Thermal Facial Analysis for Deception Detection

Bashar A. Rajoub, *Member, IEEE*, and Reyer Zwiggelaar

***Abstract*— Thermal imaging technology can be used to detect stress levels in humans based on the radiated heat from their face. In this paper, we use thermal imaging to monitor the periorbital region's thermal variations and test whether it can offer a discriminative signature for detecting deception. We start by presenting an overview on automated deception detection and propose a novel methodology, which we validate experimentally on 492 thermal responses (249 lies and 243 truths) extracted from 25 participants. The novelty of this paper lies in scoring a larger number of questions per subject, emphasizing a within-person approach for learning from data, proposing a framework for validating the decision making process, and correct evaluation of the generalization performance. A *k*-nearest neighbor classifier was used to classify the thermal responses using different strategies for data representation. We report an 87% ability to predict the lie/truth responses based on a within-person methodology and fivefold cross validation. Our results also show that the between-person approach for modeling deception does not generalize very well across the training data.**

***Index Terms*— Automated deception detection, behavioural analysis, facial analysis, thermal imaging.**

I. INTRODUCTION

VARIOUS studies have shown that both ordinary people and trained experts are poor at discriminating between liars and truth tellers [1], [4] and for an average person performance is only slightly better than chance [9]. Empirical evidence indicates that differences between cognitive processes will often make liars experience a different mental state than truth tellers [17]. Liars may experience feelings of guilt, anxiety, anger, disgust, fear, and shame more often than truth tellers [8].

Polygraph technology involves various contact sensors to measure changes in blood pressure, respiratory, cardiovascular, and electrodermal activity [11], [35], [40]. Pavlidis et al. reported that on average, the accuracy of polygraph examination is around 90% for adequately controlled specific-incident tests [31], [33]. However, manual analysis takes time and makes the outcome expert dependent [24]. Processing of a typical 10-minute interrogation session using polygraph technology may take several hours [37]. The speed limitation of the polygraph and the need for trained experts to run polygraph testing motivated researchers to investigate automated credibility assessment using non-invasive technology.

Behavioural and facial-based approaches for detecting deception are unobtrusive and do not require the subject's cooperation. Automated deception detection is based on the assumption that guilt manipulation results in measurable physiological and/or behavioural changes that distinguish truth tellers from deceivers [15], [16], [27], [32], [38]. Thermal imaging [22] offers a non-contact approach of measuring physiological features like blood flow [32], pulse rate [36], blood vessels distribution [2], and breathing rate [13]. Thermal cameras capture the surface skin temperature which can vary due to emotion-specific bio-physiological states in the human body. As such, facial thermal analysis could potentially provide reliable cues to deception because controlling emotions can be difficult [9], [12], [41].

Thermal imaging techniques have been studied to detect deception in mock-crime scenarios [30], [31], [33]. Both instantaneous and sustained stress conditions can be detected using thermal imaging since instantaneous stress brings about an increase in the periorbital blood flow while sustained stress is associated with elevated blood flow in the forehead [29]. Classifying responses as deceptive or truthful can be achieved by finding relevant patterns in the input features using statistical or machine learning techniques [18], [33]. Some of the reported classification accuracies were as high as 87% [38] and 91.7% [33].

In this paper, we investigate the potential use of thermal facial analysis to detect deception based on information gathering interviews. The novel aspects of this paper lie in adopting a robust methodology for learning the model of deception based on a larger number of test questions per person compared to existing work. We also propose a framework for validating the feature extraction and the decision making process in order to evaluate the robustness of the developed approach. We begin by presenting a review of closely related work, then cover the framework for extracting thermal features and learning the baseline for classifying deception. We also present our deception detection experiment where deception is designed around a learnt-story (i.e. based on character profiles). This design considers various aspects from the theories on deception, in particular, we exploit cognitive load [44] by requiring the participants to plan their lies before the test and by asking questions not being covered by the profile. This requires the subject to extend their lies beyond the learnt story, and as such, increasing cognitive load. We also adopt the expectancy violations theory [6] by using a within-subject approach for detecting deception. In other words, rather than looking for

Manuscript received September 19, 2013; revised January 7, 2014; accepted March 26, 2014. Date of publication April 14, 2014; date of current version May 8, 2014. This work was supported by the Engineering Physical Sciences Research Council under Grant EP/G004137/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Siome K. Goldenstein.

B. A. Rajoub is with the Centre for Applied Digital Signal and Image Processing, University of Central Lancashire, Lancashire PR1 2HE, U.K. (e-mail: balrjoub@uclan.ac.uk).

R. Zwiggelaar is with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K. (e-mail: rrz@aber.ac.uk).

This paper has supplementary downloadable material at http://ieeexplore.ieee.org, provided by the authors. The material consists of two movies that show the tracker's consistency in estimating the eye corner positions throughout the video. The material is 16.1 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2014.2317309

the deception cues that are shared between humans in a given experiment, our research attempts to find cues of deception in relation to one specific person based on their own behaviour/baseline. This proved very significant and highlights the fact that people respond differently even to the same stimuli in similar situations (see Section III for details).

II. Background and Related Work

Deception detection includes contributions from human physiology, psychology, behaviour, human factors, and machine intelligence. Behavioural and facial-based approaches predict changes in the body's internal states due to stress caused by attempted control [26]. Attempting to conceal the truth results in an emotional response which might produce measurable physiological characteristics, like blushing [19], [46]. Deception might also produce measurable changes in gaze behaviour which indicates recognition of the visual stimuli [17] by causing the eyes to almost instantly move to the visual stimulus without any cognitive effort [39].

Initial work by Pavlidis et al. investigated the potential of using thermal imaging for anxiety detection [32]. They noted that all subjects under stress had a sudden (less than 0.3 sec) increase in blood flow around the eyes that was independent of face or eye movement and cooling over the cheeks and warming over the carotid. The mean temperature in the nasal area remained the same. These changes reverted to their pre-startle state within about 1 min [32].

A. Simulating Guilt and Lying

Cues to deception exist at varying strengths and are generally weak except when lies are about transgressions or highly motivated [9]. The strength of deception cues is affected by many factors including: the design of the study, the duration of responses, and whether responses were prepared or not.

Simulating guilt and lying has been attempted using various ideas. Lying about transgressions using mock crime scenarios seems to be one of the most popular approaches that have been used in the literature. Mock crimes include concealing a banned object [19], stealing money [31], [47], stealing jewellery [21], or airline passengers attempting to smuggle contraband [25].

