Measuring cues for stand-off deception detection based on full-body non-verbal features in body-worn cameras

Henri Bouma [1], Gertjan Burghouts, Richard den Hollander, Sophie Van Der Zee, Jan Baan,
Johan-Martijn ten Hove, Sjaak van Diepen, Paul van den Haak, Jeroen van Rest

TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands

ABSTRACT

Deception detection is valuable in the security domain to distinguish truth from lies. It is desirable in many security applications, such as suspect and witness interviews and airport passenger screening. Interviewers are constantly trying to assess the credibility of a statement, usually based on intuition without objective technical support. However, psychological research has shown that humans can hardly perform better than random guessing. Deception detection is a multi-disciplinary research area with an interest from different fields, such as psychology and computer science. In the last decade, several developments have helped to improve the accuracy of lie detection (e.g., with a concealed information test, increasing the cognitive load, or measurements with motion capture suits) and relevant cues have been discovered (e.g., eye blinking or fiddling with the fingers). With an increasing presence of mobile phones and bodycams in society, a mobile, stand-off, automatic deception detection methodology based on various cues from the whole body would create new application opportunities. In this paper, we study the feasibility of measuring these visual cues automatically on different parts of the body, laying the groundwork for stand-off deception detection in more flexible and mobile deployable sensors, such as body-worn cameras. We give an extensive overview of recent developments in two communities: in the behavioral-science community the developments that improve deception detection with a special attention to the observed relevant non-verbal cues, and in the computer-vision community the recent methods that are able to measure these cues. The cues are extracted from several body parts: the eyes, the mouth, the head and the full-body pose. We performed an experiment using several state-of-the-art video-content-analysis (VCA) techniques to assess the quality of robustly measuring these visual cues.

Keywords: Deception detection, bodycam, video-content analysis, surveillance, interrogation.

1. INTRODUCTION

Deception detection is valuable in the security domain to distinguish truth from lies. It can be used in many security applications, such as suspect and witness interviews, object security, emergency response, airport passenger screening and intake interviews with refugees asking for asylum. Interviewers are constantly trying to assess the credibility of a statement, usually based on intuition without objective support. Deception detection is a multi-disciplinary research area with an interest from different fields, such as psychology and computer science. Psychological research has shown that humans obtain a deception-detection accuracy of 54%, which is hardly better than random [14]. The last decade, several developments helped to improve the accuracy of lie detection. For example, researchers developed interviewing techniques that induce cognitive load and magnify behavioral differences between truth tellers and liars [91]. A more technological development is the automated measurement of deceptive behavior, for example based on video footage [63] or full-body motion-capture data [90]. Although the previously described automated measurements have a positive effect on accuracy, these methods often decreased the practical applicability of deception detection. Full-body motion capture suits are expensive and time-consuming to put on, and video footage from static cameras will only be available in specific situations. A mobile, stand-off, automatic video analysis of interviews would allow for a much broader application of automated deception detection especially as part of an investigation with bodycams in the vicinity of an incident. An automated interview analysis could help law-enforcement agencies to stay close to the factual information and avoid subjective interpretation. Statement reliability testing is not only relevant for suspect interviews, but may also be useful in witness interviews that either take place immediately after or close to the incident, or later in time at a police station. Witness statements have often found to be unreliable and inconsistent because people unconsciously interpret

[1] henri.bouma@tno.nl; phone +31 888 66 4054; http://www.tno.nl

H. Bouma, G. Burghouts, R. den Hollander, S. Van Der Zee, J. Baan, J.M. ten Hove, S. van Diepen, P. van den Haak, J. van Rest,
"Measuring cues for stand-off deception detection based on full-body non-verbal features in body-worn cameras,"
Proc. SPIE, Vol. 9995, (2016). http://dx.doi.org/10.1117/12.2241183

and are affected by several biases. Therefore, a witnesses may honestly believe to have observed something that they did not actually observe. Making a distinction between true memories, deliberate deception or accidental false memory is therefore an important capability in witness interviewing. Although the accuracy of these distinctions may not be sufficient for evidence in court, they are valuable as steering indications for efficient usage of means in an investigation. However, to support law-enforcement agencies in making assessments of the reliability of statements, research is needed that enables automatic non-intrusive observation at a distance using video content analysis.

Our main contribution is that we study the feasibility of measuring non-verbal cues that can indicate deceit with video content analysis (VCA). In this paper, we give an overview of recent developments in two communities: in the behavioral-science community the developments that improve deception detection with a special attention to the observed relevant non-verbal cues, and in the computer-vision community the recent methods that are able to measure these cues. The cues are extracted from several body parts: the eyes, the mouth, the head and the full-body pose. We performed an experiment using several state-of-the-art VCA techniques to assess the quality of robustly measuring these cues. Our experiments and results are mainly based on static cameras, but we foresee the use in body-worn cameras, which fits the trend towards more flexible and mobile deployable sensors, such as smartphones and drones.

The outline of this paper is as follows. Section 2 gives an overview of related work. Section 3 presents the technological approaches for measuring these cues. Section 4 describes the experimental setup and shows the results. Finally, Section 5 presents the conclusions.

2. RELATED WORK

2.1 Cues for deception detection

Verbal and nonverbal cues

Deception detection is based on the assumption that lying elicits behavioral changes and that these changes are detectable. To identify behavioral changes, one can first identify baseline behavior and subsequently compare with behavior demonstrated in response to target questions. However, it has proven difficult to establish a reliable baseline of truthful behavior because behavior is affected by more factors than just deception. Alternatively, researchers can identify deceptive behaviors based on behavioral differences on group level. This research has provided us with a set of cues to deceit, behaviors people display more or less when deceiving compared to truth telling. These cues can be categorized into three different types: everything a person says (verbal cues), how he or she behaves and speaks (nonverbal cues and para-verbal cues) and physiological responses. With a multimodal approach, cues from different categories can be combined to increase deception-detection accuracy [33][93]. In this paper, we focus on nonverbal cues.

Three underlying factors: emotional response, cognitive load and behavioral control

Unfortunately, there is not one cue that is uniquely related to lying. In other words, there is not one type of behavior that people only do when lying, and not when being truthful. Instead, people may display certain behaviors more whilst lying, and other behaviors a bit less. And some behaviors are more strongly related to deceiving than others. Which deceptive behaviors a person is likely to display when lying depends on three (formerly four) factors associated with lying [93][108]. These factors are an emotional response, increased cognitive load and attempted behavioral control. First, lying can elicit an emotional response. Especially fear, guilt and delight have previously been associated with lying [33]. Emotions like fear have been found to decrease the production of illustrator gestures (i.e. movements that accompany speech) [32], and increase the production of self-adaptors and fidgeting [108]. Similarly, the excitement experienced when lying has been shown to increase the occurrence of body movements, such as smiling and illustrator gestures.

Second, lying can be more cognitively demanding than truth telling because formulating a lie whilst remembering the truth, avoiding slips of the tongue, and monitoring one’s own and the receiver’s behavior can increase cognitive load [91]. This load can be induced by the interviewer by asking reverse order, specific or unanticipated questions, or by giving a secondary task (e.g., drawing) [93] [94]. These questioning techniques increase interview difficulty more for liars than for truth tellers, thereby magnifying behavioral differences between the two. Cognitive load has been found to increase gaze aversion, while reducing hand movement [31], overall body animation [93] and eye blinks [6].

Third, liars usually do not take their credibility for granted [51] and may try to control their behavior in order to appear honest [17]. Therefore, liars may attempt to avoid those behaviors that they believe are associated with deceit. These are not necessarily the behaviors that are actually signs of lying, as many popular beliefs about cues to deceit, such as avoiding eye contact, are actually inaccurate [46][83]. Another common believe is that people move more when lying, which is likely to reduce overall movement when a liar attempts to control his lying behavior [3].

Indicators based on increased and decreased activity
These three factors explain the occurrence of an individual's nonverbal cues to deceit. However, these factors can lead to different and sometimes contradicting behaviors [3]. As a consequence, both an increase and a decrease in specific behaviors can be a sign of lying. For example, an increase in fidgeting can be caused by lie related nervousness, whilst a decrease in fidgeting can be due to increased cognitive load or attempted behavioral control. As a result, studies on nonverbal cues to deceit have found very mixed and often contradicting results. For example, Granhag and Stromwall [40] found a decrease in smiling when lying, while Fiedler and Walka [37] demonstrated an increase. Similarly, Ekman and Friesen [31] found an increase in self-adaptors when lying, while Gross and Levenson [41] demonstrated a decrease.

Consistency in large meta-analysis
Because results differ so much between studies, meta-analyses – which combine the results of several studies – provide a better insight into behavioral correlates of deceit. The meta-analysis by DePaulo et al. [25], which included 120 deception studies, demonstrated that pupil dilation, ambivalent behavior, vocal uncertainty and tension, nervousness, chin raise, lip pressing, and facial pleasantness are nonverbal cues to deceit that are consistently found across studies. Several other cues but seemed related to deceit as well, including foot movement, pupil changes, genuine smiles, indifferent behavior, planned/unspontaneous behavior, intensity of facial expressions, and direct orientation.

Objective and subjective cues
Some of these cues can be measured objectively, such as pupil dilation [27], chin raise and lip pressing. Others are more subjective and difficult to annotate, such as ambivalent behavior, vocal uncertainty, facial pleasantness and genuine smile. These cues are often rated by several people across a statement based on their impressions [25]. Subsequently, ratings of truthful and deceptive statements are compared to identify behavioral differences on group level. There are also cues for which the type of measurement varies across studies. For example, vocal tension can be measured subjectively based on impressions, or objectively based on vocal micro-tremors [62]. Expert-based rules could be designed to recognize objective cues while the subjective cues require another approach based on training by example.

