, which allows capturing of faces and eyes at good quality at all times;
- Highly configurable code for modifying interview scripts;
- Many years of building, tuning and testing the system, including during the current project with the CBSA.

It is noted that the UA team who developed the AVATAR holds two US patents related to the development of IATS / BEBS systems.

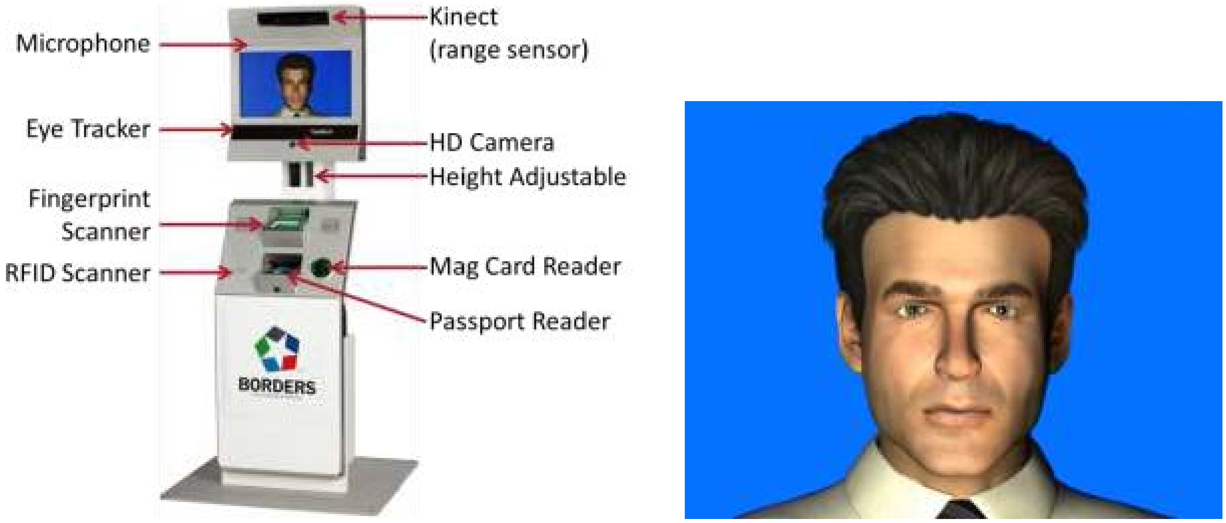


Figure 4-1 AVATAR kiosk design and computer-animated AVATAR virtual agent.

4.2 Test design

In March 2016, through the joint effort of UA, SDSU and the CBSA, an AVATAR kiosk was brought to the CBSA Science and Engineering Lab for a prototype testing exercise. Over eighty volunteers from the CBSA and other government and academic partner organizations participated in this exercise, making it possible to collect what is now the largest data-set for the research in automated behaviour screening for ABC applications. All personal identifiable information (PII) was removed from the collected data, which were then analyzed using a new evaluation framework described in the next section. No real travellers, operational systems or data have been used at any point throughout the experiments.

The test was designed to evaluate the AVATAR technology in two behaviour screening problems: detecting “imposters” and detecting “smugglers”. Each test participant was issued a plastic machine-readable ID card and was asked to pack a travel bag (see Figure 4-2). Half participants played imposters/smugglers (hereafter referred to as “Liars”); they were asked to hide a contraband item (shown on the bottom right in Figure 4-3) in their bag and had false information on their ID cards (false first name and false age). The other half of participants played regular travellers (hereafter referred to as “Non-liars”). A video describing the test has been prepared and can be provided upon request.

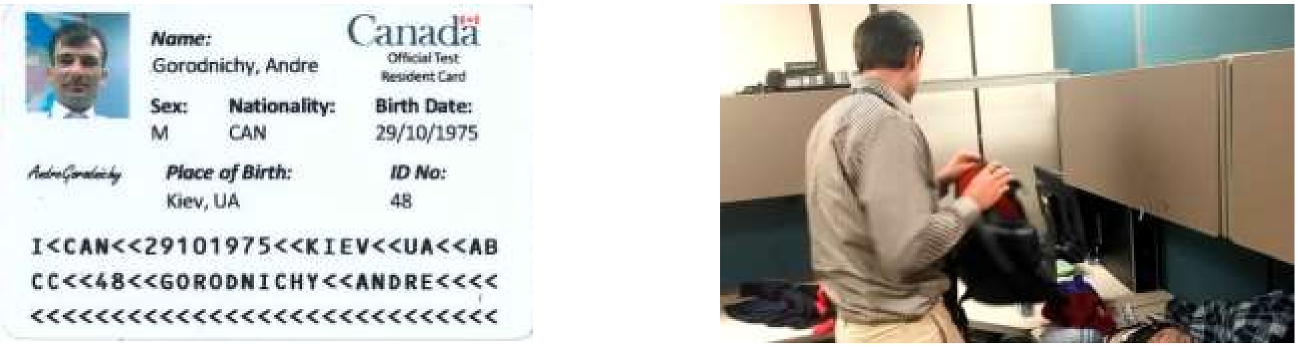
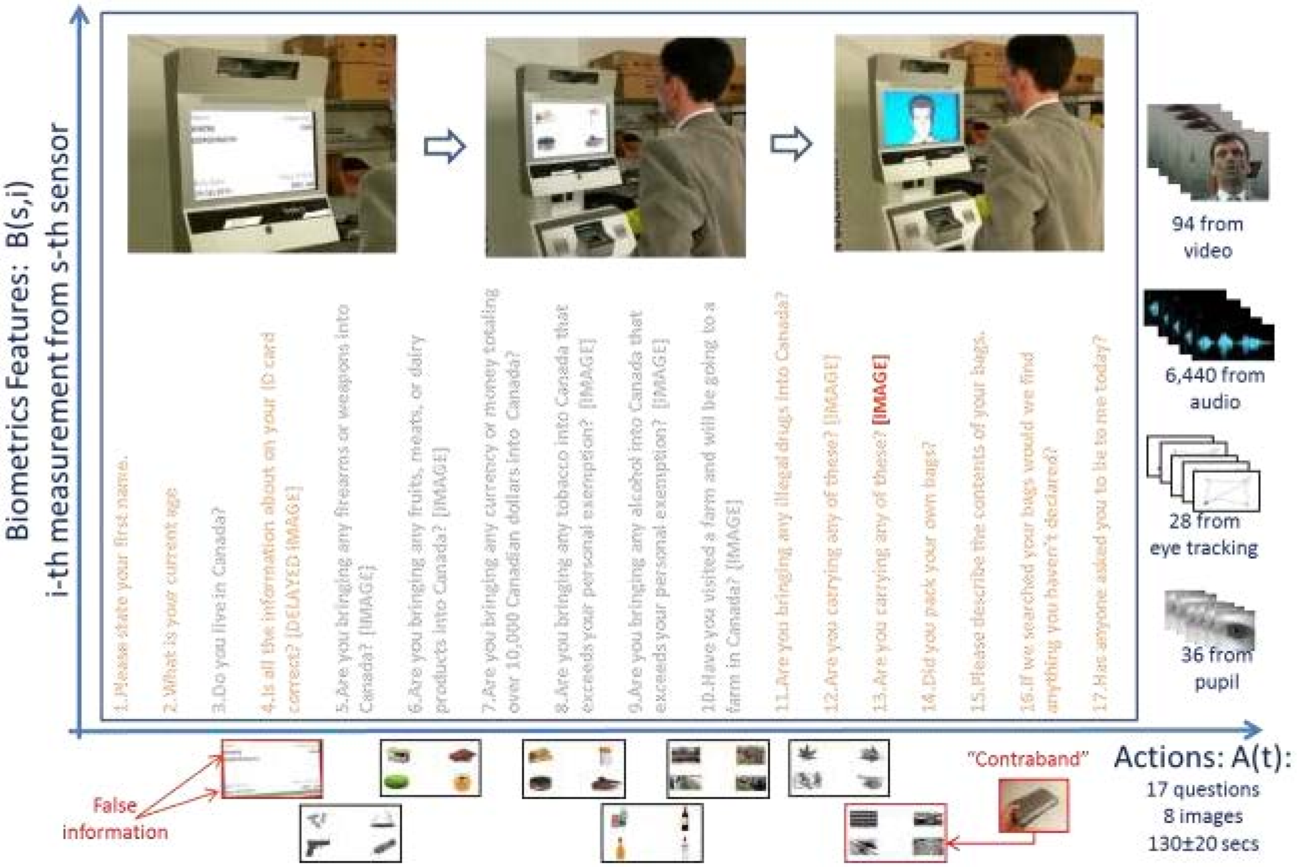