Other paradigms also exist, for example, participants might state their attitudes to a series of issues on a personality scale, then lie or tell the truth about their answers on those issues (e.g. personal feelings, beliefs, facts, opinions or academic interests) [9], [38]; describe other people [14]; lie about personal and factual information readily accessible from long term memory [44]; try to convince the interviewer that they were qualified for the job [48]; describe experiences that did or did not actually happen to them [28]; asked to watch a video then tell a lie/truth about what they had seen; or simulate entirely different emotions from the true emotions they experienced as they watched the video [9]. Some scenarios are designed more closely to operational trials [10], [46]. For example, in [46] the participants told the truth or lied about their forthcoming trip, the purpose of their visit, and the intended length of stay.

In many experiments participants were instructed to lie or tell the truth. However, there also exists scenarios where participants lied or told the truth based on their own decision (e.g. see [10] and [19]) and some studies allowed participants to practice their lies before the interview [42], [43]. Liars who prepare themselves when anticipating an interview typically show fewer cues to deceit than unrehearsed spontaneous lies; especially if liars correctly anticipated the questions [42]. Vrij et al. indicated that while the anticipated questions failed to predict the liars and truth tellers, the unanticipated questions resulted in a prediction rate up to 80% [43].

B. Detecting Deception

Studies by Pavlidis et al. suggested that the thermal eye signals from the deceptive subjects have a steeper ascend compared with the responses from the non-deceptive subjects [30], [31]. A subject was classified as deceptive if the eye signals were closer to the eye signals of the deceptive control subjects. In the search for better features that could improve classification performance, Pavlidis et al. used the product of the slopes of the "eye" curves in the question and answer segments as input features of the classifier assuming a bimodal distribution of the slope products and a threshold to obtain the baseline and achieved a classification rate of 84%.

Pollina et al. presented a more elaborate approach for extracting features [33]. The thermal recordings constituted 30 frames captured before and after the onset of the verbal response for each question-answer segment. A correct classification rate of 91.7% for a set of 24 subjects was reported based on the training performance [31].

Tsiamyrtzis et al. carried out a larger scale experiment under various environmental conditions and the average periorbital signal was used for discrimination [38]. A deceptive response was predicted based on local and global slope changes and the classification accuracy obtained using this scheme was 87% for 39 subjects.

Warmelink et al. [46] also investigated the use of thermal imaging for lie detection. Guilt and deception were simulated by asking airport passengers to lie about their intended destination. In their experiment, 64% of truth tellers and 69% of liars were classified correctly.

III. Data and Results

Figure 1 shows the data collection set-up. For data capture, we used a cooled mid-infrared camera: ORION SC7000 (FLIR). The specified resolution of the camera is 640 × 512 pixels and 14-bit grey-level resolution, a sensitivity of NEDT=20mK and operates in the 1.5-5μm waveband. In total, 492 responses (249 lies and 243 truths) were collected from twenty-five subjects. We have used a modified deception scenario [46] which required participants to learn a story and allow them to practice their lies before the interview.

All participants were instructed to read a brief description of the research project and sign an informed consent form. The participants were told that they were tested on interviewing skills, and that the skill under examination was deception as part of human communication. The facilitator explained to the

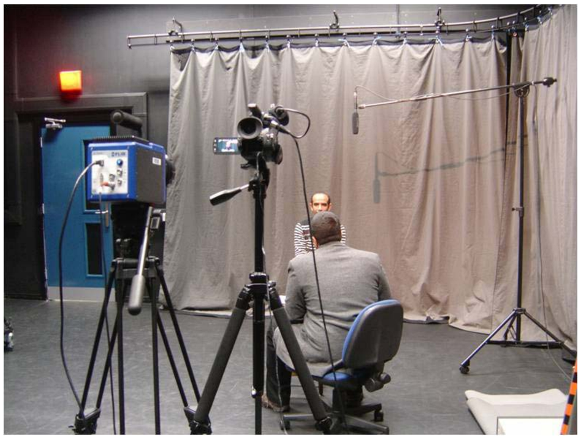

Fig. 1. Experimental set-up: the subject's thermal recordings are captured while the examiner is asking the questions.

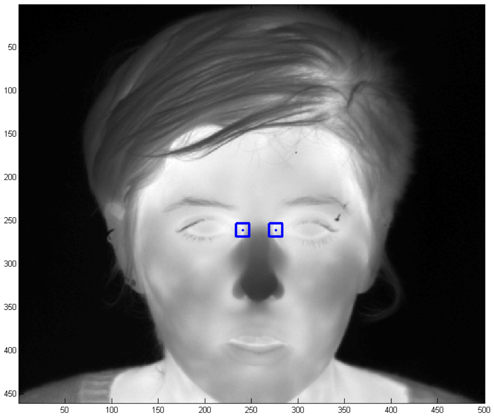

Fig. 2. A thermal image of a participant's face during questioning. The left and right eye corners were tracked for a period of 1.7 s (51 frames).

participants how the examination would be conducted, and offered them a prize (book vouchers) if they were able to convince the examiner that they were honest.

The participants were asked to lie about who they are and what they do. A character profile was provided, and the subject was allowed 5-10 minutes to learn the details before the examination. The character profile contained details on education, family, and work of an assumed person. Each participant was required to attend two separate interview sessions. The second session was conducted two hours after the first session. Each session consists of ten questions.

A. The Interview

It is possible to predict a suspect's strategies from their behaviour and thereby increase the chances of accurately predicting guilt or innocence, as liars tend to become over-controlled and show inhibited behaviour [26]. In this paper, we have two separate examination sessions, in the truth session the participant will be asked to be themselves and answer all questions truthfully (as such there is no story to learn), however in the lie session, we require the participant to learn a character profile and the participant is expected to lie in agreement with the learnt story. The lie-truth sessions were counterbalanced in the sense that half the participants lied in the first session while the other half told the truth. Before the start of each session, each subject was asked four baseline questions which were answered truthfully. The examiner was then allowed to enter the room and the interrogation session began. At the end of the interview, the facilitator verified with the participant that he/she completed the task successfully.

The used scenarios were developed in consultation with psychology and security experts from QinetiQ (www.qinetiq.com). An example of a character profile which was used in our experiment is shown next:

> *"You were born in Ireland but moved when you were 2 and grew up in Wales. Your family (mother, step-father and two elder brothers) still live in Wales, although your eldest brother who is 27 is looking at emigrating to Australia. You enjoy theatre, films and history and love to go to the cinema, museums, see shows and go shopping with your friends. You are a very caring person and love to host people at your house. So often you will arrange dinner evening at your house for friends and, as you like meeting new people, the invitation is often open! You can speak a bit of Latin and French, which has mainly been self-taught from your studies and different texts you have had to cover, although your grandparents now live in France so when you visit them you are able to practise and improve your French."*

When the subject is ready for the examination, the facilitator asks the participant four baseline questions which are answered truthfully. These will be used to register the initial thermal state of the participant. The examiner will be allowed into the room and the interview starts by asking a number of questions that have direct links to specific details in the profile (e.g. "Where do your parents live?" "Tell me about the place where you were born and/or grew up in?") while other questions were designed to force the subject to create lies on the spot by asking about details that do not exist in the character profile (e.g. "Describe the place where your parents live?" "Describe for me one of your lecturers or supervisors").