Manual analysis in the scientific community
The majority of the former mentioned cues are derived from studies in which liars were videotaped and their behaviors were subsequently manually annotated (i.e., coded or rated). There are several issues associated with manual coding [89]. Manual coding is very time consuming and there is a trade-off between the number of coded actions and the costs associated with coding these actions [68]. Manual coding also involves a predefined set of actions that are coded, which means that other actions may go unregistered. Furthermore, manual coding relies on the interpretation of the coder and therefore is subjective [75]. Last, manual coding is often expressed binary (e.g., right hand movement at this point in time, yes or no), which means that the duration and magnitude of the movement are not taken into account. For these reasons, an automatic measurement of nonverbal behaviors is gaining in popularity in the scientific community.

Reliability of statements in security community: deception and false memories
In the security community, development of the ability to automatically measure behaviors enables the real-time usage in police interviews to assess the reliability of statements of suspects and witnesses on the spot. In this paper, we focus on the measurement of cues that (in future work) can be used for assessing the reliability of statements remotely by automatic video analysis, and include cues for two types of incorrect statements: deliberate (deception) or accidental (false memory). In their search for the truth, the police do not only gather as much relevant information as possible, but they also have to determine the reliability of that information. Unreliable information from witnesses or suspects can steer the investigation in the wrong direction and cause great disruption. Both deliberate lies and involuntary false memories will lead to misinformation, but the detection possibilities for the two are very different.

Deception detection is based on the assumption that the act of deceiving affects the deceiver's physiology and behavior (nonverbal, para-verbal and verbal), and that changes in these behaviors can be identified. To circumvent reliability issues associated with human deception detection (with the low accuracy of 54%) [14], researchers started experimenting with automated detection methods based on recordings of truths and lies. Especially video recordings are of interest, because they comprise both physiological and behavioral cues to deceit. Based on visual content, nonverbal cues such as blinking, eyebrow movement, wrinkling, mouth corners [80], lip pressing, full-body motion [67], thermal variations [70] and saccades [92] can be identified. Measurements of these behaviors are based on computer vision techniques such as pose estimation, gesture recognition and facial expression recognition. Standardized multimodal deception detection techniques have proven successful: research based on trial video data [63] reached accurate detection rates up to 75%, and studies on other real life video data further increased that achievement up to 82% [64]. Although these standardized analysis techniques provide promising results, behaviors in these studies were manually annotated. A fully automated analysis of deceptive behavior would allow for a more objective and real-time analysis. Several researchers have investigated automatic deception detection based on nonverbal behaviors. For example, Meservy et al. used blob analysis of the head and hands for classification of guilty and innocent [58]. Limitations of the system are that it requires, extensive interaction and good orientation and lighting for the color segmentation. Bhaskaran et al. used an eye tracker for deception detection [12]. Unfortunately, the system requires manual initialization. Abouelenien et al. use physiological, thermal and visual CERT [54] features [1]. The thermal and physiological modalities perform slightly better than random guessing, but the visual modality suffers from a deteriorated performance. Yu et al. used visual cues that consist of head and face tracking and interactional synchrony estimation [100]. They extracted head nodding, head shaking, smiling and head direction (looking forward or looking away). Detection accuracies of these four features are between 30 and 55%. Average precision and recall in detection of truth and cheating based on the combined features is much higher (66%). Su et al. used 2D appearance-based methods to characterize nine separate facial regions using facial analysis for: eye blink, eyebrow motion, wrinkle occurrence and mouth motion [80]. A Random Forest was used to classify deceptive and truthful categories. Despite the uncontrolled factors in their dataset (illumination, head pose and facial occlusion), they achieved a high accuracy (77%). These results show that automatic measurement of behaviors, especially facial clues, are very promising. The majority of research in this area has focused on automatically measuring facial cues, although a study using full-body motion capture suits has demonstrated that movements in the entire body can be indicative of deceit [89]. Our research will further contribute to the field of VCA-based deception detection because it investigates an automated measurement method of full-body deceptive behavior, and lays the groundwork for analyses of video footage from mobile cameras, increasing the practical applicability of these analysis techniques.

A second way in which a statement can be incorrect is due to misremembering, also referred to as false memories. False memories predominantly affect the reliability of (eye) witness statements and are a common problem in the legal system. Incorrect eyewitness identifications are the largest cause of wrongful convictions in the US [39]. So far, false memory research has focused on demonstrating that false memories exist and affect the legal system [38], that certain interviewing techniques can increase the chance of forming false memories [56], and how to distinguish between true and false memories. Research on the latter has demonstrated that reports of false memories are actually very similar to reports of true memories and therefore difficult to distinguish [77]. Differentiating between true and false memories is important to determine whether someone has witnessed or done something. When a piece of familiar information is presented, an involuntary orienting reflex occurs. This reflex can be measured in several ways, including physiology, EEG, fMRI, thermal cameras (blush detection [101]) and visual cameras (eye-tracking and response latency). Based on this reflex, the Concealed Information Test (CIT) was designed [10]. A CIT is constructed of several multiple-choice questions in which crime-related details are presented to a suspect. When an orienting reflex repeatedly occurs in response to crime-related information, it indicates the suspect holds perpetrator knowledge.

Combining non-verbal cues for assessing the reliability of statements

Previous deception research on nonverbal cues to deceit has demonstrated that automated methods outperform human deception detection abilities. However, the methodologies with which these accuracies were reached are more difficult to apply in a realistic context. We aim for measuring stable cues that can be automatically extracted by video analysis. The combination of measuring multiple non-verbal cues across the whole body makes it harder for interviewees to deceive the system than with only a single cue. In summary, for the assessment of reliability of statements, important visual cues are related to eyes (e.g., pupil dilation, blinking, saccades, gaze aversion), facial expression (e.g., intensity, of expressions and wrinkling), head (e.g., chin raise), mouth (e.g., pressing lips and genuine smiles), hand and arm gestures (arm movement, hand movement) and full-body pose (e.g., foot movement and full-body motion). Furthermore, the combination of visual cues is stronger than individual cues.

2.2 Video content analysis to measure deception cues

In this subsection, we give an overview of recent developments in the computer-vision community to measure the previously described non-verbal cues that are relevant for deception detection. The cues are measured at several positions and scales. We discuss them in the following order: full body pose, hand and arm gestures, and facial behavior.

Full body pose

There are several types of techniques for full-body motion recognition. The first type is based on silhouettes without more detailed appearance information. They are using motion history [13], single depth images [78], optical flow [29].

The second is based on grids. Examples of this type are the use of Histograms of oriented gradients (HOG) with non-negative matrix factorization [84], optical flow in shape contexts [102], and space-time volumes [52]. The third type uses deep-learning. Examples of these are using deep neural nets [85], flowing ConvNets [66], deep structure [104], pooling [48][49] and a convolutional network with a graphical model [86]. Other approaches are using part-based histograms [47], search-space reduction [30][36], single depth images [78], pose shape and appearance models [50][61], iterative error feedback [22], spatio-temporal matching [61][105], dense optic flow and flowing puppets [107], multi-view poses [9], poses using temporal links [23], body models for action recognition [96], joint trajectories [2] and flexible mixtures of parts [71][98]. Recently, a reference guide for human pose estimation was written [79] and the different pose-estimation techniques have been compared in a benchmark and overview of state-of-the-art by Andriluka et al. [4]. This overview shows that the flexible mixture parts of Yang and Ramanan [98] performs well in comparison to others. Therefore, we use this method for recognizing full-body motion as a starting point in our work. The method uses a new representation of deformable parts models that jointly captures spatial relations between part locations and co-occurrence relations between part mixtures.

Hand and arm gestures

Hand gesture recognition is an active field of research, especially to enable human-computer interaction. Initially, only data gloves allowed effective capturing of hand motion. Vision-based hand-gesture recognition is an alternative to the data glove that is less expensive and provides non-contact interaction [35]. The two major categories of visual hand gesture representation are 3D-model-based methods and appearance-based methods [72]. Red-green-blue (RGB) camera-based gesture recognition can suffer from occlusions, light change, or other skin-colored objects in the background. In these complicated situations, the gesture recognition can be assisted with depth images, either extracted from stereo video cameras, or sensed directly with depth cameras such as the Microsoft Kinect or ASUS Xtion [81][99]. Recent results show that hand gestures and finger positions can be estimated in real time [106]. A review about detection, tracking and recognition of hand gestures in given by Rautaray [72]. In our work, we focus mainly on analysis in RGB cameras where we combine pose and gesture estimates.

Facial behavior (including head, eyes and mouth)

There has been an increasing interest in the automatic interpretation of social signals based on facial behavior. The face is an important part of the body that shows nonverbal communication. Facial behavior analysis includes facial landmark detection, head pose estimation, detection of facial action units, facial expression recognition, and eye gaze estimation. Comprehensive surveys can be found about head pose estimation [60], facial feature point detection [95], recognition of facial affect [74], face recognition [26], eye gaze estimation [43] and facial expression recognition [73][11][5]. Recent benchmarks and methods are proposed for eye gaze estimation [87][103][82], eye blink detection [53][45], eye localization [20], micro-expressions [57], expression recognition [21][54][88][42], sentiment analysis [69], and facial behavior analysis [8]. Examples of complete open-source packages for facial behavior analysis are the CLM framework (Cambridge Face Tracker) [7] and OpenFace [8]. Existing approaches for eye blink detection, such as the Image Flow approach [45] or the Blinky[2] approach are not suitable because they require too high resolution, controlled environment, close proximity and minimal motion. As a basis for our facial behavior analysis, we use the CLM framework (version 1.3.6), which includes the 3D constrained local model (CLM-Z) and the constrained local neural fields (CLNF) for robust facial landmark detection.

