Mining Bodily Cues to Deception

Ronald Poppe*, Sophie van der Zee†, Paul J. Taylor‡§, Ross Anderson† and Remco C. Veltkamp*
*Utrecht University, The Netherlands, {r.w.poppe,r.c.veltkamp}@uu.nl
†University of Cambridge, UK, {sophie.van-der-zee,ross.anderson}@cl.cam.ac.uk
‡Lancaster University, UK, p.j.taylor@lancaster.ac.uk
§University of Twente, Enschede, The Netherlands

***Abstract*—A significant body of literature has reported research on the potential correlates of deception and bodily behavior. The vast majority of these studies consider discrete bodily movements such as specific hand or head gestures. While differences in the number of such movements could be an indication of a subject's veracity, they account for only a small proportion of all performed behavior. Such studies also fail to consider quantitative aspects of body movement: the precise movement direction, magnitude and timing are not taken into account. In this paper, we present and discuss the results of a systematic, bottom-up study of bodily correlates of deception. We conducted a user experiment where subjects either were deceptive or spoke the truth. Their body movement was measured using motion capture suits yielding a large number of global and local movement descriptors. We present statistical results on the mining of bodily cues. Our analyses support the feasibility, and report the performance, of automatic deception classification from body movement.**

***Keywords*-Body motion; motion capture; deception**

I. Introduction

There is much interest in estimating whether a subject is telling the truth, for example in police interviews or in border security. Currently, judgements of subjects' veracity are made manually by humans. This introduces subjectivity and limited detection skills; human observers often ignore or misinterpret nonverbal behavior cues [13]. Overall, they perform only slightly better than chance in lab settings [22].

We investigate whether measuring a subject's nonverbal behavior automatically might be of use, for example in automatic screening and interviewing [17]. This could have several advantages over the manual systems used at present. First, an automated interviewing system can screen for subjects that appear more likely to be deceptive, so that human investigators personnel can focus on them. Second, automatic measurement allows for a more fine-grained analysis: instead of focusing on a limited set of behaviors, it can take many cues at various levels into account, such as the precise direction, magnitude and timing of movements [19]. Thus more subtle cues can be evaluated. Third, an objective, numerical representation of behavior allows analysis to be more free from potentially biasing factors such as (facial) appearance, clothing and ethnicity.

Researchers have therefore begun to tackle the automatic measurement of behavior. Eye gaze [7], facial behavior [16] and body movement [14] can be measured unobtrusively, which makes them suitable for automatic screening. The identification of specific cues indicate deceptive behavior has received a significant amount of research attention, but a meta-analysis revealed that the vast majority of researched cues are indicative of both deceptive and truthful accounts, which makes them unsuitable as single signals for deception detection [6]. This observation is also true for bodily movements.

There are several reasons for the current lack of suitable cues. Among those is the fact that the identification and verification of potential cues has predominantly taken a top-down approach. Candidate behavior cues are coded from video by hand, which is time-consuming and subjective; subtle but meaningful cues may be missed.

The alternative, which we explore here, is a bottom-up approach. Modern "big data" systems are built with the philosophy of collecting as much data as possible, extracting as many signals from it as possible, and then using statistical machine-learning techniques on a large training set to work out which combinations of signals have discriminatory power. Rather than a few bits of data (whether the subject averted their gaze or not) we collect many megabytes (precise position and orientation of 23 tags on the subject's body). These new analytic techniques do have some new problems, of course: the analysis of high-dimensional data (as where the number of possible cues is high compared to the number of subjects involved) has to be done carefully in order to manage the risk of false positives – chance findings that do not generalize across subjects and settings.

In the current study, we mine motion capture data for cues to deception. We consider a large number of signals – body pose and movement features – and investigate how their number and type influence detection results. Our analyses are aimed at assessing the tradeoff between false positives and false negatives for deception detection. While the data used in the mining of the features might be obtained under controlled conditions (e.g., in the lab using dedicated measurement equipment), the aim is to apply this work in less-constrained settings. We therefore explicitly analyze the factors that potentially hamper generalization.

The paper is organized as follows. We first discuss cues to deception and their automatic measurement. Then we describe the data used in our systematic analyses, which are presented in Section IV. We conlude by discussing our findings and their application in automatic screening.