B. Methodology and Results

1) Extracting Thermal Patterns: The analysis of thermal and visible facial data has advanced significantly over the past decades. However, the automatic extraction of cues is still a challenge and requires the ability to consistently track motion of specific region/points of interest [20], [45] in order to accurately infer internal emotions/behavioural states.

The first stage of data processing starts by manually identifying the two eye corners using two regions of interest (ROI), 17 × 17 pixels each (see Figure 2). The temporal interval is composed of 51 frames (or 1.7 s) formed by taking 25 frames (taking 5 more/fewer frames had little effect on the results) before and after the onset of the subject's verbal response to a given question. Both ROIs are tracked over the entire response time-line using a robust tracker that incrementally learns the appearance of the tracked region by efficiently representing it in a low-dimensional subspace [34]. This generates a $17 \times 17 \times 51$ data volume for each eye corner, which was reduced to a 51 dimensional vector by taking the average value in each 17×17 ROI. The robustness of the manual component in this tracking process and the size of the ROI are evaluated through a set of experiments which we discuss in Section III-D.

In order to compare the subject's truth responses with lie responses on an equal footing, baseline thermal correction is required to compensate for the differences in starting thermal levels. Baseline temperature correction was achieved by subtracting the mean of the four baseline questions from the recording session from all the interview responses for a given subject. Finally, the extracted signatures from both the left and right eye corners are concatenated to form a 102-dimensional feature vector which is then reduced to fewer dimensions by applying principal components analysis (the effect of the choice of number of principal components is discussed in Section III-D).

C. Learning the Model of Deception

A profile of normal/abnormal behaviour can be learned from data gathered from deceptive and truthful interactions. Machine learning can be used to train a classifier that captures the model of deceptive/non-deceptive responses based on either a between-person or a within-person approach. For classification purposes we have used the nearest neighbour classifier because this is a well-known, simple, non-linear, and generic classifier. The nearest neighbour classifier scores a response by assigning the class that receives most votes within a specified neighbourhood size, k. Alternative classification methodologies can also be used (e.g. SVM, logistic regression, decision trees [3]), however, we leave this as further work.

1) Between-Person Approach: The between person approach uses the lie/truth responses from all but one person as the training dataset while the left-out person's lie/truth responses are used as the test dataset. This is done in a round robin way to reuse all persons as the test dataset. This approach is also called leave-one-person-out and could answer the question of whether deceptive patterns generalise across the whole population.

Figure 3 shows the effect of choosing different values of k on the average between-person performance, where we have kept the two largest principal components of the transformed thermal signature (in Section III-D we discuss the effect of using more principal components). Each box represents the variation in the predictive accuracy across 25 individuals. We see that as k varies from 5 to 50, the average predictive accuracy varies between 58-62% with a maximum accuracy obtained for $k = 21$. We also tested a modified baseline thermal correction where the data is transformed to comparable scales by dividing the mean-corrected responses by the standard deviation. This did not improve the classification performance and the accuracy remained 62%. This can be explained by noting that each individual's responses were normalised based on the mean and standard deviation of that individual and as such does not affect the pool for the between-subjects dataset.

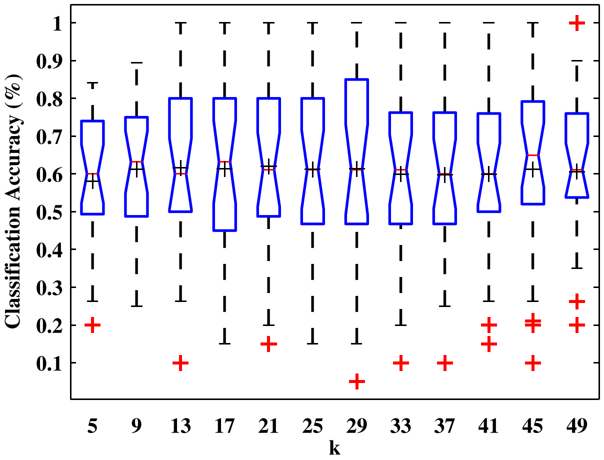

Fig. 3. The average between-person estimate of predictive performance over 25 persons obtained by the k-nearest neighbour classifier versus k. Each box represents the variation in the predictive accuracy across 25 individuals. The central mark on each box is the median (red -), the black "+" represents the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as red +'s.

TABLE I

BETWEEN PERSON CROSS VALIDATION ESTIMATE OF PREDICTIVE PERFORMANCE USING THE k-NEAREST NEIGHBOUR CLASSIFIER ($k = 21$), AVERAGE CLASSIFICATION ACCURACY 61.96%

Person id	$k = 21$	Person id	$k = 21$
1	80.00	14	15.00
2	45.00	15	60.00
3	80.00	16	70.00
4	78.95	17	55.00
5	57.89	18	45.00
6	63.16	19	60.00
7	80.00	20	26.32
8	61.11	21	94.74
9	80.00	22	20.00
10	60.00	23	100.00
11	50.00	24	90.00
12	36.84	25	75.00
13	65.00		

As an explorative example, Table I shows the estimated classification accuracy using the k-nearest neighbour classifier ($k = 21$) and again keeping the two largest principal components of the transformed thermal signature. Note that responses from subjects 2, 12, 14, 18, 20, and 22 were predicted with accuracy below 50% while the responses from subjects 1, 3, 7, 9, 21, 23, and 24 were correctly predicted with at least 80%. The average classification accuracy over all 25 persons in the dataset is 61.96%. It should be noted that these results do not differ significantly from a random chance distribution (at the $p < 0.01$ level).