3. METHODOLOGIES FOR VISUAL CUES

In this section, we describe the visual cues that we measured at several body parts, namely for the head, eyes and mouth (three components of facial behavior in Subsection 3.1, 3.2 and 3.3 respectively) and the full-body pose and gestures (in Sec. 3.4), and we describe the techniques that are used to detect their condition in the images.

3.1 Head motion

A basic step in the analysis of facial expressions is the accurate localization of the head and characteristic facial feature points. One very successful method for measuring these characteristics is the Constrained Local Model (CLM) method by Cristinacce and Cootes [24]. This method accurately finds facial feature points under varying viewing conditions. The method has been refined in later work, and the CLM-Z framework [7] is used here as baseline. Starting with a face detector, it calculates 68 face points across the face, and subsequently derives the head orientation from these points. An

[2] https://github.com/dilawar/eye-blink-detector

example of the face points is shown in Figure 1. The facial feature points allow the localization of facial parts, such as eyes and mouth. This enables further measurement of behavior specific characteristics.

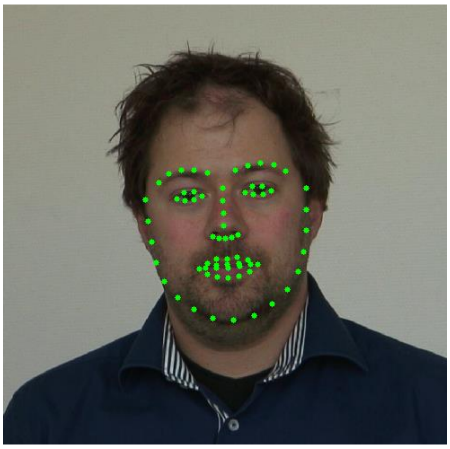

Figure 1: CLM landmarks on the face.

The implementation facilitates also the measurement of head orientation (Figure 2). The amplitude of emblem head movements (e.g., nodding 'yes') and chin raises are indicative for deception detection. The head orientation can be used for recognizing rotating actions, such as pitch (nodding 'yes'), yaw (shaking 'no') or roll. We applied several filtering steps to remove bias and noise and the four expert-based rules of Table 1 to recognize these rotational head actions.

Table 1: Expert based rules and parameter values to recognize rotational head actions.

id	**Expert-based rule and parameter description**	**Parameter value**		
		Roll	**Yes**	**No**
1	Minimum peak-to-peak amplitude (in radians) of the rotation	0.20	0.13	0.12
2	Maximum time (in seconds) between two subsequent peaks	1.6	0.8	0.8
3	Minimum number of subsequent shakes	2	3	3
4	Subsequent peaks must make sign transition (positive /negative)	False	False	True

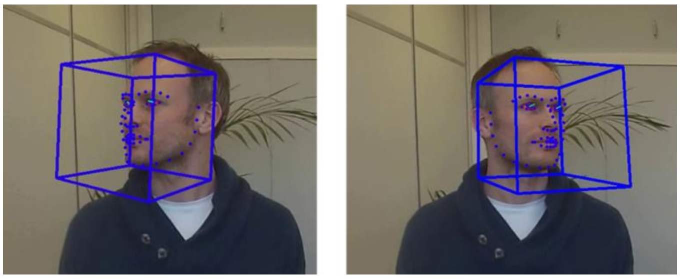

Figure 2: Head rotations with CLM.

3.2 Eye gaze estimation

Eye movements – such as saccades and gaze aversion – are cues for deception detection. One approach for the detection of eye movement is the application of gaze estimation. Although much research has been undertaken on this topic, it is still difficult to reliably estimate gaze direction in arbitrary video footage. We experimented with existing gaze estimators on our dataset (e.g., the CLM implementation of [7], which is visualized in Figure 3), and found that the performance was not yet satisfactory due to limited resolution in our recordings. The extreme eye direction towards the left and right were sometimes not followed correctly. Initial experiments with the gaze estimation with the CLM framework [7] showed a frequent underestimation of the yaw/pitch angle in our dataset, causing the detection of extreme viewing angles to be insufficiently reliable. Next, we experimented with localizing the extreme directions (left, right or upwards) directly from the intensity images by using one dimensional intensity profiles. However, the intensity profiles were defined between CLM face points, and slightly shifted face points hindered the use of intensity profiles. Our approach has therefore been directed towards explicit training on these cases of extreme viewing angles, and as such differentiating them from the cases with central gaze and near-central gaze directions, as described below.

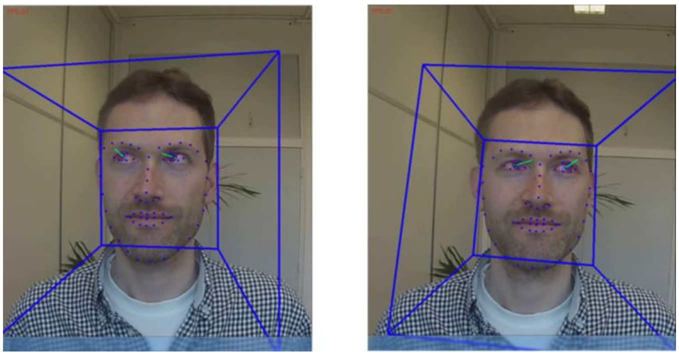

Figure 3: Eye motion with CLM.

Horizontal and vertical eye motion

We developed a more robust use of the intensity and color information, and also an improved localization of the eyes. We found this combination in the design of a dedicated eye direction detector. In particular, a separate detector was made for different gaze directions, such as looking sideways and for looking upwards. The detector was trained using the concept of Aggregated Channel Features (ACF) [28], where the training set (both positives and negatives) was made up of synthetic images. Since the number of instances of the extreme directions is limited in real datasets (including ours), we use synthetic images that assure there are enough training samples for different gaze directions and to avoid overtraining on our dataset. The synthetic images were taken from the UnityEyes tool [97], that synthetizes eyes looking in a specified direction with random variations among gender, ethnicity, age etc. Also camera angle variations and eye direction variations can be specified accordingly.

For the looking-upwards detector we have simulated 474 images with 30 degrees eye pitch angle and 5 degrees standard deviation for eye pitch/yaw and camera yaw (Figure 4). These images formed the positive training set.

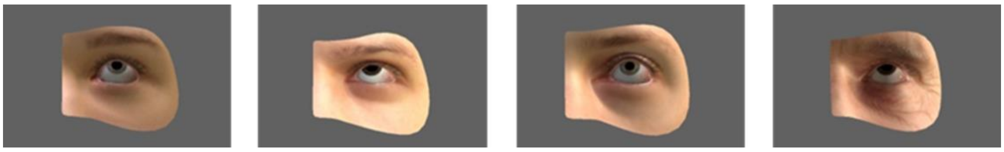

Figure 4: Some synthetic images for looking upwards using UnityEyes.

The positive training set of the sideways looking detector was made up of 474 images with eyes at 30 degree yaw angles, and with the same angle standard deviations as in the upwards case, see Figure 5 for image examples.

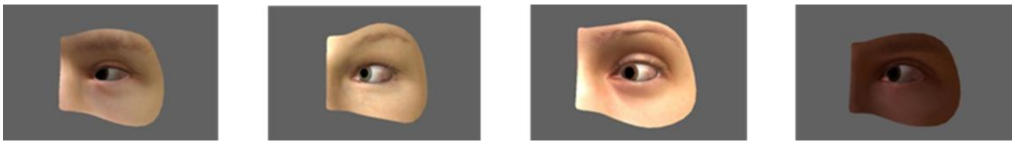

Figure 5: Some synthetic images for looking sideways.

For both detectors, the eye regions in the positive training samples was specified with a bounding box, in order to train the detector on the most relevant and distinctive part of the synthetic images, see Figure 6. The positive samples were also mirrored in order to handle both the left and right eye, and also left and right gazes. In addition, the samples were rotated within a 15 degree (upwards) and 10 degree (sideways) angle to compensate for possible head rotation. Furthermore, the bounding box for the sideways training samples was chosen smaller than for the upwards samples; the appearance change of the eye for sideways looking is less pronounced. It therefore important to include only the variable part in the positive samples, and to omit the eyelids that are more or less identical for central and sideways gazes. In contrast, upward gazes cause a change in the eyelid appearance since the eye opens up more than for a central gaze.

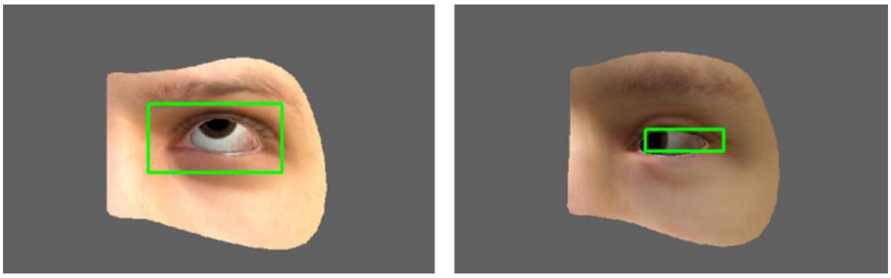

Figure 6: Eye regions in the positive training examples.

The negative training images were created from eye images with the eyes looking central or looking downwards, see Figure 7. The latter set resembles blinking of the eye, which should not confuse the detector. In total 1100 images were generated for the negative set. The detector takes 10000 negative samples with the region size from Figure 7 randomly from the negative images.