II. Cues to Deception

The first step is to consier what signals we will attempt to extract from the mas of raw data. Fortunately there is a significant literature on the analysis of deception from observable verbal and nonverbal cues [6], [22]. We do not consider verbal signals – whether language use or content analysis, or whether from written accounts or transcripts of verbal accounts – as they might be hard to get in automated applications of interest. Nonverbal signals deception can roughly be divided into bodily, facial and paraverbal cues. Being deceptive is generally assumed to be more cognitively demanding [23], and might lead to higher levels of arousal [18]. These, in turn, might affect bodily, facial and paraverbal behavior, thereby "leaking" cues to deception [11].

Unfortunately, there is much debate over which signals correlate with deceptive behavior. A meta-analysis by DePaulo *et al.* [6] reveals that very few cues are correlated consistently, which may be caused in part by the different processes associated with lying. These include emotional responses, increased cognitive load and attempted behavioral control, each of which can lead to different types of behavior [22]. For example, if people are aware that lying-induced arousal can cause an increase in movements seen as indicative of lying, the behavioral control theory predicts that liars will try to control their movements in order to appear honest. This will lead to rigid and unnatural movement [3]. Such processes may elicit contradicting cues, and several of them can occur simultaneously, leading to a subjects behavior being a combination of both controlled and uncontrolled movement. We hypothesize that, if this is the case, differences in behavior between liars and truth-tellers are subtle and may have contributed to the low accuracy rates reported in deception research [2].

A. Automatic Behavior Analysis

Some identified cues to detection can only be measured intrusively, such as fMRI and skin conductance. Others, such as eye movements [12] and respiration, require close measurement that might be unrealistic in practical settings. We therefore focus on cues that can be observed both robustly and unobtrusively. The automatic measurement analysis of facial and bodily cues from video data has seen a lot of work in recent years [15]. Notably, Bartlett *et al.* investigated the detection of deception from facial expressions; by training machine learning classifiers on automatically measured facial expression features, they improved performance significantly over human judgement (85% and 55%, respectively) [1].

Facial expression analysis benefits from existing coding schemes such as FACS [10], platforms that enable researchers to build on each others' work. For the analysis of body movement, this work is largely still to be done. It is more complex than facial expression analysis because of the large number of degrees of freedom which gives a wide range of possible body poses; we have no agreed coordinate scheme for describing these quantitatively. There have been some recent efforts in this direction [5], [19] but they are still not commonly used. There is the further complexity that while we can infer body pose from video, this is a challenging task.

We anticipate that the increasing sophistication of motion analysis software will reduce the performance gap between motion capture and video in the near future [18]. But to circumvent issues with inaccuracy for the time being, our starting point is the automatic measurement of body pose and motion using motion capture equipment. Our work is therefore related to Duran *et al.* [9], who also employ motion capture data to find patterns of behavior. In contrast, we explicitly focus on gaining insight in the type of signals that can be used in a practical application. We do not focus on obtaining the best classification rates but explicitly explore parameters that affect the quality of the features we use as cues, in terms of generalization.

III. Data Collection and Coding

We use the data described in [21]. In the experiment, pairs of two subjects (the *interviewer* and *interviewee*) were seated facing each other. Interviewees were randomly assigned to a *truth* or *lie* condition. Prior to the interview, interviewees in the truth condition played a computer game and delivered a wallet to the lost-and-found. In the lie condition, they only looked at a description of the game, and were instructed to take a 5 pound note from the wallet. In the interview, the interviewer asked the interviewee a number of questions in a fixed order. For the *game* session, these questions were in reversed chronological order, adding to the difficulty of the task [24]. In the *wallet* session, questions were asked in normal order. In both the truth and lie conditions, interviewees were tasked with convincing the interviewer that they were telling the truth. Sessions lasted about 2.5 minutes, and were then stopped.

In total, 180 students and employees, divided into $n = 90$ pairs, took part in the experiment. We explicitly included people with different cultural backgrounds (i.e., White British and South Asian). Moreover, subjects were both male and female adults. In the present study, we do not consider the cultural background and gender of the subjects.