Figure 4-2 Plastic ID card issued for a test participant. A participant packs his backpack with travel items.

An interview script that was prepared for the AVATAR is conceptualized in Figure 4-3. The horizontal axis shows the actions that a traveller had to perform. These actions were the answers to 17 questions asked by the AVATAR agent during an automated two-minute interview. Questions where Liars needed to lie are marked red. Eight questions were supplemented with images shown at bottom: one showing the person's ID information and seven showing the images related to custom declaration questions. Users were asked to look at those images to confirm their answers to the questions.

The vertical axis shows biometric features (person's facial emotions, voice characteristics, eye tracking information and pupil dynamics) that were obtained using the sensors (video camera, microphone, eye tracker) installed in the kiosk.



Biometrics Features: B(s,i)
i-th measurement from s-th sensor
1.Please state your first name.
2.What is your current age
3.Do you live in Canada?
4.Is all the information about on your ID card correct? [DELAYED IMAGE]
5.Are you bringing any firearms or weapons into Canada? [IMAGE]
6.Are you bringing any fruits, meats, or dairy products into Canada? [IMAGE]
7.Are you bringing any currency or money totaling over 10,000 Canadian dollars into Canada?
8.Are you bringing any tobacco into Canada that exceeds your personal exemption? [IMAGE]
9.Are you bringing any alcohol into Canada that exceeds your personal exemption? [IMAGE]
10.Have you visited a farm and will be going to a farm in Canada? [IMAGE]
11.Are you bringing any illegal drugs into Canada?
12.Are you carrying any of these? [IMAGE]
13.Are you carrying any of these? [IMAGE]
14.Did you pack your own bags?
15.Please describe the contents of your bags.
16.If we searched your bags would we find anything you haven't declared?
17.Has anyone asked you to lie to me today?
94 from video
6,440 from audio
28 from eye tracking
36 from pupil
False information
"Contraband"
Actions: A(t):
17 questions
8 images
130±20 secs

Figure 4-3 Graphical representation of the AVATAR experiment at the CBSA. See [1] for details.

4.3 Features measured

Third party software was used to extract features from the data recorded by the sensors. First, raw statistical measurements, as registered directly by the sensor, are obtained. Then these measurements were processed to produce the total of 19 higher-order statistical measurements (*functionals*) per question, such as mean, standard deviation, range, quartiles, up-level times, regression slope, intercept and others. In such a way, for each action (i.e., each of 17 asked questions, eight of which were supplemented by images), the following numbers of features were computed for each person during each interview:

- 5*19*17 facial emotion features – obtained from HD (30FPS 720p) video-camera: 19 functionals for five facial emotion measurements (Happy, Surprised, Neutral, Disgusted, Sad) computed for each of 17 questions using the Intraface [12];
- 6440*17 vocalic features – obtained from microphone: include basic acoustic measures such as principle frequency (f0), response latency, intensity, etc.) extracted from raw audio using Praat software [13] and Interspeech Computational Paralinguistics Challenge acoustic features extracted from raw audio using the OpenSmile software [14];
- 2*19*17 pupil and proximity data - obtained from eye tracker: 19 functionals for two measurements (pupil size and distance to camera) registered by eye tracker, computed for each of 17 questions ;
- 4*7 eye fixation measurements – obtained from eye tracker: computed from raw eye tracking data for each of 7 questions with images, indicating how long a person looked at each of four quadrants in the image.

This resulted in a total of over a million biometric and non-biometric measurements recorded for each person at each interview. One of the main contributions of the project was to show that, out of this large quantity of measurements, it is only a very small fraction (between 20 and 40 measurements) that are actually useful. Other measurements present either noise, irrelevant or highly correlated (repeated) measurements. If they are not discarded when building a system, they may lead to erroneous conclusions about the person's credibility.

In addition to the above listed measurements, the system has also provided a complete log of eye tracking and face tracking data, which is over 350,000 vertical and horizontal coordinate measurements recorded at 60 Hz by the video camera and eye tracker for each interview session. These tracking data however were not used for building the system during the duration of the project.
More details on the features used by AVATAR and the full analysis of their value for the deceit detection task are available in separate technical papers.

5 Methodology for designing and evaluating BEBS systems

5.1 Limitations of ISO biometric evaluation standards

Current biometric standards, such as those developed by the International Standards Organization (ISO) Sub-Committee on Biometrics (SC-37) and presently used by industry and academia [6], are limited to the use of biometric data for identification purposes only (such as 1-to-1 identity verification and 1-to-N identity search). As a result, no formal guidelines have been developed to date for evaluating systems that use biometric data for behaviour screening applications. In order to evaluate BEBS technology, and the AVATAR kiosk in particular, we had to develop new BEBS-related terminology and a BEBS evaluation methodology.

In this regard, a list of new biometric terms related to BEBS applications has been developed and submitted to International Standard Organization (ISO), presented in a separate report [3], and a novel framework for designing and evaluating such systems has been developed, further described below.