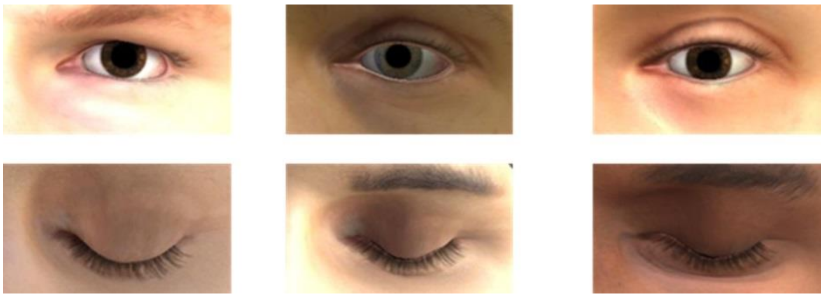

Figure 7: Examples from the negative training set for both detectors.

The detectors were run on all the video frames, and the results were analyzed on a subset of the frames where the head pose was well-conditioned; this assures that the eye regions were not blurred or seen from an extreme viewing angle. We took head poses within 10 degrees pitch and roll angles, and within 20 degrees yaw angle. The frame-to-frame displacement of the head position was limited to 0.5 pixel. In addition, the detections were also filtered on their location at the CLM eye points, in order to be sure they correspond to the eye regions.

During initial experiments (results not shown), we noticed that the sideways gaze detector had less discriminative power than the upward gaze, considering the results on central gaze images. Therefore, we extended the analysis of the sideways detections with an additional check on pixel intensity values inside the detected bounding boxes; for a sideways gaze we expect the lowest intensity values (the pupil) to be at the left/right of the bounding box. In addition, the average intensity in the middle part of the bounding box is higher for sideways gaze than for central gaze.

Blinking
The eye blink frequency is one of the cues that is directly relevant for deception detection. Unfortunately, the existing methods mentioned in Sec. 2.2 could not be used because they require too high resolution, a controlled environment, close proximity and a static camera. Therefore, we developed an approach based on the CLM landmarks. From the CLM points around the eyes, a convex hull is computed to focus on a square region of interest. From this region, an image snippet is extracted. The blinking is detected by analyzing when the eyes are closed for a short period. As a reference of a closed eye, we consider the mean snippet of all closed eyes from a training set. Within the training set, the match of each eye snippet to the mean closed eye snippet is computed by the normalized cross correlation, because it is invariant to intensity changes. Such changes are expected between persons and throughout the recordings due to varying lighting conditions. On the training set, an optimal threshold is determined for the minimum correlation value, above which the eye is considered to be closed. After training, the blinking model consists of this mean closed eye snippet and a detection threshold. On a test person, the eye snippet is extracted from each frame, after which the snippet is classified as closed eye or not. Consecutive closed eyes are merged into one eye blink detection.

3.3 Mouth analysis

The mouth can take a variety of poses of which specific combinations enable us to speak. Much research has been done on lip reading in the context of speech. For deception detection, pressing lips and smiling are important features. This subsection describes our methods for detecting pressing lips and smiles in our dataset.

Pressing lips

The way people press their lips is very personal and unique with regards to force, timing and the proportion of the hidden part of the lips. Furthermore, every mouth has its own unique characteristics with respect so size, color and shape. Finally, the subsequent actions that precede or take place after a pressing lips action vary on a personal level as well. This makes the detection of pressing lips challenging.

The CLM method, as introduced in Section 3.1, is used to detect the region of the mouth in the video image. In our experiment, early stage findings indicated that geometry features like mouth angles, height, width etc. based on CLM points are unreliable and do not provide directly usable information to detect pressing lips or smiles, because the points of the model do not follow the changing shape of the mouth accurately. Our method for detection of pressing lips uses a vertical scanline in the middle of the mouth from the top to the bottom of the lips. This scanline builds an image over time which represents the RGB values from the vertical middle line of the mouth. Figure 8 shows an example of a scanline on the middle point of the lips over time.

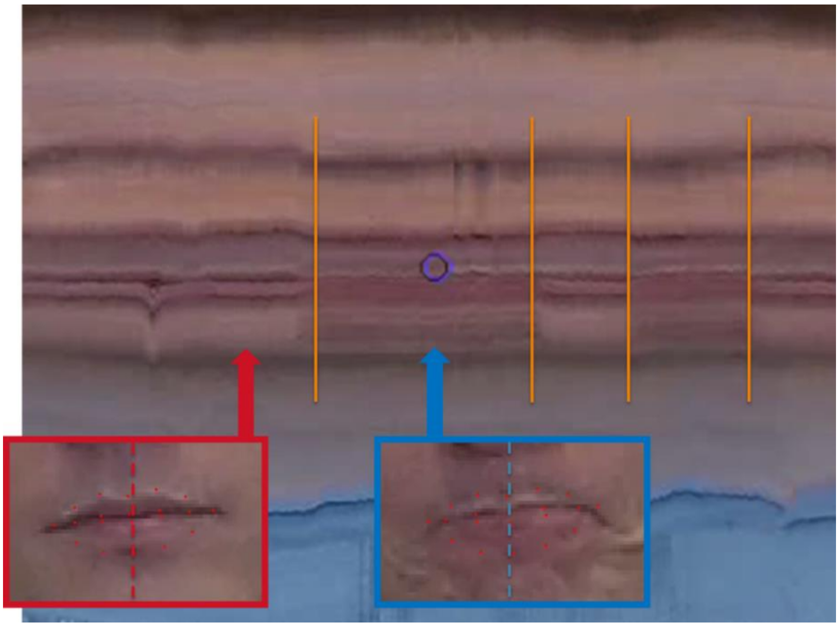

Figure 8: RGB vertical scanlines of the mouth for multiple frames (background). The left (red) overlay box indicates a frame without pressing lips and the right (blue) overlay box indicates a frame with pressing lips. The orange lines indicate transitions between the two modes.

Another example of a scanline is shown in Figure 9. The figure on the left shows the scanline from three channels in the RGB domain, and the figure on the right shows the scanline from the red channel in the RGB domain.

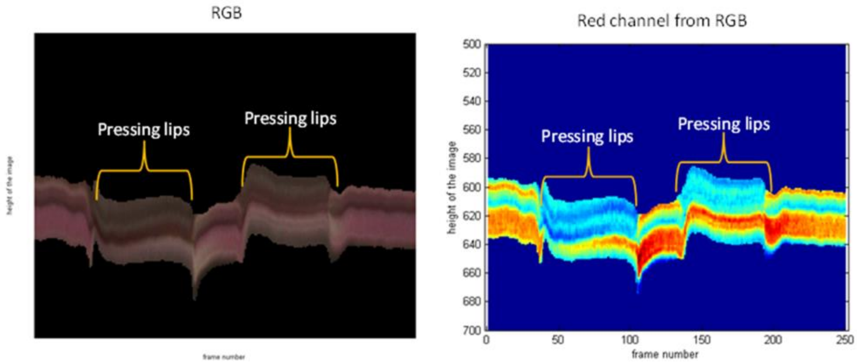

Figure 9: RGB vertical scanlines of the mouth for multiple frames (left) and the red channel vertical scanline of the mouth for multiple frames (right).

From these figures, a pressing lips action shows different color patterns compared to a situation where lips are not pressed. Looking at our subjects, the darkest point of the lips typically makes a downward movement when lips are pressed together. Our method builds a model by taking the mean of the red channel values in the vertical scanline over a number of frames where the lips are not pressed. Our pressing lips detection algorithm detects per frame whether subjects are pressing their lips by detecting whether the darkest point makes a downward movement. A downward movement is considered to indicate a pressing lips action when the distance between the darkest point in the video frame of interest and the darkest point in the model is positive and larger than the threshold. The threshold is set to 0.5 times the standard deviation of all distances where the distance is larger than 0. All distances above this threshold are considered to indicate a pressing lips action. The result of this algorithm is an array where a 0 denotes no pressing lips and a 1 denotes a pressing lips action for a certain video frame. After this, some smoothing is applied. All connected sets of 1's with a duration smaller than 0.2 seconds are removed, gaps containing less than 0.6 seconds are filled.

Smile

The way people smile differs with regards to the opening of the mouth, height increase of the mouth corners, visibility of teeth and the curve of the mouth. Furthermore, the reason because people smile and the severity of a smile can vary. For example, because they have just heard something funny, they want to be polite or they are just happy. For deception detection it is important to distinguish genuine and fake smiles.

As with pressing lips, the CLM method is used to indicate the region of the mouth. The set of initially calculated features consist of vertical RGB and intensity scanlines on the mouth corners and horizontal scanlines over the mouth from the left mouth corner to the right mouth corner over the middle height of the mouth (Figure 10). Our analysis showed that the horizontal scanline of the mouth in the intensity domain (averaged over the RGB channels) provides the most valuable information and therefore, this is used to determine whether a subject in a video frame is smiling or not.

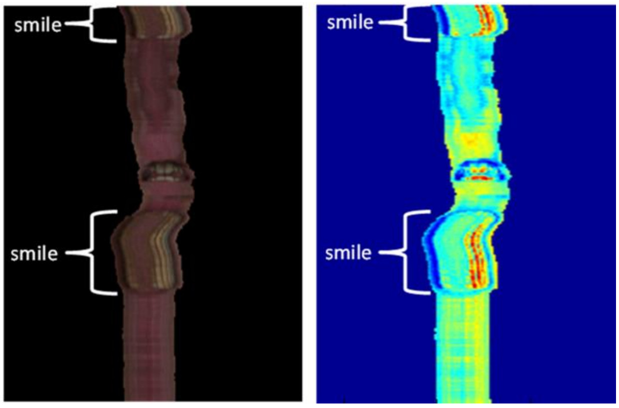

Figure 10: RGB horizontal scanline image of the mouth (left) and red channel horizontal scanline plot of the mouth (right).