The body movements of both interviewers and interviewees were recorded with Xsens MVN motion capture systems. These employ inertial sensors placed in straps around the body to measure the 3D position of 23 joints in the body. Fig. 1 shows the locations of these joints. In

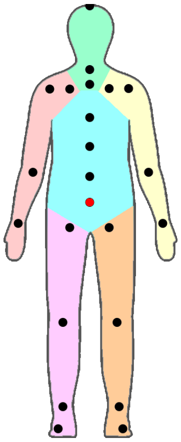

Figure 1. Location of the 23 joints. Root joint joint in red. Body parts are indicated with different colors.

this study, we use only the data of the interviewee. There might be meaningful patterns in the coordination of the behavior of both interactants (e.g. [20], [25]); but given that we consider the scenario where the interviewer could be a virtual character [17], we only want to take into account the behavior of the interviewee in this initial analysis.

A. *Space Dimension*

In line with [19], we normalized postures for global position by expressing joint postions relative to the root (i.e., pelvis). We also scaled all body parts to average lengths, to overcome differences in body dimensions between subjects. These transformations can be made without any knowledge of the subject, but help in the generalization. The resulting representation is a 66-dimensional coordinate system (22 3D joint positions). From this representation, we calculated a number of features in four different *feature types*:

- **Movement** For each of the 22 joints, we calculate the Euclidian distance between two subsequent frames. Additionally, we calculate the total amount of movement for the body parts left/right leg, left/right arm, torso and head, and for the upper and full body. The body parts are visualized with different colors in Fig. 1. The upper body contains both arms, the torso and the head. Full body contains all body parts. The total number of features of the movement type is 30.
- **Joint angle** Body movement occurs at the joints. Each joint has between one and three degrees of freedom, determined by the number of axes around which the joint can revolve. Here, we do not regard these degrees of freedom, but rather calculate the smallest angle directly by only considering the plane in which the two neighboring segments of a joint reside. We then calculate the angle between the vectors of the two neighboring segments. For example, for the left elbow we consider the vector shoulder-elbow and the vector elbow-wrist. The joint angles that we consider are those of the neck, shoulders, elbows, hips and knees. We also include the mean of these 9 angles, which brings the number of joint angle features to 10.
- **Joint distance** For a number of pairs of joints, we calculate the Euclidian distance between them. These pairs are head-left/right elbow, left hand-right hand, left hand-right elbow, right hand-left elbow, left hand-left knee, right hand-right knee, left knee-right knee, left ankle-right ankle and pelvis-right/left ankle. Including the mean over these distances, we obtain 12 features.
- **Symmetry** The joint positions are such that the x-axis runs from left to right. We mirrored the joint positions in the plane through the root, orthogonal to the x-axis. We then compared the mirrored positions of all joints to the unmirrored positions of their left/right counterpart. We use the distance between each pair as a feature. Given that these are equal for left-right counterparts, we only calculated the features for the left limbs. We finally also calculated the mean over these distances. This mean is a measure for the symmetry of the whole body pose. In sum, this 15 symmetry features.

The total number of features that we extract is 67. Although these features are somewhat arbitrarily chosen, they cover the whole body, include both local and global descriptions and carry a broad range of information. We will later investigate whether this set is optimal in terms of performance.

B. *Time Dimension*

The Xsens motion capture suits record data at a rate of 60 measurements per second. When using vision-based motion analysis, such high frequencies are currently not possible. To ensure that the findings of our study are more conviently scalable to video technology, we re-sampled our data down to 5 frames per second.

Not much is known about the time scale over which deception should be observed. We therefore include the length of our observation window as a parameter. Smaller windows allow for good representation and identification of brief, salient movements, such as a face touch or a posture shift; but they might often be devoid of discriminative movement and thus uninformative. They may also fail to capture significant longer-term behavior. For larger windows, the opposite is true. To be able to compare our findings to those reported in the literature, one window setting considers the entire session duration of approximately 2.5 minutes; we also use increasingly smaller window lengths of 1 minute, 30 seconds, 10 seconds, 5 seconds and 1 second. For each window, we calculate the mean, minimum, maximum, range and standard deviation of the feature values, which we will call *window types* in the remainder of the paper. The total dimensionality of a feature vector for each window, independent on the window size, is therefore 335 (67×5).

IV. Experimental Setup and Results

In this section, we describe the various computational experiments. We first explain the classification procedure. We then present the results, followed by a discussion of the influence of window length, feature and window type and noise.