5.2 BEBS systems vs. traditional biometric systems

The key difference of BEBS systems from traditional biometrics comes from the fact they rely on biometric features that, when used individually, are of very low discriminating power. It is shown that the distributions of matching scores obtained from voice for “Liars” vs. “Non-liars” are only very slightly different from each other. This is in sharp contrast to traditional biometrics, where density distributions of opposing classes are very different and well separated.

The reason for such low discriminating power of features in BEBS is that they are compared not to the features of the same person (as done in traditional biometrics), but rather to some “averaged” features (models) representing their class (Liars). These models are computed based on some historical data and the previous knowledge of deception signals described in Section 3, and they may not be assumed to be very precise.

To compensate for the low discriminating power of features, BEBS systems use a large quantity of those features, accumulating them over a period of time (e.g., over two minutes at a 20 Hz sampling rate, as done in AVATAR). This is illustrated in Figure 4-3, where the total number of measured features (represented by a large perimeter box in the figure) is very large - equal to the number of all measurements from all sensors B(s.j), shown along vertical axis, times the number of all actions (responses to interview questions) sampled over time A(t), shown along the horizontal axis.

Such large accumulation of features in BEBS systems further differentiates such systems from traditional biometrics, where recognition is made from only a few biometric features and mostly from a single action (such as looking into iris camera). The result of such accumulation is that most of features do not contain information related to lying, and those features that are related to lying may not be known in advance.

Based on these two critical differences of BEBS from traditional biometrics, the principles for designing and evaluating BEBS systems are developed, as summarized below.

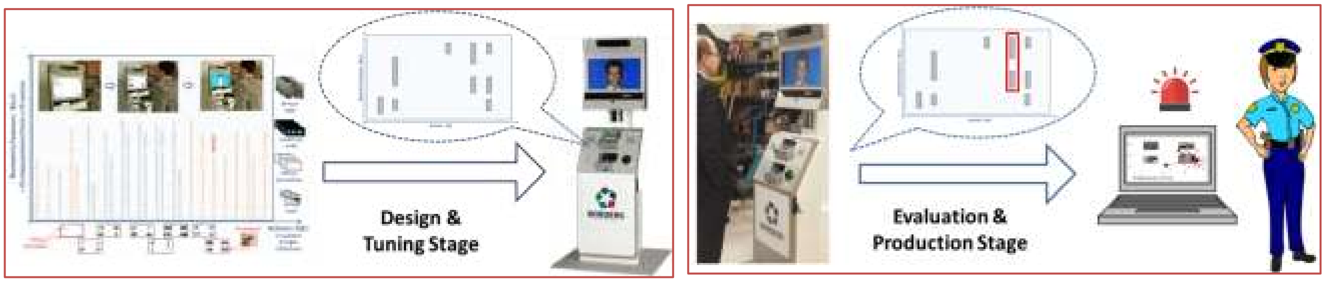


Figure 5-1 Key principles for building and testing BEBS systems.

At the Design stage, best feature-actions are identified (shown as small grey rectangles inside a large rectangle).
At the Evaluation stage, the decision is made in such a way that it does not generate many false alarms and is easy to interpret by an officer, e.g., by showing features that have been found to be abnormal (marked red, and shown on the officer’s laptop screen).

5.3 Four design principles

The main four principles for designing BEBS systems are defined as follows.

1. *Design objective:* Identify Actions and Biometric features that provide the best statistically quantifiable discrimination between lie and non-lie behaviours (shown as small grey rectangles inside large perimeter boxes in Figure 5-1);

2. *Performance metrics:* Use metrics suitable for detection of low frequency events, as done in Video Analytics applications. Specifically, instead of (or in addition to) measuring the percentage of True Positives (detected Liars) with respect to All Positives (all Liars) – referred to as Accuracy, Recall or True Positive Rate (TPR) that are commonly used in Biometrics, one should also measure the percentage of True Positive with respect to All Alarms, which is referred to as *Precision*. See Table 5-1 for more details.

3. *Criteria for success:* Measure the success in terms of likelihood to improve the status-quo recognition rates, as done in Clinical Trials, rather how this is done in traditional Biometrics;

4. *Operator-centred design:* Design recognition models and detection visualization interfaces that are easily interpretable and efficient for humans (see Figure 5-1).

Table 5-1 Confusion matrix and performance metrics used in Biometrics and Video Analytics.

Metrics used in biometrics for frequent event detection systems		**Actual Liars (42)**	**Actual Non-liars (40)**
Predicted Liars (Alarms)	For computer (52)	TP = 22 **TPR** = 22/42 = 52%	FP = 30 **FPR** = 30/40 = 75%
	For human (15)	TP = 10 **TPR** = 10/42 = 24%	FP = 5 **FPR** = 5/40 = 13%
Predicted Non-liars	For computer (30)	FN = 20 **FNR** = 20/42 = 48%	TN = 10 **TNR** = 10/40 = 25%
	For human (67)	FN = 32 **FNR** = 32/42 = 76%	TN = 35 **TNR** = 35/40 = 87%

Metrics used in video analytics for rare event detection systems	**Precision ≡ TP / All Alarms**	**Recall ≡ TPR≡ TP / All Liars**
For computer	22/52 = 42 %	22/42 = 52%
For human	10/15 = 67%	10/42 = 24%

Numbers are provided for illustration purposes only and should not be used as a reference on system/human performance.
For computer, a sample of results from AVATAR testing is used.
For human, the results are hypothesized based on visual observations of interviewees by the author of this report.

TPR (also called Recall and Accuracy), which stands for True Positive Rate, can be used for evaluation of both rare and frequent event detection systems. In contrast, FPR and FNR, which stand for False Negative Rate and False Positive Rate respectively, should not be used in evaluation of frequent event detection systems.

6 Experimental Results

Full results and insights that have been obtained from the evaluation of the AVATAR kiosk will be presented in a separate technical paper [2], which is also where the images in this section are taken from. A summary of these results is presented below.

6.1 Interpretable model results

Decisions Trees offer a simple and easy to interpret way for designing and evaluating intelligent systems. Figure 6-1 shows the best achieved accuracy rates for the AVATAR lie detection system built using Decision Trees. It is based on two voice features only: one at question #2 (about age) and the other at question #16 ("would we find anything you have not declared?").

Observations:

- A simple Decision Tree-based model achieves True Positive Rate of 56 % (32%+ 24%, shown in the blue boxes at the bottom) at the cost of True Negative Rate of 44 % (shown in the green box, at the bottom of the decision tree).
- It is clear from this model that not all features and questions contribute to lie detection.

