2021 IEEE Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-0477-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/WACV48630.2021.00019

Identity Unbiased Deception Detection by 2D-to-3D Face Reconstruction

Lê Minh Ngô[1,4] Wei Wang[1] Burak Mandira[2] Sezer Karaoğlu[1,4] Henri Bouma[3]
Hamdi Dibeklioğlu[2] Theo Gevers[1,4]
[1] University of Amsterdam [2] Bilkent University [3] TNO [4] 3DUniversum
{l.m.ngo, th.gevers}@uva.nl, burak.mandira@bilkent.edu.tr
henri.bouma@tno.nl, s.karaoglu@3duniversum.com, dibeklioglu@cs.bilkent.edu.tr

Abstract

Deception is a common phenomenon in society, both in our private and professional lives. However, humans are notoriously bad at accurate deception detection. Based on the literature, human accuracy of distinguishing between lies and truthful statements is 54% *on average, in other words, it is slightly better than a random guess. While people do not much care about this issue, in high-stakes situations such as interrogations for series crimes and for evaluating the testimonies in court cases, accurate deception detection methods are highly desirable. To achieve a reliable, covert, and non-invasive deception detection, we propose a novel method that disentangles facial expression and head pose related features using 2D-to-3D face reconstruction technique from a video sequence and uses them to learn characteristics of deceptive behavior. We evaluate the proposed method on the Real-Life Trial (RLT) dataset that contains high-stakes deceits recorded in courtrooms. Our results show that the proposed method (with an accuracy of* 68%*) improves the state of the art. Besides, a new dataset has been collected, for the first time, for low-stake deceit detection. In addition, we compare high-stake deceit detection methods on the newly collected low-stake deceits.*

1. Introduction

Deceptive behavior is frequently displayed in daily life, yet, recognition of such behavior or lies is not an easy task for humans. On average, people can correctly classify only 47% of lies and 61% of truthful statements [4].

Therefore, reliable methods for deception detection is an important need specifically for high-stakes situations such as court cases, and suspect/witness interrogations for further investigations and low-stakes situations to improve our daily communications. However, the ubiquitous polygraph, the most commonly known deception detection mechanism is unreliable [13].

Invasive approaches such as PET (positron emission tomography) and fMRI (functional magnetic resonance imaging) based methods perform better but they are neither fully reliable nor practical in deception detection where compactness or portability is required. Besides, the invasive nature of such mechanisms leaves them to be easily tricked by skilled deceivers [13]. Hence, deception detection requires non-invasive and covert methods for accurate detection. The difficulty in non-invasive deception detection lies in the weakness of external cues, since a large volume of work indicates that deceits are barely evident in behaviour [18].

Recent developments in computer vision, along with the availability of deceptive behavior videos, have increased the research interest on deceit detection from visual patterns. The driving mechanism behind this ambition is the (subconscious) leakage of behavioral cues to deception [18]. These cues are often weak, very fast, or subjective, making them hard to interpret by humans. Recent studies on automated deception detection [27] exploits different behavioral modalities such as facial actions/expressions, head pose/movement, gaze, hand gestures, and even vocal features in the analysis [1, 27]. In contrast, our work focuses solely on temporally coherent disentangled facial cues.

High-level visual features used in the literature [27] such as facial action units are prone to errors due to challenging environmental conditions (i.e. illumination, viewpoint, occlusion, etc.). Thus, features extracted under challenging conditions can be unreliable. In this paper, to cope with such issues, we propose to exploit 2D-to-3D face reconstruction to obtain an effective low-level representation for more reliable deception detection. 2D-to-3D face reconstruction aims at decomposing a face image into its components such as 3D facial geometry, expression, skin reflectance, head pose, and illumination parameters. Expression and head pose components are expected to carry important information for deceit detection [25].