A. Classification Procedure

We are interested in finding cues that can discriminate between truthful and deceptive accounts. To this end, we train classifiers for each feature individually on a training set, and subsequently evaluate the classifiers on test data. The data in the training and test sets are thus disjoint, which allows for the analysis of the generalization of the learned classifiers to unseen data, typically from other subjects. We use a leave-one-out cross-validation (LOOCV) approach, with the data of one pair (i.e., two sessions) in each fold. Specifically, we train on the data of $n = 89$ pairs and test on the remaining pair. We do this for all pairs and present results as the average scores over all test folds.

Our classifier is a Gaussian Naive Bayes Classifier [8], which models each class as a normal distribution. For each class c (truth or lie) and each feature i ($1 \leq i \leq 335$), we determine the mean value and standard deviation of the feature on all training samples. Given a feature value x_i in the test set, we can determine the most likely class $\hat{c}_i$ as:

$$\hat{c}_i = \operatorname*{argmax}_c \frac{1}{\sigma_{c,i}\sqrt{2\pi}} e - \frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2} \qquad (1)$$

We assume equal prior probabilities for the two classes. This is common in lab settings, to which we compare our work.

With per-feature classification, we can classify our test data based on a single feature. This allows us to look at the predictive quality of an individual feature; features with higher correct classification rates can be considered more promising features for deception detection. Additionally, we consider all features together using two different measures. First, we take the majority vote over the binary class estimates; the class which has been estimated by the majority of the per-feature classifiers is the guessed class. Second, we take the majority vote but only over the features whose distributions are statistically dissimilar with a probability of at least 95%. Typically, these should be the features for which the data of the two classes are further apart. We will refer to these two different classifiers as *all* and *stat-95*, respectively. At this point, we do not consider correlations between features.

B. Classification Results

Table I
Classification rates of different training and test sets (in percentages), obtained using all (left) / stat-95 (right) features.

	Training		
Test	Both	Game	Wallet
Both	60.0 / 65.0	58.9 / 61.7	62.2 / 66.1
Game	64.4 / 67.8	62.2 / 64.4	64.4 / 68.9
Wallet	55.6 / 62.2	55.6 / 58.9	60.0 / 63.3

Initially, we consider three sets of data: (1) from both tasks together, (2) only from the game sessions and (3) only from the wallet sessions. We evaluate all combinations of training and test sets, to gain insight in the potentially different nature of the three sets. We use one feature vector per session, corresponding to a window length of 2.5 minutes on average. As such, we use all available data. Classification results for both all and stat-95 appear in Table I. Overall classification performance is 60.0% when training and testing on both sets on all features, and improves with another 5% when only statistically significant features are considered. In the remainder, we will therefore focus on the stat-95 features. Compared to the baseline of 50%, there is a modest but important improvement. The classification performance is also better than that of humans: in this experiment, the interviewers also estimated the veracity of the interviewees, and their judgements were correct in 52.8% of the sessions.

The scores for the game task are higher than those of the wallet sessions (67.8% and 62.2%, respectively). This can be explained due to the more difficult nature of having to deceptively answer questions in reverse order [24]. This difficulty may have made the changes in behavior more salient than in the lies of the easier wallet sessions. The difference between the two tasks is also reflected in the human performance: 44.4% of the game and 61.1% of the wallet sessions were judged correctly.

Table II
Confusion matrices for the game (left) and wallet (right) sessions, when trained on stat-95 features of both.

	Actual	
Guessed	Truth	Lie
Truth	40.0%	22.2%
Lie	10.0%	27.8%

	Actual	
Guessed	Truth	Lie
Truth	40.0%	27.8%
Lie	10.0%	22.2%

Overall, the best results are obtained when training on the wallet sessions. Generalization typically improves when more data is available for training, which results in better classification rates. However, the additional availability of the game sessions does not improve the results. Rather, the game sessions appear to negatively affect the learning of the classifiers, as witnessed from the lower scores when training only on these sessions. This might be due to the more proncounced nature of the behavior in these sessions. The differences between truthful and deceptive accounts apparently do not generalize to other settings, specifically the wallet sessions. Table II shows the confusion matrices of the classifier, as trained on both features and tested on the game and wallet data separately.

In both cases there is a truth bias. In the game and wallet sessions respectively, 62.2% and 67.8% of the classifications is truthful. This leads to high recall rates for truthful accounts (80%), but markedly lower recall for deceptive ones (55.6% and 44.4%, respectively). We hypothesize that this truth bias is due to the more varied nature of deceptive accounts.