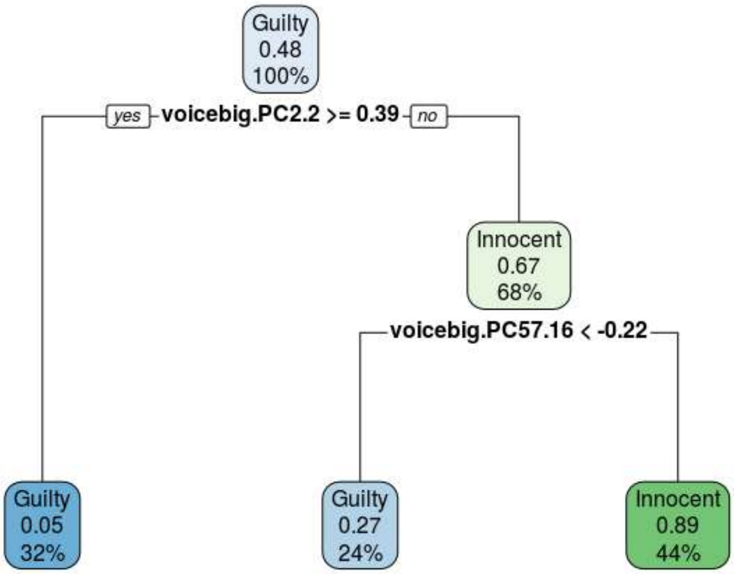


Figure 6-1 Accuracy rates obtained using an interpretable model (decision tree). Features that are used to make the decisions are shown.

6.2 Baseline results

Figures 6-2 and 6-3 show results obtained using four more advanced machine learning techniques. These techniques are commonly treated as "black boxes". Their decisions may not be understood or interpreted by humans, however their performance is expected to be better than that of a simpler interpretable model such as a decision tree shown above.

The models are built without feature optimization, i.e., using all measured features, regardless of whether they are related to lying or not. In this sense, the results obtained by these models should be considered as baseline results, upon which the system performance may potentially be further improved, should an effort be applied to select better features, as per the system design principles described in previous section.

Based on these curves, the baseline accuracy metrics for AVATAR, for each modality and combined, are obtained as presented in Table 6-1. For comparison, the table also provides Accuracy (a.k.a. Recall) results for manual lie detection by humans, which is estimated based on the information provided by UA scientists for the conditions that are similar to those used in the AVATAR test, i.e., when recognizing a lie from a two-minute interview using 17 custom-declaration questions. The table also lists advantages and disadvantages of each modality and manual lie detection.

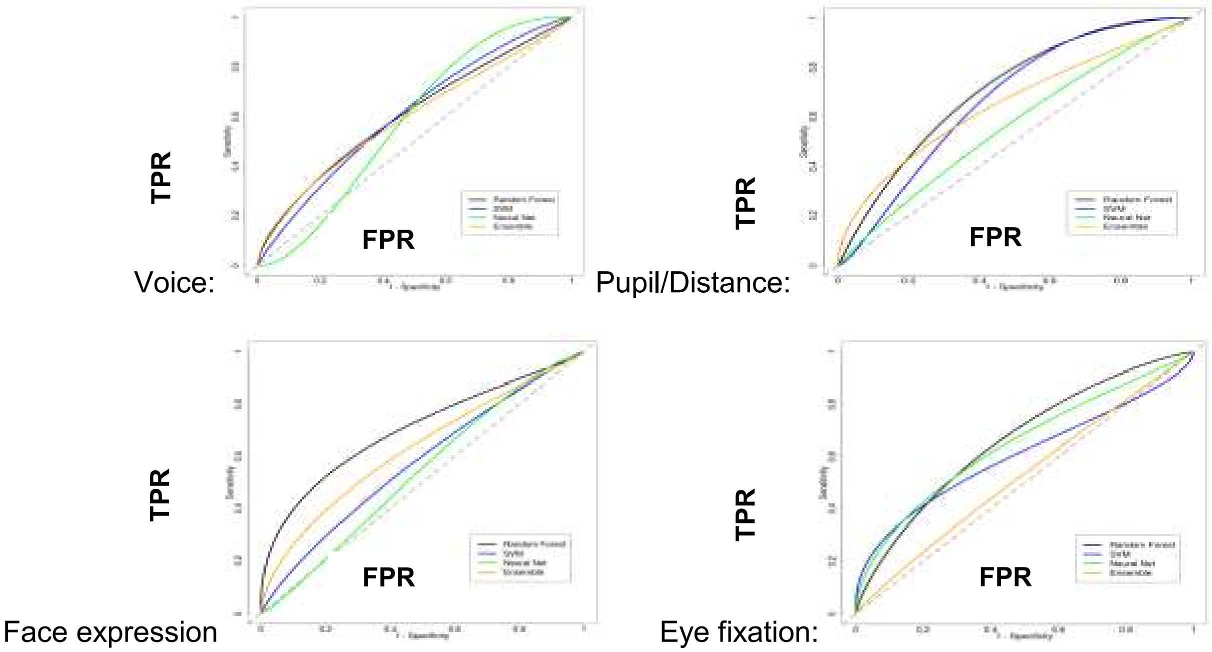


Figure 6-2 Baseline AVATAR performance - by modality, obtained using four different machine learning techniques, without feature optimization. Accuracy, marked as True Positive Rate (TPR), is shown as a function of False Positive Rate (FPR).

Observations:

- The graphs show that each modality contributes to detecting a lie - the predictions made by each of four models for each modality are correct in over 50% of cases, i.e., it is better than a random guess (shown as dashed line) . The variation of performance by model is seen. There's no one-fit-all model.
- For a combined system, it is also seen that at Recall of 60% (marked by dashed red vertical line), the Precision of up to 75% was achievable; similarly, at Precision of 60% (marked by dashed red horizontal line), the Recall in lie detection of up to 80% was achievable.
- The constructed predictive models are not easily interpretable by humans. There is no information there on the features and questions that help recognizing the deceit and that do not.

Table 6-1 Baseline AVATAR performance, limits and constraints

	With Microphone: Vocalics	**With Eye Tracker: Pupil & distance to camera**	**With Eye Tracker: Eye Fixation**	**With HD camera: Face emotions**	**All combined**	**Humans***
Recall at 60% Precision	55-78%	54-65%	60-65%	48-58%	65-68%	45-65%
Dis-advantages	Privacy issues. Requires good acoustics, absence of other voices.	Eyes need to be visible at all times	Calibration required	Lowest performance	Needs technical expertise & support	Can't monitor many features at a time. Need training. Hard to audit and scale.
Advantages	High performance	No PII recorded. Calibration not required. High performance	No PII recorded	Easiest to capture. Calibration not required.	Scalable, auditable, bias-free	Do not require procuring and maintaining equipment

**Estimated for the same conditions (i.e., using a two-minute interview and 17 custom declaration questions) by University of Arizona scientists based on over 40 years of research experience in deceit detection and behavior screening.*