Our method computes a model by taking the mean of the horizontal scanlines in the intensity domain for a number of frames where the subject does not smile. The maximum intensity level of the horizontal scanline is lower in cases where the subject does not smile. For the frame of interest, the difference between the maximum intensity level of the horizontal scanline and the maximum intensity level in the model is computed. If the difference exceeds a threshold, we classify this frame as a case where the subject smiles. The threshold is set to the standard deviation of the differences over a set of frames containing smiling and non-smiling cases.

Basically, the intensity becomes higher because the teeth become more or less visible during a genuine smile. This happens during speech as well, but in general the teeth are visible in a much shorter time range during speech compared to a smile. Our method removes detections with a duration shorter than 0.6 seconds.

3.4 Pose and gestures

To identify the movement of torso and arms, two approaches are explored, the first is based on pose estimation, and the second on moving patches.

Pose estimation

Deception detection studies have shown that the motion of arms and scratching parts of the head (cheek, ear, chin) are relevant cues. These movements can be recognized by pose estimators. Yang and Ramanan [98] proposed a method for detecting articulated people and estimating their pose from static images based on a representation of deformable part models. They use a mixture model that jointly captures spatial relations between part locations and co-occurrence relations between part mixtures, augmenting standard pictorial structure models that encode just spatial relations (Figure 11). This pose-estimation model provides 26 patches on a body, each relating to one part of the body. The patches and their movement can be used to detect vertical arm motion and thus the moments that a person scratches his/her head. It was found that it is essential to compute the poses on full resolution, at each frame, for the needed maximum quality. Two patches of the pose estimation relate to one lower arm. All the 26 patches are jittering – because the method considered for pose estimation is a single-frame method – and therefore hindering precise behavior analysis. To reduce the jitter, the two patches related to one arm are averaged to one image location. For further reduction, the location is tracked and smoothed within a one second window (polynomial fit). As a measure of vertical motion of the arm, the distance of the arm location to the head is determined. This distance is normalized by the length of the body. To determine whether an arm has moved vertically, the maximum movement of either the left or right arm is taken, in order to make the approach left-right invariant. Unfortunately, we also noticed unstable pose estimation, especially for interesting gestures when the hands are not relaxed besides the body, but in front of the body or moving upward.

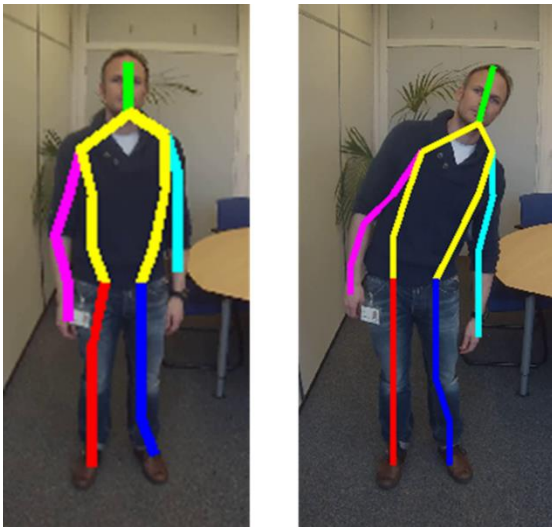

Figure 11: Pose estimation with a pose estimator [98].

Moving patches

As an alternative to pose estimation, a second approach has been explored. First, person detection is performed with ACF [28]. At each image location inside this detected region, multi-scale patches are considered. Three criteria are used to select the patches. The first requires that each patch contains significant motion. This motion is derived from image changes within the patch, relative to the image changes in that patch a little earlier. The image changes are computed with frame differences and the comparison to earlier changes are implemented as a signal-to-noise ratio (SNR). In this way, the small irrelevant motions of high-contrast edges – which may lead to large changes between frames – are counteracted. The second criterion is to avoid duplicate or overlapping patches. Non-maximum suppression (NMS) is applied to limit the number of patches per frame by comparing the SNRs. The third criterion is a minimum time duration to allow tracking. Patches are tracked to obtain speed estimates, which is used as one of the features. Patches that do not meet the three criteria are removed (Figure 12). The feature vector of each patch contains the SNR of image changes, location, horizontal and vertical speed, and speed magnitude, where the values are normalized to the patch size. Within time windows of 0.6 second, all patches are collected and a model is learned through a bag-of-words (BoW) approach and a support-vector machine (SVM) classifier. The moving patches are used to recognize arm movement that indicates scratching of cheek, ear or chin.

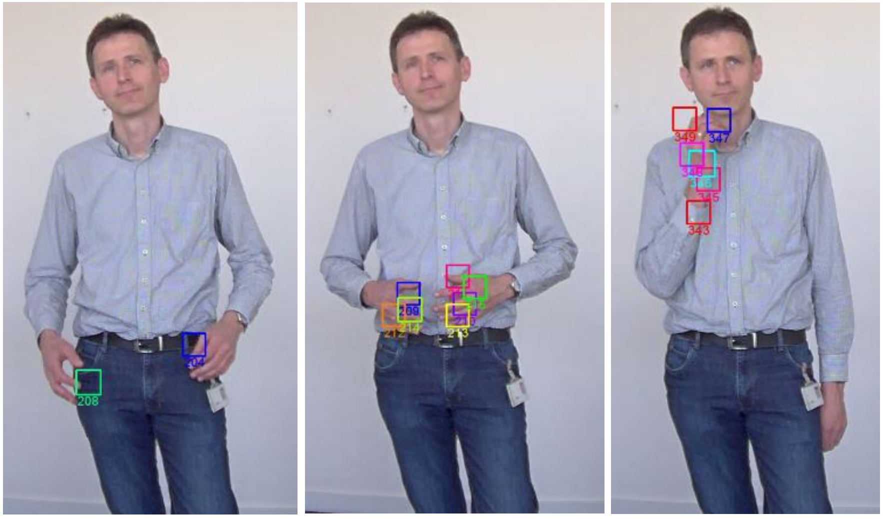

Figure 12: Moving patches.

4. EXPERIMENTS AND RESULTS

In this section, we performed a feasibility experiment using several VCA techniques to assess the quality of measuring relevant cues for deception detection.

4.1 Experiment

Dataset for evaluation

To evaluate the performance of the measured cues, a dataset was recorded with 9 volunteers (Figure 13). The volunteers were asked to perform the actions that are described in Table 2. These actions are related to cues that are relevant for deception detection, but they are not identical. For example, eye movement and head movement are relevant for deception detection, but relevant eye movement is not necessarily horizontal movement, and relevant head movement is not necessarily shaking ‘No’. We modified the actions to make them countable and suitable for quantitative performance assessment. The actions were recorded with multiple cameras, namely two AXIS CCTV cameras with 1920x1080 pixels resolution at 25 frames per second (one for an overview of the whole body and one for a close-up of the upper body), one ASUS Xtion pro RGB-D camera with 640x480 pixels resolution at 25 frames per second, and one GoPro body-worn camera with 1920x1080 pixels resolution and 25 frames per second (Figure 14). The quantitative results in this paper are based on video from the two static AXIS cameras.

Figure 13: The dataset consists of 9 volunteers.

Table 2: Actions recorded in the dataset.

Action label	Action description (translated to English)
Eye move horizontal	Move your eyes horizontal, to the left and to the right (3x)
Eye move vertical	Move your eyes vertically, up and down (3x)
Eye blinking	Blink with your eyes rapidly (3x). Blink with your eyes slowly (3x).
Head shake 'No'	There will be three head rotations. The first is: shaking 'No'.
Head nod 'Yes'	The second rotation is: nodding 'Yes'.
Head rotate Roll	The third is: move your ears to your shoulders while looking forward (roll)
Mouth smile	Pretend a fake smile, as somebody that is not really happy. Create a genuine smile. (A funny remark was made)
Mouth press lips	Press your lips (3x)
Scratch cheek/ear/chin	Move your arms to scratch your cheek. Scratch your ear. Scratch your chin.

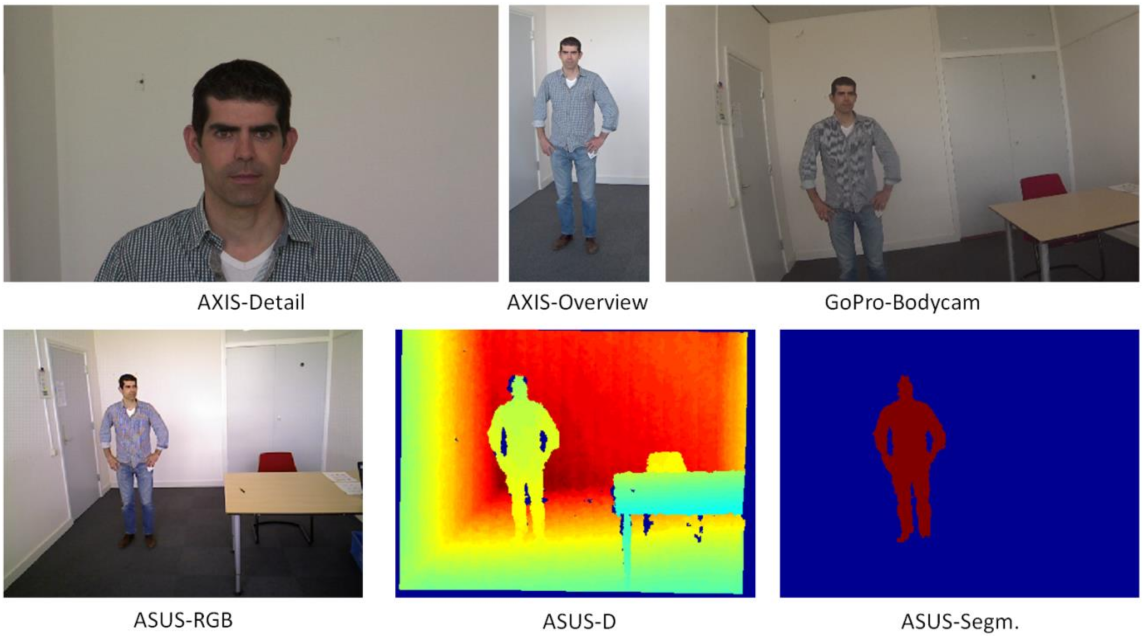

Figure 14: The dataset consists of several cameras: AXIS for details (top left), AXIS for overview (top middle), GoPro bodycam (top right), ASUS for RGB (bottom left), depth (bottom middle) and person segmentation (bottom right).