The above results indicate that deceptive behaviour does not generalise well across the whole population. This conclusion could be explained by the fact that people react differently as they lie. There are also people who are simply bad liars while others might be able to tell lies that are hard to detect. In more scientific terms, the leave-one-person-out approach results in different probability distributions for the training and test sets,

TABLE II
ESTIMATE OF PREDICTIVE PERFORMANCE FOR INDIVIDUAL PERSONS OBTAINED BY THE k-NEAREST NEIGHBOUR CLASSIFIER ($k = 5$) USING THE WITHIN PERSON APPROACH AVERAGED OVER 50 RUNS. – INDICATES NO STATISTICAL SIGNIFICANT DIFFERENCE COMPARED WITH A RANDOM CHOICE DISTRIBUTION, $p < 0.01$ (P-VALUE CALCULATED RELATIVE TO INDIVIDUAL SUBJECTS)

Person id	LOQO	5FCV	2FCV
1	70	72.60 (6.48)	69 (9.14) (–)
2	80	80 (0)	82 (3.91)
3	100	100 (0)	96.50 (2.90)
4	89.47	89.47 (3.97)	85.1579 (6.17)
5	73.68	73.26 (4.15)	62.21 (9.66) (–)
6	68.42	75.47 (7.10)	78.31 (6.94)
7	100	99.20 (1.85)	98.80 (2.15)
8	94.44	94.66 (1.09)	95.77 (2.39)
9	100	99.40 (1.64)	92.20 (8.93)
10	45	62.00 (7.62) (–)	59.40 (11.41) (–)
11	100	100 (0)	95.10 (4.89)
12	94.73	94.84 (0.74)	95.78 (2.12)
13	90	87.60 (3.38)	73.30 (7.99)
14	95	94.90 (3.57)	90.80 (4.55)
15	75	77.60 (3.23)	78 (5.05)
16	90	87.60 (3.38)	87.80 (4.06)
17	45	56.50 (8.22) (–)	54.20 (10.11) (–)
18	100	97.90 (2.87)	98.10 (3.33)
19	95	95.80 (1.85)	96 (2.25)
20	89.47	85.68 (3.98)	77.47 (6.38)
21	100	94.73 (4.75)	86.10 (8.15)
22	65	68.10 (5.33)	60.40 (7.41) (–)
23	100	100 (0)	100 0
24	95	95 (0)	91.90 (2.45)
25	55	56.90 (4.15) (–)	61.00 (6.77) (–)
average	84.40	85.56	82.61

and as such, the model obtained from the training data will not be a good predictor on the test set. Therefore, it makes more sense to compare the person's own lie/truth responses based on the within-person cross-validation approach and this is discussed next.

2) Within-Person Approach: The within-person approach, inspired by the expectancy violations theory [6], is concerned with comparing behavioural profiles against the expected norms. Therefore it is more concerned with classifying a response according to whether it includes deviations from a baseline or discrepancies among indicators [6].

In order to guard against over-fitting, it is necessary to implement n-fold cross-validation to monitor the classification performance: First, we partition the dataset of the lie/truth responses from a specific subject into n folds using stratified sampling. The training set is then constructed by concatenating $n - 1$ folds used to train the model, while using the remaining fold to test the model. This is done in a round-robin approach to generate a total of n local models and n test sets. We then compute the percentage of correctly classified responses by examining the predictions within each test fold and compare it with the ground truth.

Table II shows the average classification accuracy for each subject (using two principal components from the thermal signature) measured over 50 runs and tested using leave-one-question-out (LOQO), five-fold cross-validation (5FCV) and two-fold cross-validation (2FCV). For example, using five-fold cross validation, the average classification accuracy for Subject 1 is 72.60% with a standard deviation of 6.48% (calculated over 50 runs). The performance across all persons using five-fold cross validation is 85.56%. On the other hand, the leave-one-question-out cross validation, results in an average classification accuracy of 84.40%. These results indicate that using a training set comprising ∼95% (LOQO), ∼80% (5FCV), or ∼50% (2FCV) of the data (questions) does produce minor differences (84.40, 85.56, and 82.61 correct classification, respectively), which indicates the robustness of the developed approach. It should be noted that all three sets of results differ significantly from a random chance distribution (at the $p < 0.01$ level).

D. Effect of Parameter Settings

In this section, we investigate the robustness of the developed approach and look at the effect of the tracker and the size/localisation of the region of interest on the classification results.

1) Choice of Eye-Corner Positions: In thermal imaging, the eye-corner regions appear as a smooth (diffused) surface with a relatively higher grey level appearance. Figure 4 (left image) shows the variation in eye-corners positions as a result of this manual selection process. This indicates that due to the manual selection process there can be variation in the start position of the tracker, but that the tracked region (Figure 4 right image shows the position in the last frame) is stable. It is expected that this is also the case if one uses automatic detection of the eye corners. It is clear that the tracker maintains reasonable estimates over time of what it thinks the eye corner region is and successfully copes with image displacements.

We investigated the effect of variation in the eye corner positions on the extracted signatures and the classification performance. We repeated the tracking experiment five times where each time a new location for the eye corner was chosen independently of the previous trial (a fixed region of interest of size 17×17 pixels was used). We did this for all subjects and use all the questions. The corresponding thermal signatures are shown in Figure 5 where the "blue curves" represent the truth signatures and the "red curves" represent the lie signatures. It is clear that there can be a considerable effect on the thermal levels and this is especially clear for the truth responses which are extracted from the right eye corner (i.e. frames 52-102). In particular, the figure shows that the bottom two truth responses are very close to the temperature levels of the lie responses. As such, it is clear that different initialisations could potentially affect the classification accuracy (e.g. such as the bottom two truth responses extracted from the right eye region).

To show the effect on the classification accuracy we again use the within-person approach with five fold cross-validation. The resulting average classification accuracy over five repetitions (trials) for all subjects are shown in the first column of Table III. The results indicate that the choice of eye-corner position could have a significant impact on performance. The statistical significance of the combination of the five corners is estimated using a meta-analysis approach [5], which for each subject takes all five mean and standard deviations into account. This analysis shows that for subjects 10 and 22 the results are not significant. Not all the variation is purely

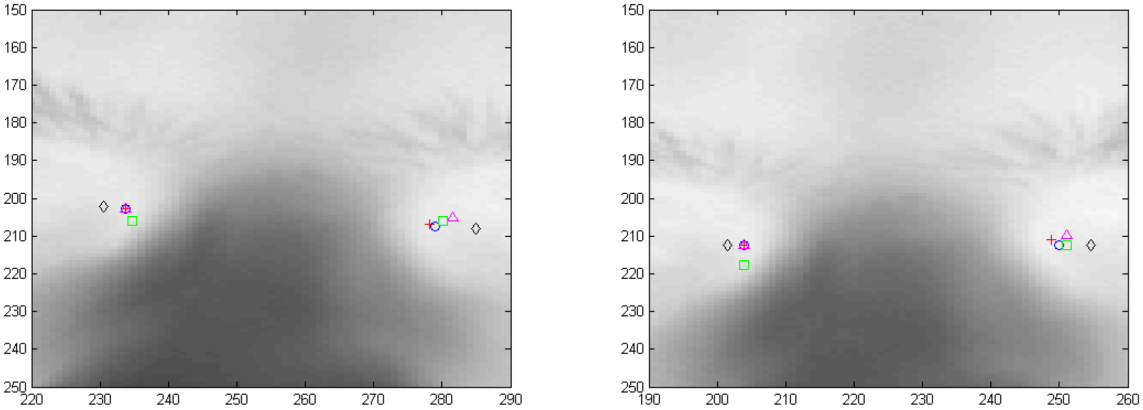

Fig. 4. Five corner positions for subject #1 obtained using five independent trials. Left: initial positions of corners. Right: tracked corners at the last frame of the sequence (the 51^{st} frame).