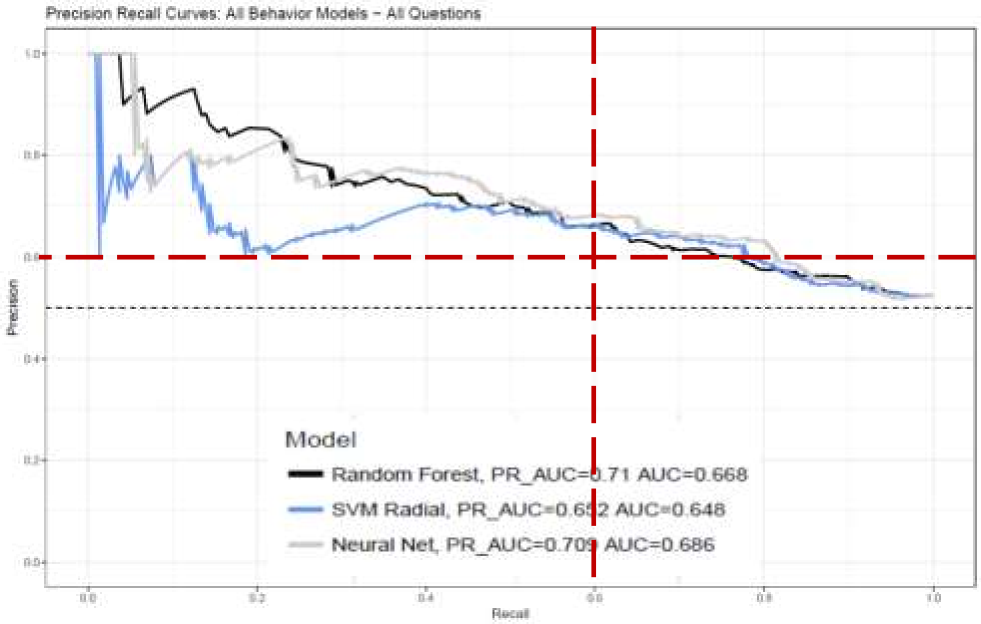


Figure 6-3 Baseline AVATAR performance - aggregated (all modalities combined), obtained using four different machine learning techniques, without feature optimization. Precision is shown as a function of Accuracy, marked as Recall in the graph.

6.3 Accuracy by question

A major effort of the study was put in analyzing the value of questions asked by AVATAR. A result of this analysis is presented in Figure 6-4. The figure shows lie detection accuracy obtained separately for each question.The accuracy is measured in terms of True Positive Rate (TPR), True Negative Rate (TNR) and Combined Accuracy Rate (Tcombined), which is the average between TPR and FPR.

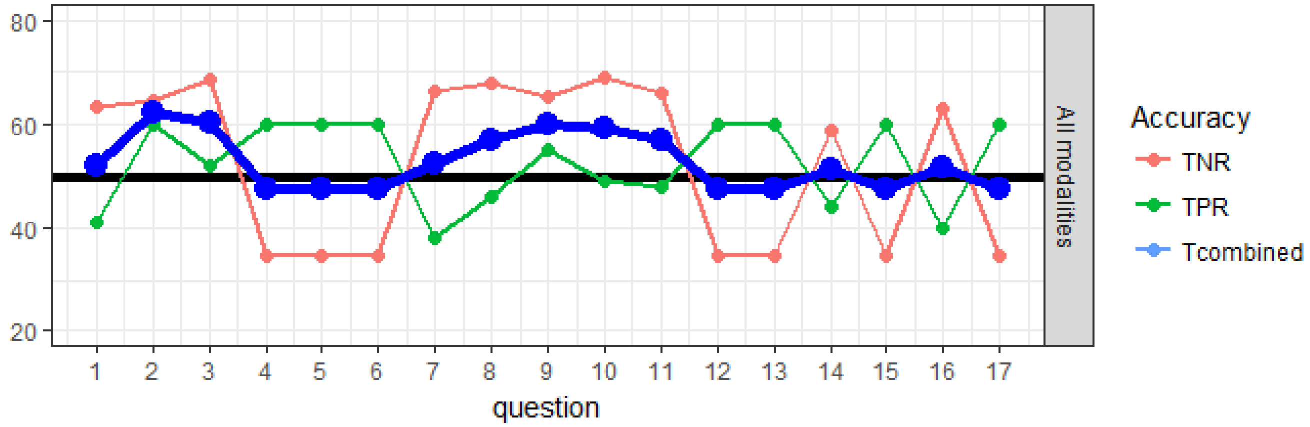


Figure 6-4 Accuracy by question, combined over all modalities.

Observations:

- It is seen that some questions help detecting lies more than others, with several of them having Accuracy over 60 %. The best-performing question is the one about age (Question #2) .
- At the same time, it is also seen that some questions do not help at all, having the prediction accuracy less than 50%, i.e., worse than flipping a coin.

- It is also found that besides guilty-knowledge-related questions (where smugglers have to lie), other questions (where smugglers do not need to lie) may also contribute to deceit detection.

6.4 Best and worst deception indicators

The key scientific contribution of this study relates to the analysis of features to be used as deceit indicators. Up till the present, the knowledge available on this subject has been obtained manually, prone to misconceptions (discussed in Section 2) and not validated or quantified by an automated behaviour screening system. Certain physiological body reactions (deceptions signals) have been linked to lying (summarized in Table 3-1), however it was not known prior to this study how those signals can be measured and when they should be measured. For example, should they be measured when a person is lying or when s/he is telling the truth but has some guilty knowledge? This study allowed us to answer this question.

In a typical interview, there will always be multiple questions asked to a person who is being screened for a bad intention. Some of these questions will not require a person to lie, like those from AVATAR interview marked blue in Figure 4-3. Others will do, like those marked orange in Figure 4-3. Through the analysis of features measured by the AVATAR on 82 test participants, it has been found that these two different sets of questions cause two very different, almost disjunctive, reactions in liars. Many features that are good indicators of guilty knowledge at the first type of questions are very bad indicators if used during the second type of questions, and vice versa.

This is illustrated in Figure 6-5 which shows the result of voice features analysis done by applying a common two-sample statistical hypothesis test [15] and a more scientifically rigorous approach called mixed-effects regression analysis [16]. The bars shows the probability of the causal relationship between various voice features and being a liar when replying to questions that do not require persons to lie, and between the same features and being a liar when replying to questions that require persons to lie, computed using the two types of analysis. A summary of modality specific results follows.