Performance measures

The performance of the different cues is measured by calculating the precision = TP/(TP+FP) and recall = TP/(TP+FN), where T=true, F=false, P=positive and N=negative. We also report the FP per minute ('FP/min'), to give an indication for the sensitivity of the system over time for the different cues. The FP per minute is calculated by dividing the amount of FP's by the number of minutes of the dataset selection.

We evaluate the different cues by comparing time segments of the system with time segments of the annotations. A time segment is treated as a positive. A true positive is registered when a system time segment overlaps for at least one frame with an annotation time segment. Only one system segment can match with an unique annotation segment, which is the one with highest overlap ratio, other segments of the system that also match with this annotation segment are registered as false positives. An example is shown in Figure 15, where the blue annotation segment matches only one system segment, while the other system segments are registered as false positives.

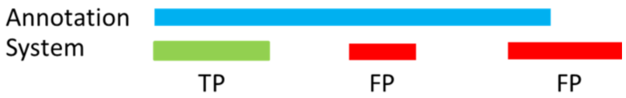

Figure 15: Comparing time segments.

Experimental setup for training and testing

Each of the methods was evaluated in a different way on the evaluation dataset (Table 3) and each way is explained in the following paragraphs.

The method for recognizing vertical and horizontal eye movements is trained on a synthetic dataset that is completely independent from the evaluation set. The method for recognizing head rotations (Yes, No and Roll) does not require training or matching either. Therefore evaluation of these methods on the dataset is straightforward. These methods are applied once to evaluation set and the scores are computed as described earlier.

The methods for eye blinking and moving patches (for arm movement to scratch body parts) are trained in a leave-one-actor-out cross-validation setup. In the cross-validation setup, each persons is tested, one by one, while the method is trained on all other persons. Hence, the person on which the test will be performed, is not within the training set. For blinking, a model of a mean closed eye is created on the training persons, and on a test person, the eye snippet is extracted from each frame and compared with normalized cross correlation to the model. For arm movement detection related to scratching a cheek, ear or chin, a model is learned through a bag-of-words approach on the moving patches on the training persons, and applied to the unseen person in the test set.

The methods for pressing lips and smiles makes a split in the dataset between parts of the dataset that are used for evaluation and parts that are used for creation of a 'normal' model by taking the mean of the values in the scanlines over a number of frames where the mouth is not pressed and not smiling. The scanlines in the evaluation part are compared to this normal model.

Table 3: The methods are evaluated in a different ways on the dataset, due to differences in training and matching strategies.

Action label	**No training and matching**	**Train on synthetic data**	**Train leave-one-actor-out cross-val.**	**Match to normal part**
Eye move horizontal		X		
Eye move vertical		X		
Eye blinking			X	
Head shake 'No'	X			
Head nod 'Yes'	X			
Head rotate Roll	X			
Mouth smile				X
Mouth press lips				X
Scratch cheek/ear/chin			X	

4.2 Results

In Table 4 we summarize the performances on the actions that are related to the cues. This table also includes the total duration of positives in seconds (Time pos), the duration of the evaluation set in seconds (Time evalset), the number of positives (# pos; i.e. system output), the number of ground-truth segments (# GT; i.e. human annotations), the number of true positives (# TP), the number of false positives (# FP), the number of false negatives (# FN). Typically, the duration of the evaluation set (see 'Time evalset') is smaller than the entire dataset and (7x) larger than the relevant action. The results in the table are based on video from a static camera.

Table 4: Performance on different cues.

Action label	Time pos (sec)	Time evalset (sec)	# pos	# GT	# TP	# FN	# FP	Precision (%)	Recall (%)	FP/ min
Eye move horizontal	53	1.9E4	333	46	26	20	30	46.4	56.5	0.09
Eye move vertical	20	1.9E4	39	39	25	3	14	64.1	89.3	0.04
Eye blinking	12	100	130	125	109	16	21	83.8	87.2	12.6
Head shake ‘No’	25	267	9	9	8	1	1	88.9	88.9	0.2
Head nod ‘Yes’	29	267	11	9	9	0	2	81.8	100	0.4
Head rotate Roll	43	234	9	9	9	0	0	100	100	0.0
Mouth smile	48	234	29	26	21	5	8	72.4	80.8	2.1
Mouth press lips	54	119	29	28	21	7	8	72.4	75.0	4.0
Scratch cheek/ear/chin	50	225	42	44	39	5	3	92.9	88.6	0.8

The evaluation in Table 4 shows that, on average (or median resp.), the precision is 78% (82%), the recall is 85% (89%) and the FP/min is 2.2 (0.4).

In the table, we see that the vertical eye movement performs much better than the horizontal eye movement. This can be explained by the fact that vertical movement leads to larger and more consistent changes of the eye. The precision appears low, but this is only due to the large evaluation set of the eye movements; the rate of FP per minute is actually very low. Blinking obtains good precision and recall. The number of FP per minute seem very high and this is caused by the training and evaluation on a part of the dataset that is rich of eye blinks (130 per minute). On average, individuals are blinking only 10 time per minute (factor 13 lower). If the detector would have been trained on more natural data, we would expect FP rates of approximately 1.0, which is also a factor 13 lower.

The results show that the performance scores of head rotations, mouth actions and scratching are above 70%. For smiling, the results per person (not shown) indicate that this algorithm either works very well or completely fails for a certain person. There is room for improvement if we could automatically identify persons that need different thresholding or different detection methods and use a hybrid model which differentiates the detection method based on personal mouth characteristics. Our datasets consist of a number of smiles per subject, but it should be noted that it is not easy to ‘act’ a genuine smile. Each of the actions was indicated with a start and a stop signal. For the genuine smiles, we noticed that a relaxed laugh appeared after the stop signal.

In Table 4, the scratching arm movements are compared to neutral arm poses. For the arm movements that are detected with moving patches, we also made additional analysis related to several different arm movements. The confusion matrix for these arm-related actions is shown in Table 5. The overall accuracy is 92%. Note that patch-based approach also helps to find subtle motions, such as fiddling and scratching.

Table 5: Confusion matrix for various arm actions with moving patches (accuracy = 92%).

Arm actions	Besides	Fiddle	Scratch	Nervous	Pick up
Arms besides the body	**100**				
Fiddle with hands in front of body		**100**			
Scratch cheek	20		**60**	20	
Nervous torso motion				**100**	
Pick up					**100**

4.3 Body-worn cameras

Several of the methods – such as the facial behavior analysis and pose estimators – were tested on body-worn video. Initial experiments (results not shown) indicated that transfer from static cameras towards body-worn cameras is feasible when the motion blur is limited and the resolution is sufficiently high.

4.4 Discussion

This paper focusses on the observability of relevant cues that can be used for the automated detection of truthful and deceptive statements. We developed and improved automated analysis techniques for several known cues to deceit, including pressing lips, eye gaze and pose. However, testing how well these automated measurements can distinguish truths from lies was not within the scope of this paper. Therefore, future work should primarily focus on determining and improving the effectiveness of actually detecting truths and lies using these cues extracted from video of static and body-worn cameras. Using mobile, body-worn cameras would further increase the practical applicability of this automated, stand-off deception detection method. Future work should also focus on a more gender balanced and representative dataset; the current group is all-male, in a limited age group. Experiments on representative 'deception' datasets are necessary to generalize and estimate the overall performance of the complete deception-detection system.

Automatic deception detection will eventually rely on the presence of several cues, meaning that different cues will be gathered before any conclusion about deceit can be given. This means that the overall true positive rates may decrease, but that the overall false positive rate will decrease as well. Some of the cues that are currently measured separately, are actually physiologically related to others. From earlier work [18][55] we learned that a combination of features improves the recognition quality because some actions correlate and some actions mutually exclude each other. For example, it is well known that eyes tend to close during a smile. Combining cues can improve individual cues and the overall deception detection. Furthermore, the frequency or intensity of observed cues can be person specific, implying that a relative change in frequency of observed cues may be more indicative than any absolute measurements. Detection of relative frequency changes will put less stress on the individual false positive rates, since it concerns a gradual change in detection characteristics. Another possible improvement could be to replace the traditional bag-of-words approach and the ACF detector by a convolutional neural network (CNN)-based deep-learning strategy [16][76], since the CNNs appears to defeat the traditional detectors in many benchmarks.