TABLE III

THE FIRST COLUMN SHOWS THE EFFECT OF THE VARIATION IN INITIAL EYE CORNER POSITIONS ON THE PREDICTIVE PERFORMANCE (USING A FIXED 17 × 17 ROI). THE SECOND COLUMN SHOWS THE EFFECT OF THE VARIATION IN THE SIZE OF THE ROI ON THE PREDICTIVE PERFORMANCE (USING A FIXED EYE CORNER POSITION). THE RESULTS IN THIS TABLE ARE ALL BASED ON A WITHIN PERSON APPROACH AND 5FCV AVERAGED OVER 50 RUNS. – INDICATES NO STATISTICAL SIGNIFICANT DIFFERENCE COMPARED WITH A RANDOM CHOICE DISTRIBUTION, $p < 0.01$ (P-VALUE CALCULATED RELATIVE TO INDIVIDUAL SUBJECTS USING A META-ANALYSIS APPROACH [5])

Person id	using five corners	using four ROI sizes
1	78.22 (8.76)	72.22 (5.31)
2	63.14 (10.56)	78.77 (7.81)
3	100 (0)	100 (0)
4	71.15 (14.06)	78.15 (7.68)
5	60.71 (9.40)	58.86 (6.92) (–)
6	80.10 (18.37)	82.94 (3.04)
7	77.88 (19.49)	97.05 (3.61)
8	83.19 (9.31)	73.77 (9.13)
9	95.78 (5.79)	77.62 (16.99)
10	48.74 (14.75) (–)	54.70 (15.09)
11	76.78 (13.15)	80.65 (13.23)
12	72.16 (18.94)	77.60 (21.25)
13	85.06 (6.27)	82.67 (3.19)
14	66.84 (18.51)	74.25 (6.08)
15	67.30 (19.38)	76.57 (4.21)
16	75.28 (14.83)	72.00 (14.63)
17	72.18 (6.78)	66.22 (8.06)
18	89.64 (9.95)	87.65 (5.33)
19	81.44 (10.65)	88.05 (2.17)
20	62.10 (9.24)	73.57 (17.81)
21	91.19 (4.82)	97.57 (2.18)
22	50.84 (7.47) (–)	51.75 (7.39) (–)
23	95.86 (4.51)	99.22 (1.55)
24	89.28 (4.10)	92.60 (2.83)
25	83.06 (6.93)	77.25 (7.61)
average	76.71	78.87

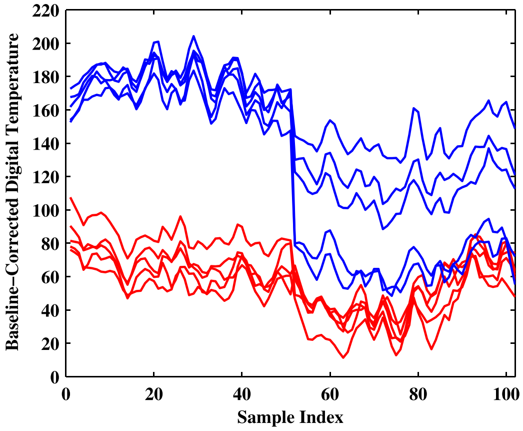

Fig. 5. The corresponding thermal signatures for five independent initialisation of the eye region tracker. This figure shows the effect of manual variation of the eye tracker starting point on the extracted thermal levels. The "blue curves" represent the truth signatures and the "red curves" represent the lie signatures. Note that the sample index 1-51 represents the left eye region and 52-102 the right eye region, and as such there can be a discontinuity at the 51-52 transition. In addition, the bottom two truth responses (i.e. frames 52-102 which are extracted from the right eye corner) are very close to the temperature levels of the lie responses.

attributed to variation in initial corners location. On average a 5% standard deviation is due to the effect of different possible ways one can split the data into five folds (e.g. see second column of Table II which shows effect of repeating data partitioning 50 times).

2) Effect of the Size of the Region of Interest: The size of the region of interest determines how many pixels are involved in computing the mean temperature. The smaller the size of the region of interest the fewer pixels and the average temperature will haver higher frequency variations over time while a larger region of interest will produce smoother signatures. Figure 6 shows the extracted mean signatures for using window sizes of 9 × 9 and 65 × 65 pixels, respectively, and in the second column of Table III we show the corresponding variation in the classification results. The statistical significance of the combination of the four ROIs is estimated using a meta-analysis approach [5], which for each subject takes all four mean and standard deviations into account. This analysis shows that for subjects 5 and 22 the results are not significant.

3) Marginalising Ad Hoc Parameters: Table III indicated that the choice of ad hoc parameters, in particular the location of the eye corner region and the size of the measurement region of interest, has an effect on the classification performance. However instead of using a pre-specified size for the ROI and a single estimate of the eye corner position, it would be more reasonable to extract the thermal signature by averaging across

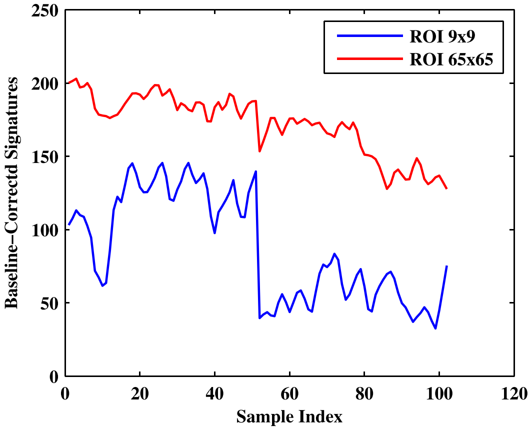

Fig. 6. Extracted signature using two sizes for the region of interest (9 × 9, and 65 × 65). Note that the sample index 1-51 represents the left eye region and 52-102 the right eye region, and as such there can be a discontinuity at the 51-52 transition.

different corner positions and window sizes, which will reduce the noise aspects. This is expected to have a significant impact on the robustness of the thermal deception detection classifier: Table IV (left column) shows the resulting classification accuracies, using (as before), two principal components. The right column, on the other hand, shows the effect of using eight principal components. Using two principal components the average classification accuracy for subject 1 over 50 runs and using 5FCV is 78.64% with a standard deviation of 2.05% and a retained signal variance of 97.22%. It should be noted that both sets of results differ significantly from a random chance distribution (at the $p < 0.01$ level).