C. Window Length

In the previous section, we used a single feature vector that covered the entire duration of the session. Ideally, we would like to consider smaller windows as they would reduce the time needed to make a decision regarding the truthfulness of a subject's account. We thus evaluate several window lengths. Results are summarized in Fig. 2. In the figure, we also included stat-99 results: the majority vote over the features that have a different distribution between truth and lie samples with 99% chance.

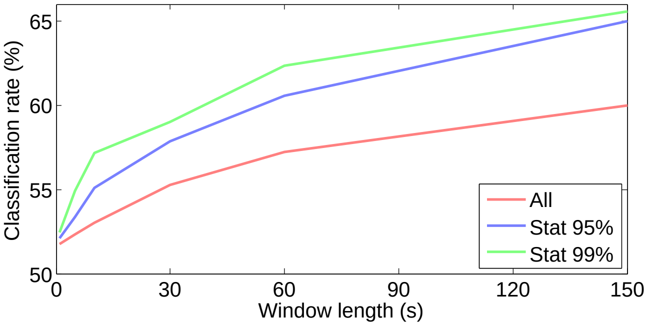

Figure 2. Classification scores in percentages for different window sizes (in second), obtained using all, stat-95 and stat-99 features.

Classification scores increase with increasing window length: the additional information that is accumulated over time is beneficial for decision performance. So decisions about a subject's veracity become more reliable when the subject's behavior is observed longer. The fact that there is more training data available for smaller windows does not help in the classification. Between the smallest (1 second) and the largest (2.5 minute) windows, there is a factor 150 more training samples. We hypothesize that many of these windows are uninformative, which might reduce the effectivity of the classifier. For windows of 1 second, the performance is barely above chance level. This increases steadily as the windows become larger, although with diminishing returns; an upper bound to the performance is to be expected.

Compared to stat-95 scores, approximately twice the window size is needed to achieve similar results using all features. For smaller windows, a similar trend can be observed between stat-99 and stat-95. As the number of features decreases from all to stat-95 to stat-99, it appears that fewer features is beneficial to the classification. To test this, we systematically varied the number of selected features from 1 to 200. Features were sorted on the significance level of the difference between the truth and lie feature distributions. Fig. 3 shows the classification rate as a function of the number of features used. Note that this number is constant over all cross-validation folds, whereas the number of features selected in stat-95 and stat-99 is generally slightly different between folds. For a window size of 2.5 minutes, the number ranges between 127 and 153 for stat-95, and between 85 and 110 for stat-99. The optimal number seems to be around 30, judging from Fig. 3. Including more features appears to decrease the classification performance as these features are less discriminating.

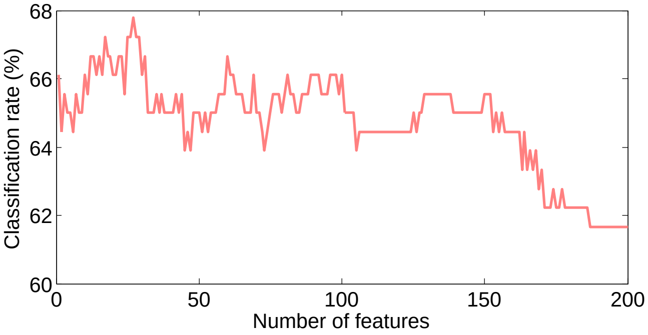

Figure 3. Classification scores in percentages for different numbers of ordered features, obtained when training and testing on both sessions with a window size of 2.5 minute.

Table III
Percentage of selected features (in stat-95) and the average classification rate per body part.

Body part	Selected	Classification rate
Left arm	70.0%	60.2%
Right arm	62.8%	58.6%
Left leg	29.3%	57.0%
Right leg	28.6%	56.8%
Head	51.3%	58.2%
Torso	44.8%	58.6%

D. Feature Type

The pool of features we evaluated covers all parts of the body. We first analyze whether some body parts are more informative than others in the detection of deception. To this end, we indicated for each feature which body parts it considers. For example, a left elbow angle considers the left arm, whereas the distance between the left hand and the right knee considers both the left arm and the right leg. Averages over all joints or distances take into account all body parts. Given that we thus linked features to body parts, we analyze how often these features contribute to the classification. We calculated, for each body part, the percentage of features linked to it that occur in stat-95. Results are summarized in Table III and visually represented in Fig. 4(a).