Cues with accuracies of 70% may already be valuable in a deception-detection system in practical security situations. The weak cues can be combined to create stronger deceit signals [100] and deception detection scores of 'only' 80% may already be valuable for screening. For example, deception detection may be used for rewarding reliable statements, instead of punishing unreliable statements at an airport by pre-selection of people for a fast lane, which allows a focus of scarce resources to an enriched group. A key question is how to use this type of technology in practice. Privacy and data protection should be embedded throughout all development phases of this kind of technology. Second, the creation of a baseline of behavior per person could help, but may not be practical in police work on the street. Third, detecting truthful statements might be more useful than detecting lies. For example, witnesses that made verifiable truthful statements on the street do not have to be invited to a police station for later interviewing, whereas others – without implying they lied – might. Fourth, the deception detection may be too unreliable to be used as evidence in court for convictions, but sufficient as tactical information to steer the scarce resources in the security domain.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we gave an overview of recent developments in two communities: in the behavioral-science community the developments that improve deception detection with a special attention to the observed relevant non-verbal cues, and in the computer-vision community the recent methods that are able to measure these cues. For the assessment of reliability of statements, important visual cues are related to eyes (e.g., pupil dilation, blinking, saccades, gaze aversion), facial expression (e.g., intensity, of expressions and wrinkling), head (e.g., chin raise), mouth (e.g., pressing lips and genuine smiles), hand and arm gestures (arm movement, hand movement) and full-body pose (e.g., foot movement and full-body motion). Several cues are implemented and measured at different body parts: at eyes, mouth, head, and full-body poses. We performed an experiment using multiple state-of-the-art video-content-analysis (VCA) techniques to assess the quality of robustly measuring these cues. The results showed a median precision score of 82%, a recall of 89% and a false-alarm rate of 0.4 per minute.

Future work could improve the current performance scores by combining features. The current work only focused on the measurement of cues. Future research could use these cues to perform the automatic detection of deception and false memories. Furthermore, the type of sensors can be extended from static RGB cameras to body-worn cameras and thermal cameras, and the selected cues can be extended (e.g., response latency and blushing).

ACKNOWLEDGEMENT

The work for this paper was conducted as part of the program ‘Security’ in the project ‘Sensors and systems for security’ in the Dutch top sector ‘High Tech Systems and Materials’.

REFERENCES

[1] Abouelenien, M., Mihalcea, R., Burzo, M., “Trimodal analysis of deceptive behavior,” ACM WMDD, 9-13 (2015).

[2] Ali, S., Shah, M., “Human action recognition in videos using kinematic features and multiple instance learning,” IEEE Trans. PAMI, (2008).

[3] Akehurst, L., Vrij, A., “Creating suspects in police interviews 1,” J. Appl. Soc. Psych. 29(1),192-210 (1999).

[4] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014, June. “2D human pose estimation: new benchmark and state of the art analysis,” IEEE CVPR , 3686-3693 (2014).

[5] Anil, J., Suresh, L., “Literature survey on face and face expression recognition,” IEEE Int. Conf. Circuit Power and Computing Techn., (2016).

[6] Bagley, J., Manelis, L., “Effect of awareness on an indicator of cognitive load,” Percept. Mot. Skills 49, 591-594 (1979).

[7] Baltrusaitis, T., Robinson, P., Morency, L., “3D constrained local model for rigid and non-rigid facial tracking,” IEEE CVPR, 2610-2617 (2012).

[8] Baltrusaitis, T., Robinson, P., Morency, L., “OpenFace: an open source facial behavior analysis toolkit,” IEEE Winter conf. Appl. Computer Vision, (2016).

[9] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N. and Ilic, S., “3D pictorial structures for multiple human pose estimation,” IEEE CVPR, 1669-1676 (2014).

[10] Ben-Shakhar, G., Elaad, E., “The validity of psychophysiological detection of information with the guilty knowledge test: a meta-analytic review,” J. Appl. Psych. 88(1), 131-151 (2003).

[11] Bettadapura, V., “Face expression recognition and analysis: the state of the art,” arXiv 12036722, (2012).

[12] Bhaskaran, N., Nwogu, I., Frank, M., Govindaraju, V., “Lie to me: Deceit detection via online behavioral learning,” IEEE Autom. Face and Gesture Recognition, 24-29 (2011).

[13] Bobick, A., Davis, J., “The recognition of human movement using temporal templates,” IEEE Trans. PAMI 23 (3), 257–267 (2001).

[14] Bond, C., DePaulo, B., “Accuracy of deception judgements,” Personality and Social Psychology Review 10(3), 214-234. (2006).

[15] Bouma, H., Baan, J., Burghouts, G., et al., “Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall,” Proc. SPIE 9253, (2014).

[16] Bouma, H., Eendebak, P., Schutte, K., et al., “Incremental concept learning with few training examples and hierarchical classification,” Proc. SPIE 9652, (2015).

[17] Buller, D., Burgoon, J., “Interpersonal deception theory,” Communication theory 6(3), 203-242 (1996).

[18] Burghouts, G., Schutte, K., Bouma, H., Hollander, R., “Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos,” Machine Vision and Appl. 25(1), 85-98 (2014).

[19] Burghouts, G., Schutte, K., Hove, J., et al., “Instantaneous threat detection based on a semantic representation of activities zones and trajectories,” SIVP 8(1), 191-200 (2014).

[20] Campadelli, P., Lanzarotti, R. and Lipori, G., “Precise eye localization through a general-to-specific model definition,” BMVC, 187-196 (2006).

[21] Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K., “A 3D facial expression database for visual computing,” IEEE Trans. Visualization and Computer Graphics 20(3), 413-425 (2014).

[22] Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., “Human pose estimation with iterative error feedback,” CVPR, (2016).

[23] Cherian, A., Mairal, J., Alahari, K., Schmid, C., “Mixing body-part sequences for human pose estimation,” IEEE CVPR, 2353-2360 (2014).

[24] Cristinacce, D., Cootes, T., “Feature detection and tracking with constrained local models,” BMVC 1(2), (2006).

[25] DePaulo, B., Lindsay, J., Malone, B., et al., “Cues to deception,” Psychological bulletin 129(1), (2003).

[26] Ding, C., Tao, D., “A comprehensive survey on pose-invariant face recognition,” ACM Trans. Intell. Systems and Techn. 7(3), (2016).

[27] Dionisio, D., Granholm, E., Hillix, W., Perrine, W., “Differentiation of deception using pupillary responses as an index of cognitive processing,” Psychophysiology 38(2), 205-211 (2001).

[28] Dollar, P., Appel R., Belongie S., Perona P., “Fast feature pyramids for object detection,” IEEE Trans. PAMI 36(8), 1532-1545 (2014)

[29] Efros, A., Berg, A., Mori, G., Malik, J., “Recognizing action at a distance,” ICCV 2, 726–733 (2003).

[30] Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V., “Articulated Human Pose estimation and search in almost unconstrained still images,” ETH Zurich Tech report, (2010).

[31] Ekman, P., Friesen, W., “Hand movements,” J. Comm. 22(4), 353-374 (1972).

[32] Ekman, P., “Lying and nonverbal behavior: theoretical issues and new findings,” J. Nonverbal Behavior 12(3), 163-176 (1988).

[33] Ekman, P., “Why lies fail and what behaviors betray a lie,” Credibility assessment 47, 71-81 (1989).

[34] Ekman, P., Rosenberg, E., “What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system FACS,” Oxford University press USA, (1997).

[35] Erol, A., Bebis, G., Nicolescu, M., Boyle, R., Twombly., X., “Vision-based hand pose estimation: a review,” CVIU 108(1), 52–73 (2007).

[36] Ferrari, V., Marin-Jimenez, M., Zisserman, A., “Progressive search space reduction for human pose estimation,” IEEE CVPR, (2008).

[37] Fiedler, K., Walka, I., “Training lie detectors to use nonverbal cues instead of global heuristics,” Human Comm. Research. 20(2), 199-223 (1993).

[38] Frenda, S. J., Nichols, R. M., Loftus, E., “Current issues and advances in misinformation research,” Current Directions in Psychological Science 20(1), 20–23 (2011).

[39] Garrett, B., “Convicting the innocent,” Harvard University Press. Cambridge USA, (2011).

[40] Granhag, P., Strömwall, L., “Repeated interrogations: verbal and non-verbal cues to deception,” Appl. Cogn. Psych. 16(3), 243-57 (2002).

[41] Gross, J., Levenson, R., “Emotional suppression: physiology, self-report, and expressive behavior,” J. Personality and Soc. Psych. 64(6), 970-986 (1993).

[42] Gudi, A., Tasli, H., den Uyl, T., Maroulis, A., “Deep learning based FACS action unit occurrence and intensity estimation,” IEEE Automatic Face and Gesture Recognition, (2015).

[43] Hansen, D.W. and Ji, Q., “In the eye of the beholder: A survey of models for eyes and gaze,” IEEE Trans. PAMI 32(3), 478-500 (2010).

[44] Heaps, C., Nash, M., “Comparing recollective experience in true and false autobiographical memories,” J. Experimental Psychology Learning Memory and Cognition 27(4), 920-930 (2001).

[45] Heishman, R., Duric, Z., “Using image flow to detect eye blinks in color videos,” Proc. IEEE WACV, (2007).

[46] Hurley, C., Griffin, D., Stefanone, M., “Who told you that? uncovering the source of believed cues to deception,” Int. J. Psych. Studies 6(1), 19 - 32 (2014).

[47] Ikizler, N., Cinbis, R., Pehlivan, S., Duygulu, P., “Recognizing actions from still images,” ICPR, (2008).

[48] Ionescu, C., Carreira, J. Sminchisescu, C., “Iterated second-order label sensitive pooling for 3D human pose estimation,” IEEE CVPR, 1661-1668 (2014).

[49] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., “Human3.6M Large scale dataset and predictive methods for 3D human sensing in natural environments,” IEEE Trans PAMI 36(7), 1325-1339 (2014).

[50] Johnson, S., Everingham, M., “Clustered pose and nonlinear appearance models for human pose estimation,” BMVC 2(4), (2010).

[51] Kassin, S., Gudjonsson, G. “The psychology of confessions a review of the literature and issues,” Psych. Science in the Public Interest 5(2), 33-67 (2004).