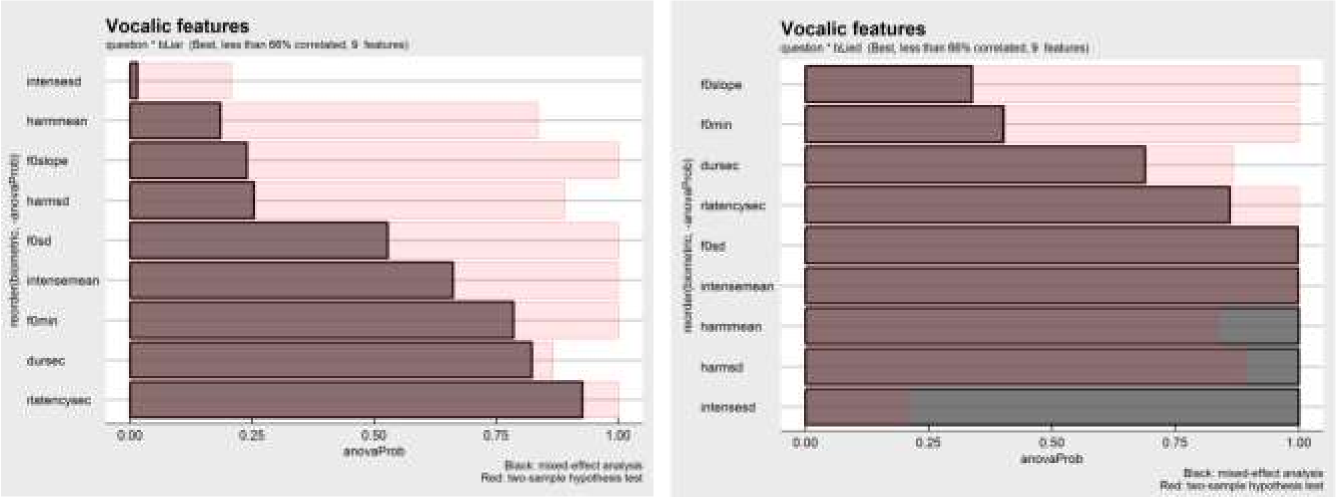


Figure 6-5. Best and worst deceit indicators in voice modality: for questions that do not require a person to lie (left) and for questions that require him/her to lie (right). The bar length indicates the probability that a feature can be used as a deceit indicator.

Results obtained using a simple approach (two-sample statistical hypothesis test) are shown as red bars. Results obtained using a more advanced approach (mixed-effects regression analysis) are shown as dark bars. The more advanced approach uncovers the previously unknown best and worst deceit indicators.

Observations:

- In the voice modality (see Figure 6-5), the feature *intensesd* (which signifies the variation in voice intensity while responding to a question) was found to be the most powerful for detecting lying (with P>0.9). It is however a very poor discriminator when a person does not lie (P < 0.1). At the same time, this is the opposite for voice features *dursec* (which signifies the delay between the question and response) and *f0min* (which is the primary voice frequency), which are found to be good for flagging liars when answering questions where they did not need to lie (P=0.85 and P=0.80), but worse for detecting the actual moment of lying (P=0.7 and P=0.32).

- In the pupil dynamics modality, *pupilrange* (which is the difference between the minimum and maximum pupil size observed during a question) is seen highly valuable for detecting the moment of lying (P>0.9), but not at all when a liar does not lie (P < 0.1), in which case *pupilskew* (which describes the shape of the dynamics in pupil change during the response) is found most useful (P>0.95).

- The proximity features (which characterize the dynamics of the distance between the person and the kiosk) have not been found indicative of lying (all of them have low probability in relationship to either lying or being a liar).This disproves the earlier belief that such distance matters.

- The importance of using more advanced techniques (such as a mixed-effect regression used in this study) in analyzing AVATAR systems is also seen. Simpler approaches (such as previously used two-sample statistical hypothesis tests) could lead to opposite conclusions on the value of various features for lie detection.

These new insights on how a person with guilty knowledge reacts on different questions will assist both the systems designers and human interrogators.

7 Conclusions

7.1 Limitations of the study

The results obtained from the AVATAR test at the CBSA appear better than that by humans in identifying liars. However the following limitations of these results need to be highlighted:

1. They are obtained on a very small population size (82 volunteers who participated in the experiment). This is a very small size to be able to extrapolate them to any large size population with any statistical significance.
2. These results are obtained, after several months of tuning the algorithms, following several algorithm tuning iterations. Only the best performing algorithms are shown in this report. In real deployment, once the system is deployed, it may not be tuned.
3. They are obtained on the same dataset that was used to tune the algorithm. As mentioned in Section 5, this creates an “optimistic” bias in the reported accuracies.
4. Volunteers had no advance knowledge of the system. They were not allowed to see the kiosk prior to their interview with it. In real life, this may not be expected from people who come prepared to lie.
5. Furthermore, volunteers were actors, rather than real smugglers / impersonators. Their motivation and skills in lying is questionable. Results of this experiment may not represent real world results.
6. Volunteers were also guided by a study team member, who helped them to start the interaction with the kiosk and who was by their side at all times to resolve any technical difficulties related to operating the kiosk.
7. Results are obtained for a particular scenario and setup that was developed for this test. This scenario involved the same actions for all non-compliant travellers (hiding the same contraband item and falsifying their name and age). They may be quite different for another scenario or setup.

8. In general, evaluation of lie detection technologies should never be expected to be precise, because in real life one should never expect to label correctly the data used for training the system, i.e., knowing exactly who is a liar and who is not.
9. Finally, the study did not consider any CELP (Cultural, Ethical, Legal, Privacy) issues related to the use of this technology. These may potentially impose additional constraints on the choice of sensors and data collected, further limiting the accuracy of system performance.

Despite a seemingly large number of limitations listed above, this study has identified several ways to further improve the performance of BEBS systems, and make it less dependent on a particular scenario or setup. With another test (or pilot in the field) it should be possible to further improve the accuracy results, using the methodology and new insights presented in this report.

At the same time, the study has also highlighted the fact that simply adding more sensors or more questions does not necessarily improve the recognition results, which is a popular “the more, the better” stereotype. In fact, this can make them worse, if features are not properly selected and filtered out.

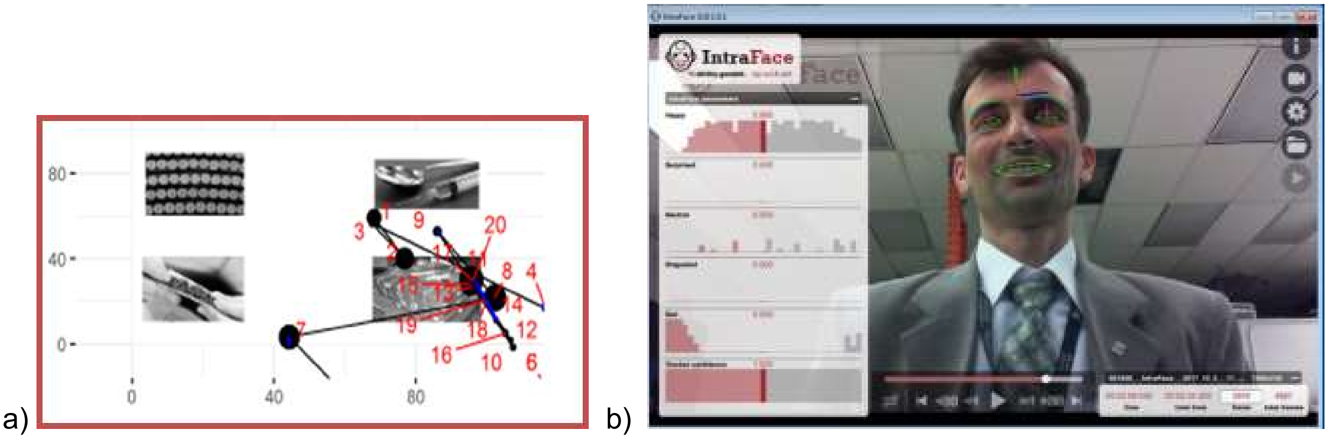


Figure 7-1. Through the eyes of AVATAR: Visual Interfaces showing the result of eye tracking (a), and facial analysis (b) performed by the kiosk.