Although a successful decomposition has been a backbone for many face-related computer vision tasks (*e.g.* face recognition, emotional expression recognition, head pose

978-1-6654-0477-8/21/$31.00 ©2021 IEEE
DOI 10.1109/WACV48630.2021.00019

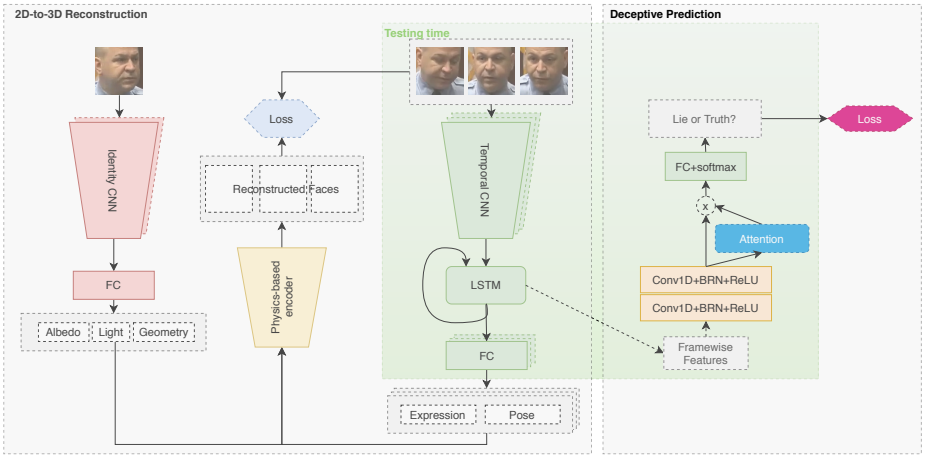

Figure 1. Architecture overview. Our proposed method decomposes temporally related features (expression and pose) from identity and environment properties by simultaneously training two CNNs (Identity and Temporal CNNs) to produce two sets of features using 2D-to-3D reconstruction. Features from the Temporal CNN are used for Deceptive Prediction.

estimation, etc.), this work is the first one that exploits face reconstruction for deceit detection. To this end, we propose an identity (i.e. facial geometry and skin reflectance) and environment (i.e. illumination) unbiased deceit detection system. Unbiasedness is achieved by conditioning on facial expression and head-pose related features alone. Facial expression and head-pose feature space are disentangled from other properties by simultaneously learning two separate networks, one to predict the identity and environment parameters and another for temporally related features (i.e. expression and head pose). Our results show that the proposed novel method for deception detection improves the state of the art high-stakes deceit detection, as well as it provides comparable results with the methods which make use of manually annotated facial attributes (*e.g.* facial actions/expressions, gaze, and head movement).

All prior automatic methods have been focusing on high-stakes deceit detection. There is no study available for automatic low-stakes deceit detection also because there is no low-stakes deceit detection dataset available. In our work, a novel Low-Stakes Deceit dataset has been collected with 624 high-res recordings of 312 subjects. To the best of our knowledge, the Low-Stakes Deceit dataset is the first and the only dataset available for low-stakes deceit detection. Besides, we use the dataset also to evaluate the existing automatic high-stakes deceit detection methods on the full spectrum of deceit.

To summarize, our contribution is four-fold:

- A novel method is proposed for deception detection on videos. The proposed method disentangles head pose and facial expression from facial identity (i.e. skin reflectance and 3D facial geometry) and illumination, using 2D-to-3D face reconstruction.
- The Real-Life Trial dataset has been cleaned and state-of-the-art high-stakes deceit detection methods have been re-evaluated using Leave-One-Person-Out (LOPO) validation.
- The proposed method outperforms the existing state-of-the-art and outperforms professional experts on the high-stakes deceit detection task.
- A new Low-Stakes Deceit (LSD) dataset is introduced. To our knowledge, it's the first visual dataset for low-stakes deceit detection. For the first time, we create a benchmark for state-of-the-art automatic high-stake deceit detection methods on low-stake deceit detection. The dataset will allow further research to be done on low-stake deceit detection.

2. Related work

2.1. Deception detection

At the basis of deception detection through nonverbal cues stands the leakage hypothesis, which states that –if the stakes of a lie are high enough– involuntary, subconscious cues of deceit will emerge from a liar [18]. One can divide observable cues into physiological cues, body language cues, and facial cues. One of the problems with intangible constructs such as deceit is that these cues range from highly objective ones (vocal pitch) to highly subjective measurements (facial pleasantness). Hence, this section aims to provide an overview of objective, non-verbal cues that are relevant to the scope of using visual features for deception detection.