Adding more features by using more than two principal components seems to improve the classification accuracy (and the difference is just statistically significant at $p < 0.01$), however, the number of training examples per subject is relatively small with respect to the number of features and these effects with respect to the generalisation of the classifier are future work.

4) Tracking Performance: To show the effect of tracking fluctuations on the extracted signature we initialised two independent trackers to start at the same position and observe the Euclidean disagreement between the two estimated positions for a time period of 300 frames. The non-deterministic aspect of the used tracker [34] means it will not necessarily track the same path along the video sequence. Supporting video material (file: tracking_same_point.mov in the online resources at http://ieeexplore.ieee.org) shows that the trackers maintain a stable position within the eye corner region. The Euclidean displacement error between the two trackers is shown in Figure 7. The disagreement between the two trackers can sometimes reach four pixels, however, this disagreement is not propagated in future frames and in general both tend to agree to a reasonable degree. The average disagreement between the two estimates of the future eye corner positions over 300 frames is around 1.21 pixels which is well within the variation produced by manual (human) estimate for the eye corner positions (see Figure 4).

We now show the effect of this displacement on the extracted thermal signature. We used the tracker output of the previous experiment as the starting point. The first frame of the sequence is replicated to generate a full 300 frame video, with each frame displaced according to the tracker

TABLE IV

THE PREDICTIVE PERFORMANCE FOR INDIVIDUAL PERSONS OBTAINED BY THE k-NEAREST NEIGHBOUR CLASSIFIER ($k = 5$) USING THE WITHIN PERSON APPROACH AND 5FCV AVERAGED OVER 50 RUNS. FIVE DIFFERENT CORNER POSITIONS WERE TRACKED AND SIGNATURES WERE THEN EXTRACTED BY AVERAGING THE RESPONSES OVER FOUR ROI WINDOW WIDTHS (9, 17, 33, AND 65 PIXELS). THE FIRST COLUMN SHOW RESULTS WHEN TWO PRINCIPAL COMPONENTS WERE USED (– FOR SUBJECTS 4 AND 10 INDICATE STATISTICALLY NOT SIGNIFICANT, $p < 0.01$). FOR EXAMPLE, USING TWO PRINCIPAL COMPONENTS, THE CLASSIFICATION ACCURACY FOR SUBJECT 1 IS 78.64% WITH A STANDARD DEVIATION OF 2.05% AND A RETAINED ENERGY OF 97.22%. SECOND COLUMN SHOWS RESULTS USING EIGHT PRINCIPAL COMPONENTS, WHERE ALL RESULTS ARE STATISTICALLY SIGNIFICANT. THE P-VALUES ARE CALCULATED WITH RESPECT TO INDIVIDUAL SUBJECTS

Person id	two principal components	eight principal components
1	78.64 (2.05) 97.22	81.66 (1.91) 98.81
2	76.14 (2.65) 94.98	78.36 (2.30) 97.83
3	100 (0) 99.89	100 (0) 99.96
4	55.85 (2.67) 97.06 (–)	58.50 (2.89) 98.59
5	72.12 (2.03) 99.18	73.20 (2.89) 99.64
6	86.35 (1.67) 95.52	86.73 (1.50) 98.13
7	95.34 (0.91) 95.43	96.62 (0.80) 99.36
8	90.51 (1.23) 97.00	92.11 (1.37) 98.82
9	99.00 (0) 99.13	99.06 (0.23) 99.69
10	52.96 (3.74) 92.97 (–)	68.06 (2.42) 98.97
11	79.48 (1.92) 88.76	91.76 (2.33) 98.83
12	89.17 (1.53) 95.03	87.89 (1.08) 98.11
13	93.04 (0.85) 95.65	100 (0) 99.76
14	67.28 (2.91) 90.01	90.04 (1.80) 97.97
15	78.88 (1.62) 95.95	78.88 (1.69) 98.25
16	86.82 (1.20) 95.73	94.22 (0.84) 99.21
17	75.84 (1.94) 99.01	77.92 (1.57) 99.55
18	95.02 (0.62) 97.72	95.40 (0.80) 99.31
19	62.88 (2.58) 94.71	82.76 (2.35) 98.30
20	72.71 (2.85) 98.94	79.20 (2.98) 99.67
21	97.89 (0) 98.64	97.89 (0) 99.56
22	64.16 (2.55) 94.48	75.86 (2.10) 99.13
23	100 (0) 99.72	100 (0) 99.92
24	98.98 (0.14) 98.80	99.00 (0) 99.83
25	87.14 (1.14) 99.11	87.06 (1.34) 99.75
average	82.24 (14.21)	86.88 (11.18)

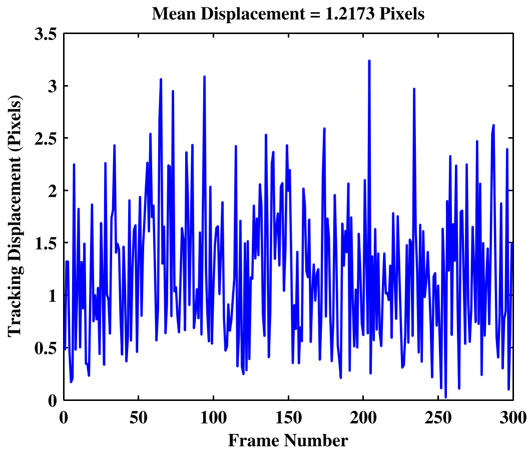

Fig. 7. Disagreement between two trackers initialised at the same image point.