7.2 Additional insights

In addition to the recognition accuracy metrics presented in this report, a number of other results important for the development and deployment readiness assessment of AVATAR-like technologies have been obtained, summarized below:

- **Use of eye tracking data:** Despite claims in the literature, full automated analysis of eye tracking has not been shown useful yet. Furthermore, eye-tracking devices require user-specific calibration, which may not be possible in real life. At the same time, it is shown that eye tracking data can be made easily interpretable by humans using *Visual Interfaces* (as shown in Figure 7-1.a). Such visualization of features may allow officers to see deception signals that are otherwise not visible to naked eye. For example, if a person was attracted by an object shown on a screen and his/her pupil dilated at the same time, an officer may be able to use this information to ask additional questions.

- **Use of facial cameras**: Similarly, full automated analysis of facial information has not been shown useful yet either. At the same time, it is found that tracking results obtained from a regular HD video camera are comparable to that of commercial eye trackers and they do not require calibration (Figure 7-1.b). Taking into account that cameras are already embedded in many ABC kiosks, this makes such cameras an attractive alternative to eye trackers which suffer from calibration issues.

- **Use of iris camera:** iris cameras, such as those used in ABC kiosks, have been found to not be capable of streaming eye tracking data and therefore cannot be presently used for behaviour screening purposes.

- **The importance of an animated agent:** The presence of an animated agent (shown in Figure 4-1) is highlighted as an important factor for the success of the technology. First, it creates an important impression on clients who will be able to relate to AVATAR as to a real and very powerful (AI-equipped) border officer. Second, the research on visual attention and saliency showed that a blinking face of AVATAR will subconsciously attract and control the direction of person's view. This can help to calibrate eye tracking, resulting in less noise in tracking data, potentially further improving the system accuracy.

- **Other sensors and modalities:** In addition to the sensors currently used in AVATAR (shown in Figure 4-3), there are a number of other sensors readily available that can also be potentially embedded into the technology to further improve its accuracy and versatility. Of particular interest are gyroscopes (on portable tablets that travellers may hold in their hands) and weight sensors (such as those used in some e-gate systems). Finally, eye blinking patterns (which can be reliable extracted from video) are not used yet but appear very promising.

7.3 Technology readiness

Using the semaphore-like PROVE-IT technology readiness assessment methodology developed in previous projects [5], the Technology Level Readiness (TRL) of the BEBS systems is assessed as being "yellow", i.e., ready for piloting in the field.

Compared to human performance, BEBS technology is shown to offer an increase in the likelihood of detecting deceit. Even though the observed increase may not appear large, it does make this technology very attractive for applications, such as ABC, where a large number of subjects need to be screened and where human factors such as fatigue and bias need to be minimized.

One may expect these results to improve with time as more sensors and biometric features are added. However, as this study demonstrated, it is critical to remember that if not properly designed and tested, adding new sensors and biometric features, may also lead to worsening of the system performance. This is why the development of good standards and building scientific expertise in the area of automated biometric-enabled behavior screening by industry and government stakeholders will be critical for the further development of these systems.

To recapitulate, high scientific integrity will need to be exercised prior to deployment of such systems in the field, because of the elevated risk of falsely flagging travellers due to technology limitations and the absence of public standards for the evaluation of such technology. The principles of ethically designed AI systems, which are being currently actively discussed within IEEE community [17], will need to be followed when designing and deploying BEBS systems.

7.4 Other considerations

Automated screening systems offer benefits beyond just better Precision / Accuracy rates. They contribute to better transparency, scalability and integrity of ABC decisions. The kiosks that are already used for ABC can do more than what they are doing now. With little modification to their hardware, they may automatically extract the information related to the credibility of people in front of them, thus assisting border officers to make better decisions.

Offering an exciting and pleasant travel experience for people entering the country and being seen by the public as a champion driving technological advances to serve better its clients is another factor not to be discounted, when considering the advantages of AVATAR-like systems.

However there is yet another important factor which we would like to mention in the conclusion of this study. It has to do with the psychological power of people. Using parallels with clinical trials made earlier, it can be called a placebo effect. Just like with medication, when people believe that the technology works, it may work better for them. That is, just by seeing a new and powerful system in front of them, travellers will likely behave in a different and more pronounced way, which may allow the machine and human officers standing by to detect something, which otherwise may not be detected. Extrapolated to large traffic of travellers processed by the Agency, even a very small increase in probability of catching someone will result in dozens of additionally caught smugglers and terrorists.

Table 7-1 Applications potentially suitable for BEBS:
Using the semaphore-like technology readiness assessment methodology, these applications are assessed as being "yellow", i.e., ready for piloting in the field.

	Application constraints	Sensors that can be used
Frontline applications		
1. Standalone automated (as within PIK)	**< 3 mins, No audio Operated by travellers**	**Video-camera + eye-tracker in kiosk**
2. Semi-automated (as a tool in PIL booth)	**No eye tracking, Operated by BSO**	**microphone + video-camera in booth**
3. At Secondary examination	**None**	**Video-camera + microphone + eye-tracker**
4. Remote ports	**VoIP audio quality telecom video quality**	**Microphone + eye-tracker**
Inland applications		
5. Interviews for trusted traveller applications	**None**	**All combined**
6. Interview for officer recruitment	**None**	**All combined**
7. Self-reporting over the phone	**Audio only Lower quality**	**Telephone microphone**

7.5 Recommendations for next steps

For the Agency, the following possibilities for further testing of the technology are seen. A good opportunity is seen in using the no-voice IATS / BEBS as part of the next-generation Primary Inspection Kiosks (PIK). Kiosks, such as those shown in Table 7-1, are already capable of automatically aligning cameras to the traveller's face, which is the main condition for automated behavior screening. This means that these kiosks may potentially be also programmed to perform certain traveller screening tasks based on the captured facial data. The deceit detection accuracy by such kiosks can be expected to be at least comparable to that of humans. If additionally an eye tracking sensor is installed in such kiosks, then deceit detection accuracy can be expected to be higher than that of humans.