Concerning facial cues, a multitude of signals have been identified to correlate with deceit, such as lip pressing [7],

smiling and pupil dilation, and facial rigidity [29]. However, the studies often find contradictory results [5, 38]. Besides, performance is highly dependent on the data used for training and validation, with some datasets being noticeably easier than others [39]. Secondly, the circumstances under which the lies were elicited are influential: multiple studies indicate that deceptive cues increase in magnitude with increased cognitive load [37]. Hence, the final application and training data should have comparable cognitive load during data recording.

Micro-expressions pose another viable source of information [41], even though other studies have shown that only a small amount of people exhibit micro-expressions when lying [10]. Facial action units (AUs) are also found to be informative for deceit detection [27].

One of the most recent methods of automated deceit detection is proposed by Morales *et al*. [27]. This method fuses information from audio-visual modalities, where visual features in the form of 408 cues, including gaze, orientation, and FACS information, are extracted using OpenFace [2] and later fused with verbal and acoustic features. Fusion occurs through a concatenation of statistical functional vectors, after which random forests and decision trees are used for deception classification. Differently, [30] presents a baseline method for their introduced Real-Life Trial dataset, which models manually coded visual features such as expression, head movement, and hand gestures together with speech transcriptions using random forests and decision trees.

In literature, deceit is typically categorized into high-stakes (hold severe consequences for the liar) and low-stakes (simple lies that individuals get away with most often). All prior automated deceit detection methods, to our knowledge, have been focusing on the high-stakes deceit detection problem. Thus, there was no research has been done on low-stakes deceit detection. Low-stakes deceit detection is considered more challenging than high-stakes deceit detection since people in high-stakes situations are expected to behave more nervously [25]. In more than 30 human behavior studies on low-stakes and high-stakes deceit conducted by other researchers an average accuracy of 55% has been achieved by *professional experts* on low-stakes in comparison with 67% for high-stakes [28].

2.2. Monocular face reconstruction

The decomposition of image components requires inverting the complex real-world image formation process. The reconstruction by inverting image formation is an ill-posed problem because an infinite number of combinations can produce the same 2D image [3]. In general, we can categorize face reconstruction methods into two groups, namely, iterative [3, 14, 34, 35] and deep learning based [33]. Iterative approaches try to optimize parameters by minimizing the error between projected (reconstructed face) and the original image in an iterative (analysis-by-synthesis) manner [3]. The energy functions are mostly non-convex. The good fitting can only be obtained by close initialization to the global optimum, which is only possible with some level of control during image capture. Since these approaches are computationally expensive they are not preferred in this paper.

Deep learning based methods, to reconstruct a face from a single monocular image, typically uses either data augmentation techniques to regress prediction to be close to the ground truth [15, 21, 32] or applies the similar analysis-by-synthesis approach to train the neural network using a physically plausible image formation model [9, 15, 24, 33]. These methods produce sufficient reconstruction quality for certain tasks, however, they sacrifice details in order to be tractable for challenging, unconstrained images. Since such methods cannot avoid expression information to be leaked in 3D facial geometry, it is likely that there is an information loss while capturing expression. To reliably capture facial movements, the separation of 3D facial geometry and expression components are quite important.

Some works have been proposed to overcome such issues by using RGB videos instead of single monocular images [14, 34, 35]. However, these works are based on the iterative optimization approach (requires energy minimization for new input data). Convolutional Neural Network (CNN) architectures are recently explored for video-based dense real-time face reconstruction. In this paper, we present a novel identity-aware, dense, and real-time face reconstruction CNN pipeline which receives RGB videos as input. Unlike previous monocular reconstruction methods, our method disentangles identity-related features (i.e. 3D facial geometry and reflectance) and illumination from temporally dependent parameters (i.e. expression and head pose) by simultaneously learning two CNNs for those sets of parameters using 2D-to-3D reconstruction. Disentanglement of temporally dependent features is important for deception detection since it allows our method to be unbiased towards subject identity and recorded environment.