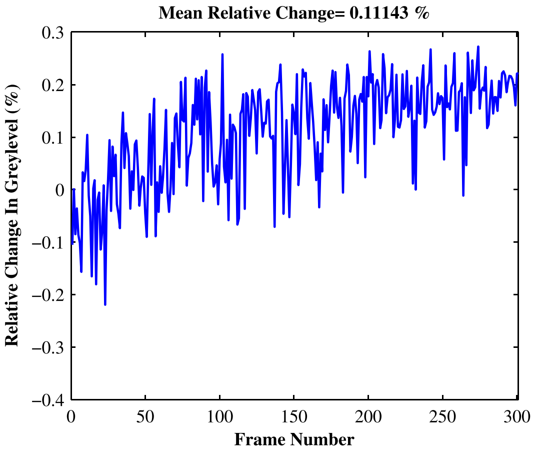

Fig. 8. Percentage difference between the ground truth temperature level and the extracted thermal response.

displacement in the original sequence. In addition, each frame is rotated by $n\pi/30$, where n is the frame number in the sequence. We subsequently initialise the tracker using the same starting point as in the original sequence. The robustness of the tracker can be estimated by evaluating the difference between the thermal response at the initial tracker point and the subsequent tracker position. Supporting video material (file: Synthesised_frames.mov in the online resources) shows that the tracker maintains a stable position within the eye corner region. Figure 8 shows the percentage difference of the extracted thermal responses which shows a maximum deviation of at most ±0.3%.

IV. Discussion and Conclusions

Skin surface temperature can be affected by other factors than deception including: facial expressions [23], body metabolism, changes in the underlying musculo-thermal activities, thermal emissions from the surrounding environment and illness. Therefore, it is important to compensate for such effects by taking into account the different initial baseline temperatures of each individual. Variations in the extracted thermal signatures can be due to how different people respond to the interview question but also due to the choice of different initialisations and parameter settings. We have addressed the variations that might occur as a result of choosing approximate location of the eye corner and as well as selecting different region sizes. It should be noted that overall there is a clear effect on the classification performance and it is recommended to extract more robust thermal signatures by averaging across corner positions and ROI sizes.

Using the generic anxiety level as a measure of deception can produce high misclassification rates - observing what appears to be agitation or over-controlled behaviour does not necessarily indicate deception, in addition, self-regulatory demands of lying do not always exceed those of truth telling [9]. In our experiment we verified this and showed that a between-person approach has poor predictive performance. Learning models of deception based on a leave-one-person-out methodology assumes that the same behaviours and body responses are shared globally across humans of various ages, genders, culture, etc. which so far has not been shown to be accurate. The average classification accuracy using a leave-one-person-out methodology varies between 58-62%. The poor predictive performance is due to the fact that the joint probability distribution of the inputs (thermal data) and outputs (class labels) of the test person is different from the joint distribution formed by all other persons in the training set.

The within-individual approach requires us to learn a classifier for each subject based only on a subset of the subject's responses which are used as ground truth. This scheme agrees with the expectancy violations theory [7]. By only considering a person's baseline model, one is able to detect abnormal variations that might have been triggered by a deceptive response. Our results indicate that the performance across all persons using five-fold cross validation and eight principal components is robust and is statistically significant for all subjects ($p < 0.01$) and with a total average of 86.88%.

As the case of various proposed scenarios for simulating deception in the laboratory which are reported in the literature, our scenario is not very different and is still considered a laboratory-based simulation of deception. As such it is not a true representation of a real-life situations where the stakes are high and individuals are highly motivated to succeed, as otherwise, severe consequences might be faced (e.g. loss of fortune / jail / etc.) which applies for both liars and truth tellers. Nevertheless, this study identifies various factors that contribute to successful implementation/testing of automated deception detection in the real world.

Acknowledgement

The authors would like to thank V. Doherty and colleagues from QinetiQ for providing the experimental design. They would like to thank Prof. Ugail and Dr. M. Yap and the examiner Mr. H. Alawar from the University of Bradford for their help in running the experiment.

References

[1] M. G. Aamondt and H. Custer, "Who can best catch a liar? A meta-analysis of individual differences in detecting deception," *Forensic Examiner*, vol. 15, no. 1, pp. 6–11, 2006.

[2] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, "Face recognition by fusing thermal infrared and visible imagery," *Imag. Vis. Comput.*, vol. 24, no. 7, pp. 727–742, 2006.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.

[4] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality Social Psychol. Rev.*, vol. 10, no. 3, pp. 214–234, 2006.

[5] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Hoboken, NJ, USA: Wiley, 2009.

[6] J. Burgoon *et al.*, "An approach for intent identification by building on deception detection," in *Proc. IEEE 38th HICSS*, Jan. 2005, p. 21a.

[7] J. K. Burgoon, "A communication model of personal space violations: Explication and an initial test," *Human Commun. Res.*, vol. 4, no. 2, pp. 129–142, 1978.

[8] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein, "Lying in everyday life," *J. Personality Social Psychol.*, vol. 70, no. 5, pp. 979–995, 1996.

[9] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, 2003.

[10] D. C. Derrick, A. C. Elkins, J. K. Burgoon, J. F. Nunamaker, and D. D. Zeng, "Border security credibility assessments via heterogeneous sensor fusion," *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 41–49, May/Jun. 2010.

[11] P. D. Drummond and J. W. Lance, "Facial flushing and sweating mediated by the sympathetic nervous system," *Brain*, vol. 110, no. 3, pp. 793–803, 1987.

[12] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics and Marriage*. New York, NY, USA: Norton, 1992.

[13] J. Fei, Z. Zhu, and I. Pavlidis, "Imaging breathing rate in the CO_2 absorption band," in *Proc. 27th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vols. 1–7. Sep. 2005, pp. 700–705.

[14] M. G. Frank and P. Ekman, "Appearing truthful generalizes across different deception situations," *J. Personality Social Psychol.*, vol. 83, no. 3, pp. 486–495, 2004.

[15] J. J. Fuready and G. Ben-Shakhar, "The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge," *Psychophysiology*, vol. 28, no. 2, pp. 163–171, 1991.

[16] J. J. Fuready and R. J. Heslegrave, "Validity of the lie detector: A psychophysiological perspective," *Criminal Justice Behavior*, vol. 15, no. 2, pp. 219–246, 1988.

[17] P. A. Granhag and M. Hartwig, "A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading," *Psychol., Crime Law*, vol. 14, no. 3, pp. 189–200, 2008.

[18] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1334–1345, 2007.

[19] K. Harmer, S. Yue, K. Guo, K. Adams, and A. Hunter, "Automatic blush detection in 'concealed information' test using visual stimuli," in *Proc. Int. Conf. Soft Comput. Pattern Recognit.*, Dec. 2010, pp. 259–264.

[20] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[21] U. Jain, B. Tan, and Q. Li, "Concealed knowledge identification using facial thermal imaging," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 1677–1680.

[22] B. F. Jones and P. Plassmann, "Digital infrared thermal imaging of human skin," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 6, pp. 41–48, Nov./Dec. 2002.