IATS / BEBS can also be used as a tool for manual screening at PIL booths, where vocal and facial signals can be automatically detected using cameras and microphones installed in the booths and shown to the officers for their information.

Other opportunities for the use of IATS / BEBS are seen in pre-screening applications (such as for trusted travellers applications and staff recruitment) and post-screening applications (such as in secondary examination), where voice recorders can be used and where longer interviews are permissible.

IATS / BEBS may also be potentially suitable for remote screening applications (as in self-reporting over a telephone or at remote unmanned ports of entry), provided that audio and video signals are of sufficient quality.

Critically, it should be mentioned that in all of these applications, IATS / BEBS may be used in either automated or semi-automated mode. In the first case, the signals detected by sensors are interpreted by the machine. In the second case they are interpreted by humans.

To conclude, IATS / BEBS technology is shown to be suitable for further testing in a variety of border control applications, with remaining challenges being mainly technical - related to properly tuning and evaluating the technology to be deployed. By establishing the methodology for designing and evaluating biometric-enabled behaviour screening technology, this project helps to address these challenges.

References

1. Dmitry O. Gorodnichy, Analysis of Risks and Trends in Automated Border Control: CSSP-2013-CP-1020 Final Report, http://cradpdf.drdc-rddc.gc.ca/PDFS/unc256/p804885_A1b.pdf. Executive Summary, http://cradpdf.drdc-rddc.gc.ca/PDFS/unc229/p803869_A1b.pdf
2. Aaron Elkins, Dmitry O. Gorodnichy, Judee Burgoon, Elyse Golob, Jay Nunamaker, Bradley Walls, Design and Evaluation of Biometric-enabled Credibility Assessment Systems, Submission to IEEE journal/conference (in preparation)
3. Dmitry O. Gorodnichy, New age glossary of biometrics terms for automated border control and video surveillance applications, CBSA Border Technology, Division Report 2016-05
4. “ACLU sues TSA over behavior screening program,” USA TODAY, March 19, 2015 http://www.usatoday.com/story/news/2015/03/19/aclu-tsa-behavior-detection-spot/24974421
5. D. Bissessar, E. Choy, D. Gorodnichy, T. Mungham, “Face Recognition and Event Detection in Video: An Overview of PROVE-IT Projects” , Technical Report DRDC-RDDC-2014-C167, http://cradpdf.drdc-rddc.gc.ca/PDFS/unc157/p800402_A1b.pdf
6. Dmitry O. Gorodnichy, Diego Macrini, Robert Laganiere, “Video analytics evaluation: survey of datasets, performance metrics and approaches ", Technical Report DRDC-RDDC-2014-C248. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc198/p800521_A1b.pdf
7. AVATAR: The Interrogation Bot, Wired, February 2013.
8. Deception detection, American Physiological Association, March 2016, Vol 47. No 3. www.apa.org/monitor/2016/03/deception.aspx
9. Jay F. Nunamaker, JR., Judee K. Burgoon, Aaron C. Elkins, Mark W. Patton, Douglas C. Derrick, Kevin C. Moffitt, Embedded Conversational Agent-Based Kiosk for Automated Interviewing, Filed: January 30, 2013, Publication date: October 10, 2013
10. Nathan W. Twyman, Jay F. Nunamaker, Automated Scientifically Controlled Screening Systems (ASCSS), Filed: May 27, 2015, Publication date: May 4, 2017
11. IntraFace, free research software for facial image analysis: www.humansensing.cs.cmu.edu/intraface
12. ISO/IEC 2382-37:2012, Information Technology “Vocabulary: Part 37: Biometrics.” Free copies at http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html
13. Praat: doing phonetics by computer: http://www.fon.hum.uva.nl/praat/
14. The openSMILE feature extraction tool. The Munich Versatile and Fast Open-Source Audio Feature Extractor: http://audeering.com/technology/opensmile/.
15. Wilcoxon signed-rank test, https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test
16. B. Bolker,, et al (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution, 24(3), 127---135.
17. Ethically Aligned Design, A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2: IEEE http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

DOCUMENT CONTROL DATA

*Security markings for the title, authors, abstract and keywords must be entered when the document is sensitive

1. ORIGINATOR (Name and address of the organization preparing the document. A DRDC Centre sponsoring a contractor's report, or tasking agency, is entered in Section 8.)

Canada Border Services Agency
280–14 Colonnade Rd.
Ottawa, ON K2E 7M6
Canada

2a. SECURITY MARKING
(Overall security marking of the document including special supplemental markings if applicable.)

CAN UNCLASSIFIED

2b. CONTROLLED GOODS

NON-CONTROLLED GOODS
DMC A

3. TITLE (The document title and sub-title as indicated on the title page.)

Design and Evaluation of Biometric-enabled Interview Assisting Traveller Screening Technology

4. AUTHORS (Last name, followed by initials – ranks, titles, etc., not to be used)

Gorodnichy, D.

5. DATE OF PUBLICATION
(Month and year of publication of document.)

October 2018

6a. NO. OF PAGES
(Total pages, including Annexes, excluding DCD, covering and verso pages.)

24

6b. NO. OF REFS
(Total references cited.)

17

7. DOCUMENT CATEGORY (e.g., Scientific Report, Contract Report, Scientific Letter.)

Contract Report

8. SPONSORING CENTRE (The name and address of the department project office or laboratory sponsoring the research and development.)

DRDC – Centre for Security Science
NDHQ (Carling), 60 Moodie Drive, Building 7
Ottawa, Ontario K1A 0K2 Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)

CSSP-2015-TI-2158

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. DRDC PUBLICATION NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)

DRDC-RDDC-2018-C223

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

Division Report 2018 – 04 (TR)

11a. FUTURE DISTRIBUTION WITHIN CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

Public release

11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

12. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Use semi-colon as a delimiter.)

Biometrics; Border Security

13. ABSTRACT/RÉSUMÉ (When available in the document, the French version of the abstract must be included here.)

This report presents the deliverables for the “Roadmap for Biometrics at the Border" (CSSP-2015-TI- 2158) study conducted by the Canada Border Services Agency (CBSA) in partnership with University of Arizona (UA) and San Diego State University (SDSU) through support from the Defence Research and Development Canada, Canadian Safety and Security Program (CSSP). The main objective of this study was to generate critical knowledge related to the use of biometric-enabled Interview Assisting Traveller Screening (IATS) technology, such as AVATAR kiosks developed by UA and SDSU. The deliverables include: overview of manual behaviour screening limitations, overview of challenges related to designing biometric-enabled behaviour screening (BEBS) systems, development of a novel framework for designing and evaluating BEBS systems, conducting a mock-up experiment with the AVATAR kiosk at the CBSA, and recommendations based on the insights gained from the conducted evaluations.