3. Proposed method

Our goal is to predict if a talking person is lying based on visual input i.e. face image. A sequence of RGB face images $\{\mathbf{I}_i\} \in \mathbb{R}^{W\times W\times 3}$ is passed to the Convolutional Neural Network (CNN) backbone to predict head-pose and facial-expression related features. Expression and head-pose are disentangled from other properties using 2D-to-3D reconstruction, which simultaneously learns latent face attributes together with environmental conditions. Constraining prediction on expression and head-pose alone allows us to be unbiased from facial identity and environment conditions which are irrelevant for deceit detection. Prior psy-

chology studies have shown expression and pose-related behaviors such as eye contact, facial twitching, pauses, stuttering, and hesitance to be indicative of lie detection [25]. Temporal features (i.e. expression and head pose) are used further in the second CNN to produce the final deceit detection. An overview of our method is shown in Fig. 1.

3.1. Modeling deceptive behaviour

We model lie detection as a Multiple Instance Learning problem [19]. Given features extracted from video frames, our model assigns a single label (lie/truth) for the entire video. For a video annotated as a lie, we assume that there is at least one sub-sequence, where the person shows a deceptive behavior. For a video annotated as a truth, we assume that everything in the video is a truth. Thus, any sequence of frames which contains a lie sub-sequence is labeled as a lie. Given expression and pose related frame-wise features our deception prediction model extracts local temporal features $\mathbf{h_t} \in \mathbb{R}^{T \times C}$ using two layers of 1D-convolutions over the temporal dimension, where T is a sequence length, and C is the number of filters. Attention block *Att* weighs features based on their usefulness for the final task. Final linear layer *fc* with sigmoid is used to produce final prediction $\mathbf{y}$.

$$\mathbf{y} = \sigma\left(\text{fc}\left[\sum_t softmax(Att(\mathbf{h_t})) \cdot \mathbf{h_t}\right]\right). \quad (1)$$

3.2. Expression and pose features disentanglement

A 2D face image $\mathbf{I}_i$ can be described using latent parameters $\mathcal{P} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{t}\} \in \mathbb{R}^{257}$ from which the original face can be reconstructed. We use CNN to predict those parameters. $\boldsymbol{\alpha} = \{\alpha_i\}$, $\boldsymbol{\beta} = \{\beta_i\} \in \mathbb{R}^{80}$ and $\boldsymbol{\delta} = \{\delta_i\} \in \mathbb{R}^{64}$ are parameters correspond to 3D face geometry, albedo and expression; $\boldsymbol{\gamma} \in \mathbb{R}^{9 \times 3}$ describes scene illumination; $\boldsymbol{\omega} \in \mathbb{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$ describe face rotation and translation.

Our model consists of two CNN backbones. The first component, **Identity CNN**, is used to predict identity and environment related parameters (identity geometry $\boldsymbol{\alpha}$, albedo $\boldsymbol{\beta}$ and lighting condition $\boldsymbol{\gamma}$). Face image is passed to the MobileNetV2 backbone [31]. Its last layer is replaced by a fully connected layer with linear activation to predict $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ parameters. The second component, **Temporal CNN**, is used to predict face expression $\boldsymbol{\delta}$ and object transformations $\boldsymbol{w}, \boldsymbol{t}$ based on a sequence of RGB face images $\{\mathbf{I}_i\} \in \mathbb{R}^{W \times W \times 3}$. MobileNetV2 backbone is followed by a recurrent layer LSTM and a fully connected layer with linear activation to predict $\boldsymbol{\delta}, \boldsymbol{w}, \boldsymbol{t}$.

We use LSTM to capture temporal relations between video frames and as an expression and pose-related feature space for the Deceptive Prediction network.

Disentanglement of expression and pose feature space from other properties is achieved by simultaneously learning all latent parameters $\mathcal{P}$ using 2D-to-3D reconstruction via *Physics-based encoder*. Disentanglement is an important property for our deception framework since it allows it to be unbiased towards identity and environment properties by assuming that lie cues are dependent on temporally related properties (expression, pose) only.