There are large differences between body parts in the percentage of features that are selected. Approximately 70% of the features in the arms are selected, whereas a mere 30% of the leg features is found statistically different between the

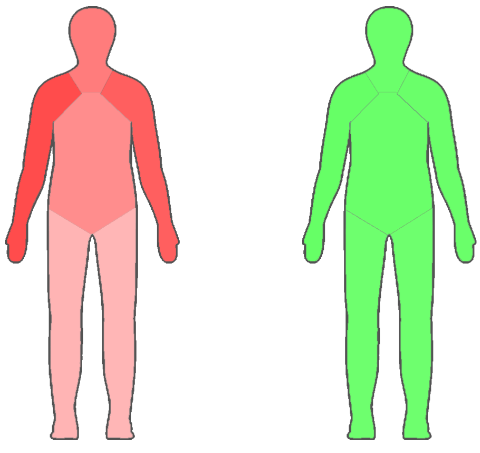

Figure 4. Visual representation of the percentage features significant at the 0.05 level (a) and their classification performance (b). Darker colors correspond to higher percentages.

truth and lie conditions. We can therefore conclude that the upper body plays a more important role in distinguishing truthful and deceptive accounts. However, the mere selection of a feature does not say anything about its quality in the classification. To this end, we present the classification rates for each body part in Table III and Fig. 4. There is little difference between body parts. The features in the arms remain the most reliable but it appears that features selected in the other body parts add to the classification. When using features from all body parts, the classification rate is 65.0%. None of the body parts alone achieves comparable rates so it is likely that different body parts are partly complementary in terms of the classification.

Table IV
PERCENTAGE OF SELECTED FEATURES (IN STAT-95) AND THE AVERAGE CLASSIFICATION RATE FOR DIFFERENT NUMBERS OF INVOLVED BODY PARTS.

Extent	Selected	Classification rate
Single body part	27.9%	57.6%
Two body parts	49.8%	57.0%
Three or more body parts	70.6%	61.2%

Some feature take into account a single body part whereas others use the postions of joints in two or more parts. We analyzed whether the extent of the feature, expressed in the number of body parts it takes into account, is of influence to the classification performance. See Table IV for a breakdown of the results. Clearly, the probability that a feature is selected increases with the number of body parts involved. This can be explained as the variance of a single feature is probably larger than the average ofr a number of features, possibly in different body parts. Consequently, differences between truthful and deceptive accounts are more often significant when considering more than a single body part. For classification performance, the involvement of three or more body parts seems to be advantageous. The most discriminating features seem to be those that average over all body parts, such as the average movement or the average symmetry. Apparently, global information is more reliable than local information.

Table V
PERCENTAGE OF SELECTED FEATURES (IN STAT-95) AND THE AVERAGE CLASSIFICATION RATE PER FEATURE TYPE.

Feature type	Selected	Classification rate
Movement	35.5%	58.4%
Joint angle	22.8%	57.1%
Joint distance	47.5%	58.2%
Symmetry	58.4%	56.4%

We used four feature types: movement, joint angle, joint distance and symmetry features. In Table V, we summarize the number of selected features and their average classification rates. Joint angles prove to be the least selected, whereas the majority of the symmetry features are selected in stat-95. Differences in classification rate between these feature types are again small. All types appear to contribute to the classification. Again, given that these individual types all score below the combined score of 65.0%, we expect that they are partly complementary.

Table VI
PERCENTAGE OF SELECTED FEATURES (IN STAT-95) AND THE AVERAGE CLASSIFICATION RATE PER WINDOW TYPE.

Window type	Selected	Classification rate
Mean	53.2%	57.8%
Maximum	43.9%	58.4%
Minimum	26.8%	52.7%
Range	37.9%	59.4%
Standard deviation	42.6%	58.0%

E. Window Type

Besides the evaluation of different window sizes, each feature was evaluated per window, for which we used five types: mean, maximum, minimum, range and standard deviation. From Table VI, it becomes clear the minimum value of a feature is often not significantly different between truthful and deceptive accounts. Especially for longer windows, the probability that values of movement are near-zero at some point is rather high. As such, it is difficult to distinguish between the two conditions. The performance of minimum features alone is also lower compared to the other window types. In contrast, more mean, maximum and standard deviation features are selected, and they also appear more promising in the classification of truths and lies. Still, the features are complementary in terms of performance.