[52] Ke, Y., Sukthankar, R., Hebert, M., “Event detection in crowded videos,” ICCV, (2007).

[53] Krolak, A., Strumillo, P., “Eye-blink detection system for human-computer interaction,” Univ Access Inf Soc 11, 409-419 (2012).

[54] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M., “The computer expression recognition toolbox CERT,” IEEE Automatic Face Gesture Recognition and Workshops, 298 –305 (2011).

[55] Lefter, I., Burghouts, G., Rothkrantz, L., “Recognizing stress using semantics and modulation of speech and gestures,” IEEE Trans. Affective Computing 7(2), 162-175 (2016).

[56] Loftus, E., “Planting misinformation in the human mind: A 30-year investigation of the malleability of memory,” Learning and Memory 12(4), 361–366 (2005).

[57] Martinez, B., Valstar, M., “Advances challenges and opportunities in automatic facial expression recognition,” Advances in Face Detection and Facial Image Analysis, 63 – 100 (2016).

[58] Meservy, T., Jensen, M., Kruse, J., Burgoon, J., Nunamaker, J., “Deception detection through automatic unobtrusive analysis of nonverbal behavior,” IEEE Intelligent Systems, 36 – 43 (2005).

[59] Minkov, K., Zafeiriou, S., Pantic, M., “A comparison of different features for automatic eye blinking detection with an application to analysis of deceptive behavior,” IEEE ISCCSP, (2012).

[60] Murphy-Chutorian, E., Trivedi, M., “Head pose estimation in computer vision: A survey,” IEEE Trans. PAMI 31(4), 607-626 (2009).

[61] Niebles, J., Fei-Fei, L., “A hierarchical model of shape and appearance for human action classification,” IEEE CVPR, (2007).

[62] O’Hair, D., Cody, M., “Gender and vocal stress differences during truthful and deceptive information sequences,” Human Relations 40(1), 1-13 (1987).

[63] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M., “Deception detection using real-life trial data,” Proc. ACM Int. Conf. Multimodal Interaction, 59-66 (2015).

[64] Pérez-Rosas, V., Abouelenien, M., Mihalcea, et al., “Verbal and nonverbal clues for real-life deception detection,” Proc. Empirical Methods in Natural Language Processing, 2336–2346 (2015).

[65] Pérez-Rosas, V., Mihalcea, R., Morency, L., “Utterance-level multimodal sentiment analysis,” Proc. Assoc. Computational Linguistics (1), (2013).

[66] Pfister, T., Charles, J., Zisserman, A., “Flowing ConvNets for human pose estimation in videos." IEEE ICCV, 1913-1921 (2015).

[67] Poppe, R., Van der Zee, S., Taylor, P., Anderson, R., “Mining bodily cues,” Proc. HICSS, (2015).

[68] Poppe, R., Van der Zee, S., Heyen, D., Taylor, P., “AMAB: Automated measurement and analysis of body motion,” Behavior research methods 46(3), 625-633 (2014).

[69] Poria, S., Cambria, E., Gelbukh, A., “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” Emp. Meth. Nat. Lang. Process., 2539-2544 (2015).

[70] Rajoub, B. A., & Zwiggelaar, R., “Thermal facial analysis for deception detection,” IEEE Trans. Information Forensics and Security 9(6), 1015-1023 (2014).

[71] Ramanan, D., “Learning to parse images of articulated bodies,” Adv. NIPS, 1129-1136 (2006).

[72] Rautaray, S., Agrawal, A., “Vision based hand gesture recognition for human computer interaction: a survey,” Artificial Intelligence Review 43(1), 1-54 (2015).

[73] Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L., “Static and dynamic 3D facial expression recognition: A comprehensive survey,” Image and Vision Computing 30(10), 683-97 (2012).

[74] Sariyanidi, E., Gunes, H., Cavallaro, A., “Automatic analysis of facial affect: A survey of registration representation and recognition,” IEEE Trans. 37(6), 1113-1133 (2015).

[75] Scherer, K., Ekman, P., “Methodological issues in studying nonverbal behavior,” Handbook of methods in nonverbal behavior research, 1-44 (1982).

[76] Schutte, K., Bouma, H., Schavemaker, J., et al., “Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation,” IEEE CBMI, (2015).

[77] Shaw, J., Porter, S., “Constructing rich false memories of committing crime,” Psychological Science 26(3), 291-301 (2015).

[78] Shotton, J., Sharp, T., Kipman, A., et al., “Real-time human pose recognition in parts from single depth images,” Communications ACM, 56(1), pp.116-124 (2013).

[79] Sigal, L., “Human pose estimation,” Computer Vision A reference guide Springer, 362-370 (2014).

[80] Su, L., Levine, M., “Does ‘lie to me’ lie to you? An evaluation of facial clues to high-stakes deception,” CVIU 147, 52-68 (2016).

[81] Suarez, J., Murphy, R., “Hand gesture recognition with depth images: a review,” IEEE Symp. Robot and Human Interactive Communication, 411 – 417 (2012).

[82] Sugano, Y., Matsushita, Y. and Sato, Y., 2013. “Appearance-based gaze estimation using visual saliency,” IEEE Trans. PAMI 35(2), 329-341 (2013).

[83] Taylor, R., Hick, R., “Believed cues to deception: Judgements in self-generated trivial and serious situations,” Legal and Criminological Psych. 12, 321-331 (2007).

[84] Thurau, C., Hlavác, V., "Pose primitive based human action recognition in videos or still images," IEEE CVPR, (2008).

[85] Toshev, A., Szegedy, C., "Deeppose: Human pose estimation via deep neural networks," IEEE CVPR, 1653-1660 (2014).

[86] Tompson, J., Jain, A., LeCun, Y., Bregler, C., "Joint training of a convolutional network and a graphical model for human pose estimation," Adv. NIPS, 1799-1807 (2014).

[87] Valenti, R., Sebe, N., Gevers, T. "Combining head pose and eye location information for gaze estimation," IEEE Trans. Image Processing 21(2), 802-815 (2012).

[88] Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M. and Cohn, J.F., "FERA 2015 - Second facial expression recognition and analysis challenge," IEEE Automatic Face and Gesture Recognition 6, (2015).

[89] Van der Zee, S., Poppe, R., Taylor, P., Anderson, R., "To freeze or not to freeze: a motion-capture approach to detecting deception," Proc. HICSS, (2015).

[90] Van der Zee, S., Kleij, R. van der, Rest, J. van, Bouma, H., "Actuele ontwikkelingen in leugendetectie," Security Management, (2016).

[91] Vrij, A., Fisher, R. P., Blank, H., "A cognitive approach to lie detection: A meta-analysis," Legal and Criminological Psychology, (2015).

[92] Vrij, A., Oliveira, J., Hammond, A., Erhrlichman, H., "Saccadic eye movement rate as a cue to deceit," J. Applied Research in Memory and Cognition 4(1), 15-19 (2015).

[93] Vrij, A., Mann S., Fisher R., et al., "Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order," Law and human behavior 32(3), 253-265 (2008).

[94] Vrij, A., Leal, S., Granhag, P. et al., "Outsmarting the liars: The benefit of asking unanticipated questions," Law and Human Behavior 33(2), 159-66 (2009).

[95] Wang, N., Gao, X., Tao, D., Li, X., "Facial feature point detection: A comprehensive survey," arXiv 14101037, (2014).

[96] Weinland, D., Ronfard, R., Boyer, E., "A survey of vision-based methods for action representation segmentation and recognition," CVIU 115(2), 224-241 (2011).

[97] Wood, E., Baltrusaitis, T., Morency, et al., "Learning an appearance-based gaze estimator from one million synthesised images," Proc. ACM Eye Tracking Research and Appl., 131-138 (2016).

[98] Yang, Y., Ramanan. D., "Articulated human detection with flexible mixtures of parts," IEEE Trans. PAMI 35(12), 2878-2890 (2013).

[99] Yao, Y., Fu, Y., "Contour model-based hand-gesture recognition using the Kinect sensor," IEEE Trans. Circuits and Systems for Video Technology 24(11), 1935-1944 (2014).

[100] Yu, X., Zhang, S., Yan, Z., et al., "Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues," IEEE Trans. Cybernetics 45(3), 506 – 520 (2015).

[101] Yue, S., Harmer, K., Guo, K., et al., "Automatic blush detection in concealed information test using visual stimuli," Int. J. Data Mining Modelling and Management 6(2), 187 – 201 (2014).

[102] Zhang, Z., Hu, Y., Chan, S., Chia, L., "Motion context: a new representation for human action recognition," ECCV LNCS 5305, 817–829 (2008).

[103] Zhang, X., Sugano, Y., Fritz, M., Bulling, A., "Appearance-based gaze estimation in the wild," IEEE CVPR, 4511-4520 (2015).

[104] Zhao, L., Gao, X., Tao, D., Li, X., "A deep structure for human pose estimation," Signal Proc. 108, 36-45 (2015).

[105] Zhou, F., Torre, F., "Spatio-temporal matching for human pose estimation in video," IEEE Trans. PAMI 38(8), 1492-1504 (2016)

[106] Zhou, Y., Jiang, G. and Lin, Y., "A novel finger and hand pose estimation technique for real-time hand gesture recognition," Pattern Recognition 49, 102-114 (2016).

[107] Zuffi, S., Romero, J., Schmid, C., Black, M., "Estimating human pose with flowing puppets," ICCV, 3312-3319, (2013).

[108] Zuckerman, M., DePaulo, B., Rosenthal R., "Verbal and nonverbal communication of deception," Adv. in Exp. Soc. Psych. 14, 1-59 (1981).