3.3. 2D-to-3D reconstruction

Albedo and geometry. 3D face geometry and albedo are parametrized using a multi-linear PCA model [16]. Face geometry is represented as a point cloud $\mathbf{X}$ in the Euclidean space with the corresponding albedo attributes $\mathbf{B} \in \mathbb{R}^{N \times 3}$.

$$\mathbf{X} = \mathbf{A_{geom}} + \mathbf{P_{id}}[\boldsymbol{\alpha} \cdot \boldsymbol{\sigma}_{id}] + \mathbf{P_{exp}}[\boldsymbol{\delta} \cdot \boldsymbol{\sigma}_{exp}], \quad (2)$$

$$\mathbf{B} = \mathbf{A_{alb}} + \mathbf{P_{alb}}[\boldsymbol{\beta} \cdot \boldsymbol{\sigma}_{alb}], \quad (3)$$

where $\mathbf{A_{geom}}, \mathbf{A_{alb}} \in \mathbb{R}^{N \times 3}$ are the mean face geometry and skin albedo; $\mathbf{P_{id}}, \mathbf{P_{alb}} \in \mathbb{R}^{N \times 3 \times 80}, \mathbf{P_{exp}} \in \mathbb{R}^{N \times 3 \times 64}$ are principal components of PCA models for face identity, albedo and expression respectively; together with their standard deviations $\boldsymbol{\sigma}_{id}$, $\boldsymbol{\sigma}_{alb} \in \mathbb{R}^{80}$, $\boldsymbol{\sigma}_{exp} \in \mathbb{R}^{64}$.

Face transformation. We model face movement in the scene using 6DoF transformation $\mathbf{T}$. Rotation matrix $\mathbf{R}(\boldsymbol{w}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{3 \times 3}$ is represented in $\boldsymbol{\omega} \in \mathbb{R}^3 \in \mathbb{SO}(3)$, and translation $\boldsymbol{t} \in \mathbb{R}^3$ in x, y, z directions.

Illumination model. Illumination changes are modeled using the first 3 bands of spherical harmonics basis function $\mathbf{H}_j$ assuming face is a Lambertian surface [36]. The intensity of the i-th vertex c_i is defined as a product of vertex reflectance b_i and a shading component.

$$c_i = b_i \sum_{j=1}^{3^2} \gamma_j \mathbf{H}_j\Big(\mathbf{R}(\boldsymbol{\omega})\mathbf{n}_i\Big), i \in 1..N, \quad (4)$$

where $\mathbf{n}_i$ is a vertex normal of the i-th vertex. We define illumination parameters γ_j separately for each RGB channels, and consequently have 27 parameters in total. Vertex normal is estimated based on 1-ring triangle neighbors. Triangle topology is known from the face morphable model.

Projection model. An obtained 3D point cloud $\mathbf{X}$ is mapped into a 2D plane by applying a rigid transformation $\mathbf{T}$ and perspective transformation $\mathbf{\Pi}$ which is a product of projection $\mathbf{V}$ and viewport $\mathbf{P} \in \mathbb{R}^{4 \times 4}$ matrices:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \\ \hat{d} \end{bmatrix} = \underbrace{[\mathbf{V}] \times [\mathbf{P}]}_{\mathbf{\Pi}} \times \underbrace{\begin{bmatrix} \mathbf{R}(\boldsymbol{\omega}) & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}}_{\mathbf{T}} \times \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \quad (5)$$

$\hat{u}, \hat{v}$ coordinates, and depth can be obtained by division by the homogeneous coordinate $\hat{d}$. The focal length is assumed to be fixed and principal points to be in the middle of

the projection screen. $\hat{u}, \hat{v}$ together with vertex color c are used for producing the final reconstructed face.