It should be noted that different window types might become relevant for different window sizes. For smaller windows, the maximum or standard deviation might be more meaningful as these reflect sudden movement better in the feature types used.

F. Amount of Noise

While training data can typically be obtained in controlled settings, it is more likely that test data will be obtained in less-controlled settings, using convential sensors such as cameras. The accuracy of vision-based body measurements is typically lower [18]. One way to model the less accurate measurement in our analysis is to add noise on the motion captured test data. We add, to each feature, Gaussian noise with a zero mean and a standard deviation r times the standard deviation of the feature in the training data. Adding Gaussian noise is somewhat artificial as noise is typically correlated in space and time, but it shows how robust the classification is to inaccurate measurements.

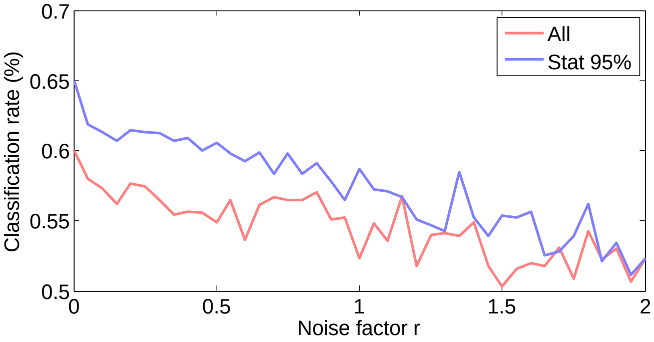

Figure 5. Classification scores in percentages for added noise with different factors r, obtained using all and stat-95 features. Scores are averaged over 5 repetitions.

Fig. 5 shows an approximately linearly decreasing classification rate for increasing noise factor r. The stat-95 classification continues to outperform a full set of features even with noise added. It is likely that a smaller set of features would be even more robust, in line with the findings reported in Section IV-C. The robustness to noise is reasonable. Even when adding noise with a standard deviation equal to that of the original data, the classification performance is still approximately 57% for stat-95. It should be noted that, especially for larger windows, the mean features are not affected much as the added noise has a mean of zero.

V. Conclusion

This paper reports an initial experiment in mining bodily cues to deception. Based on a large set of features, obtained using motion capture equipment, we have derived simple statistical classifiers to distinguish between truthful and deceptive accounts. Overall classification with all features yielded a classification performance of 60.0%, compared to a baseline of 50% and human performance of 52.8%. The selection of features based on their statistical difference between the two conditions resulted in a smaller set with an improved classification rate of 65.0%. We observed that a reduction to around 10% of the most statistically significant features would lead to a further improvement of 2-3%.

We found that features in the upper body, especially the arms, were more often significantly different between the truth and lie conditions. However, they proved to be as informative as features from other body parts in terms of classification performance. Features of a single body part scored 5-8% lower compared to the combination of features of all body parts. We therefore believe that different body parts contain complementary information for the classification. A similar observation can be made for the different feature types (e.g., movement or joint distance) and window types (e.g., mean and maximum). Despite different ratios of selected features for each type, the classification performances are comparable but consistently lower than when using all features. We also evaluated the effect of noise and found that classification rates decrease linearly in the standard deviation of the noise.

The main limitation of the current work is that we investigated features individually, and only considered classification using majority voting. We have thus ignored potential correlations between these features. Moreover, we have not looked specifically at features with complementing information. We have found that feature type, place on the body, window length and window type each carried partly complementary information that improved the classification. Exploiting combinations of features could yield more robust and, importantly, better classification rates. Currently, we considered features at a single temporal scale: we have not combined features across window sizes. It is likely that some discriminative movements are more saillant in one temporal scale while other movements are more prominent in another. There might also be patterns of behaviors over time. These patterns can be mined automatically as well, and have been shown to be promising in distinguishing truthful from deceptive accounts [4], [9]. A combination of our work with the mining of patterns seems promising.

While our analyses followed a bottom-up approach, we have only touched upon the possibilities offered in big data analysis. There is a wealth of machine learning techniques that are suitable for the type of high-dimensional data that we have considered. The current paper can aid in the selection of relevant features, time scales and design of the machine learning classifier. Eventually, these results might be used in practice, for example in automated border screening. We explicitly consider a scenario with a virtual immigration officer, and automatic measurement and analysis of the veracity of a traveller's account. This would require a different way of balancing false positives and negatives as the occurrence of deception is generally much lower than the 50% in lab settings. In a practical setting, we also underline the importance of the combination of modalities. For example, the fusion with facial expressions and eye gaze appears fruitful [12].