[23] M. M. Khan, M. Ingleby, and R. D. Ward, "Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations," *ACM Trans. Auton. Adapt. Syst.*, vol. 1, no. 1, pp. 91–113, 2006.

[24] J. Knight, "The truth about lying," *Nature*, vol. 428, no. 6984, pp. 692–694, 2004.

[25] R. E. Kraut and D. B. Poe, "Behavioral roots of person perception: The deception judgments of customs inspectors and laymen," *J. Personality Social Psychol.*, vol. 39, no. 5, pp. 784–798, 1980.

[26] M. Lakhani and R. Taylor, "Beliefs about the cues to deception in high- and low-stake situations," *Psychol., Crime Law*, vol. 9, no. 4, pp. 357–368, 2010.

[27] D. T. Lykken, "The GSR in the detection of guilt," *J. Appl. Psychol.*, vol. 43, no. 6, pp. 385–388, 1959.

[28] B. E. Malone, R. B. Adams, D. E. Anderson, M. E. Ansfield, and B. M. DePaulo, "Strategies of deception and their correlates over the course of friendship," in *Proc. Annu. Meeting Amer. Psychol. Soc. Poster*, 1997.

[29] I. Pavlidis, J. Dowdall, N. Sun, C. Puri, J. Fei, and M. Garbey, "Interacting with human physiology," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 150–170, 2007.

[30] I. Pavlidis and J. Levine, "Monitoring of periorbital blood flow rate through thermal image analysis and its application to polygraph testing," in *Proc. 23rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 23. Oct. 2001, pp. 2826–2829.

[31] I. Pavlidis and J. Levine, "Thermal image analysis for polygraph testing," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 6, pp. 56–64, Nov./Dec. 2002.

[32] I. Pavlidis, J. Levine, and P. Baukol, "Thermal imaging for anxiety detection," in *Proc. 2nd IEEE Workshop Comput. Vis. Beyond Vis. Spectr., Methods Appl.*, Jun. 2000, pp. 104–109.

[33] D. A. Pollina *et al.*, "Facial skin surface temperature changes during a 'concealed information' test," *Ann. Biomed. Eng.*, vol. 34, no. 7, pp. 1182–1189, 2006.

[34] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, 2008.

[35] E. N. Sokolov and J. T. Cacioppo, "Orienting and defense reflexes: Vector coding the cardiac response," in *Attention and Orienting: Sensory and Motivational Processes*. New York, NY, USA: Taylor & Francis, 1997, pp. 1–22.

[36] N. Sun, M. Garbey, A. Merla, and I. Pavlidis, "Imaging the cardiovascular pulse," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 416–421.

[37] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. G. Frank, and P. Eckman, "Lie detection-recovery of the periorbital signal through tandem tracking and noise suppression in thermal facial video," in *Proc. SPIE*, vol. 5778, pp. 555–566, Mar. 2005.

[38] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. T. Pavlidis, M. G. Frank, and P. Ekman, "Imaging facial physiology for the detection of deceit," *Int. J. Comput. Vis.*, vol. 71, no. 2, pp. 197–214, 2007.

[39] N. W. Twyman, K. Moffitt, J. K. Burgoon, and F. Marchak, "Using eye tracking technology as a concealed information test," in *Proc. Credibility Assessment Inf. Qual. Government Bus. Symp.*, 2010.

[40] J. M. Vendemia, M. J. Schilliaci, R. F. Buzan, E. P. Green, and S. W. Meek, "Credibility assessment: Psychophysiology and policy in the detection of deception," *Amer. J. Forensic Psychol.*, vol. 24, no. 4, pp. 53–85, 2006.

[41] A. Vrij, K. Edward, K. P. Roberts, and R. Bull, "Detecting deceit via analysis of verbal and nonverbal behavior," *J. Nonverbal Behavior*, vol. 24, no. 4, pp. 239–263, 2000.

[42] A. Vrij, P. A. Granha, S. Mann, and S. Leal, "Outsmarting the liars: Toward a cognitive lie detection approach," *Current Directions Psychol. Sci.*, vol. 20, no. 1, pp. 28–32, 2011.

[43] A. Vrij *et al.*, "Outsmarting the liars: The benefit of asking unanticipated questions," *Law Human Behavior*, vol. 33, no. 2, pp. 159–166, 2009.

[44] J. J. Walczyk, K. S. Roper, E. Seemann, and A. M. Humphrey, "Cognitive mechanisms underlying lying to questions: Response time as a cue to deception," *Appl. Cognit. Psychol.*, vol. 17, no. 7, pp. 755–774, 2003.

[45] J. J. Wang and S. Singh, "Video analysis of human dynamics—A survey," *Real-Time Imag.*, vol. 9, no. 5, pp. 321–346, 2003.

[46] L. Warmelink, A. Vrij, S. Mann, S. Leal, D. Forrester, and R. P. Fisher, "Thermal imaging as a lie detection tool at airports," *Law Human Behavior*, vol. 35, no. 1, pp. 40–48, 2011.

[47] A. K. Webb, C. R. Honts, J. C. Kircher, P. Bernhardt, and A. E. Cook, "Effectiveness of pupil diameter in a probable-lie comparison question test for deception," *Legal Criminol. Psychol.*, vol. 14, no. 2, pp. 279–292, 2009.

[48] B. Weiss and R. S. Feldman, "Looking good and lying to do it: Deception as an impression management strategy in job interviews," *J. Appl. Social Psychol.*, vol. 36, no. 4, pp. 1070–1086, 2006.

Bashar A. Rajoub (M'02) received the B.Sc. degree in communications engineering from the Hijjawi Faculty of Engineering Technology, Yarmouk University, Jordan, and the Ph.D. degree from the General Engineering Research Institute, Liverpool John Moores University, Liverpool, U.K., in 1999 and 2007, respectively. He joined the Vision Graphics and Visualization Group at Aberystwyth University in 2008, and recently moved to the University of Central Lancashire to join the Centre for Applied Digital Signal and Image Processing. His current research interests include thermal imaging, biomedical engineering, digital signal/image processing, nondestructive testing, computer vision, and third-generation machine learning.

Reyer Zwiggelaar received the I.R. degree in applied physics from State University Groningen, Groningen, The Netherlands, and the Ph.D. degree in electronic and electrical engineering from University College London, London, U.K., in 1989 and 1993, respectively. He is currently a Professor with Aberystwyth University, U.K. He has authored and coauthored more than 180 conference and journal papers. His current research interests include medical image understanding (in particular, focusing on mammographic and prostate data), pattern recognition (in particular, applied to thermal facial analysis), statistical methods, texture-based segmentation, and feature-detection techniques.