In sum, we have shown that truthful and deceptive accounts can be distinguished based on bodily cues that we can mine automatically. We have also shown that improvements can be made. With the directions of further research we have outlined, we expect that bodily cues to deception can improve the current performance of automatic screening.

Acknowledgments

The authors would like to thank Lieke Rotman, MSc, and Prof. Ellen Giebels for the important role they played in the design and data collection of this study. This publication was supported by the Dutch national program COMMIT. Dr. Sophie van der Zee was supported by the UK Engineering and Physical Sciences Research Council.

References

[1] M. S. Bartlett, G. C. Littlewort, M. G. Frank, and K. Lee. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, 24(7):738–743, 2014.

[2] C. F. Bond Jr. and B. M. DePaulo. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234, 2006.

[3] D. B. Buller and J. K. Burgoon. Interpersonal deception theory. *Communication Theory*, 6(3):203–242, 1996.

[4] J. K. Burgoon, J. G. Proudfoot, R. Schuetzler, and D. Wilson. Patterns of nonverbal behavior associated with truth and deception: Illustrations from three experiments. *Journal of Nonverbal Behavior*, 38(3):325–354, 2014.

[5] N. Dael, M. Mortillaro, and K. R. Scherer. The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior*, 36(2):97–121, 2012.

[6] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003.

[7] D. C. Derrick, K. Moffitt, and J. F. Nunamaker Jr. Eye gaze behavior as a guilty knowledge test: Initial exploration for use in automated, kiosk-based screening. In *Proceedings of the Hawaii International Conference on System Sciences*, Poipu, HI, 2010.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, second edition, 2000.

[9] N. D. Duran, R. Dale, C. T. Kello, C. N. H. Street, and D. C. Richardson. Exploring the movement dynamics of deception. *Frontiers in Psychology*, 4(A140):1–16, 2013.

[10] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[11] P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[12] A. C. Elkins, S. Zafeiriou, M. Pantic, and J. K. Burgoon. *The Oxford Handbook of Affective Computing*, chapter Unobtrusive Deception Detection. xoford, UK: Oxford University Press, to appear.

[13] T. R. Levine, R. K. Kim, and J. P. Blair. (In)accuracy at detecting true and false confessions and denials: An initial test of a projected motive model of veracity judgments. *Human Communication Research*, 36(1):82–102, 2010.

[14] S. Lu, G. Tsechpenakis, D. N. Metaxas, M. L. Jensen, and J. Kruse. Blob analysis of the head and hands: A method for deception detection. In *Proceedings of the Hawaii International Conference on System Sciences*, page 20c, Waikoloa, HI, 2005.

[15] D. Metaxas and S. Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31(6–7):421–433, 2013.

[16] N. Michael, M. Dilsizian, D. N. Metaxas, and J. K. Burgoon. Motion profiles for deception detection using visual cues. In *Proceedings of the European Conference on Computer Vision - Part IV*, pages 462–475, Hersonissos, Greece, 2010.

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

[18] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1–2):4–18, 2007.

[19] R. Poppe, S. van der Zee, D. K. J. Heylen, and P. J. Taylor. Amab: Automated Measurement and Analysis of Body motion. *Behavior Research Methods*, 46(3):625–633, 2014.

[20] S. van der Zee. *The effect of cognitive load on nonverbal mimicry in interview settings*. PhD thesis, Lancaster University, Lancaster, UK, 2013.

[21] S. van der Zee, R. Poppe, P. J. Taylor, and R. Anderson. To freeze or not to freeze: A motion-capture approach to detecting deceit. In *Proceedings of the Hawaii International Conference on System Sciences*, Kauai, HI, 2015.

[22] A. Vrij. *Detecting Lies and Deceit: Pitfalls and Opportunities*. John Wiley and Sons, 2008.

[23] A. Vrij, P. A. Granhag, and S. Porter. Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11(3):89–121, 2010.

[24] A. Vrij, S. Mann, R. Fisher, L. Leal, B. Milne, and R. Bull. 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.

[25] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. E. Dunbar, M. L. Jensen, J. K. Burgoon, and D. N. Metaxas. Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues. *IEEE Transactions on Cybernetics*, to appear.