3.4. Training losses

We use cross-entropy loss between ground-truth labels $\mathbf{y_{gt}} \in \{0, 1\}$ and predictions $\mathbf{y} \in [0, 1]$ to train our Deceptive Prediction pipeline.

$$\mathcal{L}_{dec} = \mathbf{y_{gt}} \cdot \log \mathbf{y} + (1 - \mathbf{y_{gt}}) \cdot \log(1 - \mathbf{y}). \quad (6)$$

For 2D-to-3D reconstruction we employ the energy minimization strategy of [33]. In total our loss consists of 3 main components: landmark loss E_{land}, vertex-wise photometric loss E_{vert} and regularization term E_{reg}.

$$\mathcal{L} = w_{land}E_{land} + w_{vert}E_{vert} + E_{reg}. \quad (7)$$

L_2 difference between landmark projections p from a predicted 3D face model and ground truth landmark l_j are used. In total, we use $|\mathcal{F}| = 48$ landmarks for optimization covering eyebrows, eye corners, nose, mouth, and chin.

Figure 2. Sample video frames from the RLT dataset. The dataset contains videos of trials under different lighting conditions, pose, with multiple people in the scene. Some of the videos are heavily occluded and don't contain visible facial features.

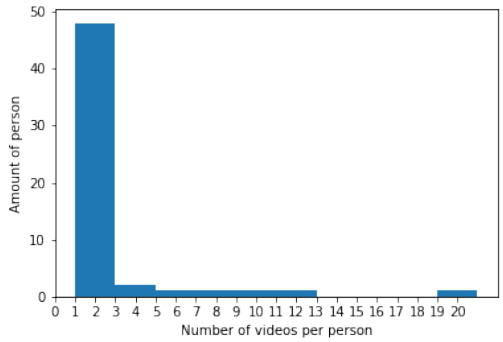

Figure 3. Distribution of videos per person in the RLT dataset. RLT dataset is imbalanced, with a few identities with a large number of videos.

$$E_{land} = \frac{1}{|\mathcal{F}|} \sum_{j \in \mathcal{F}} \|p_{k_j} - l_j\|_2^2, \quad (8)$$

where we define k_j as an annotated vertex index of the j-th landmark on the 3D model.

We define photometric loss as a $L_{2,1}$ difference [11] between vertex intensity color and its corresponded color from the original image. To find an intensity color on image space we perform interpolation. We filter out vertices which contribute to the photometric loss based on normal direction, $|\mathcal{V}|$ is the number of vertices.

$$E_{vert} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \|\mathbf{c}_i - \mathbf{X}_{\hat{u}_i, \hat{v}_i}\|_2, \quad (9)$$

We use Tikhonov regularization [36] to enforce parameters to be in the plausible range.

$$E_{reg} = w_\alpha \sum_{i=1}^{80} \alpha_i^2 + w_\beta \sum_{i=1}^{80} \beta_i^2 + w_\delta \sum_{i=1}^{64} \delta_i^2. \quad (10)$$

4. Datasets

Real-Life Trial dataset. We employ the Real-Life Trial dataset [30] which contains 121 videos from real-life high-stakes scenarios that are publicly available. See Fig. 2 for visual samples from dataset. It has 61 deceptive and 60 truthful trial clips of 21 female and 35 male subjects whose ages vary between 16 and 60. The average duration of videos is about 28 seconds. When constructing the dataset, Perez-Rosas *et al.* [30] enforce some visual constrains for videos such as the defendant or witness and his or her face should be identified during most of the footage.

Nonetheless, the video quality is noisy: the defendant's face is not always clearly visible in the video, the defendant and witnesses may appear both in the scene. Previous works, which rely on the confidence of the face detector alone, extract visual features from both defendant and witnesses for deceptive prediction. In addition, the dataset is unbalanced: the amount of videos per identity differentiates significantly (Fig. 3). One performing K-fold validation might include videos of the same person both in testing and training split, and hence achieving high accuracy. Consequently, for each video in the dataset, we have manually annotated all witnesses and removed them from the video sequence. If multiple faces appear in the scene, we remove all faces except the defendant. 5 videos without faces in the scene / occluded faces have been removed which leaves 116 videos for LOPO validation.

Low-Stakes Deceit (LSD) dataset. We have collected a new dataset of low-stakes deceit which contains 624 high-res recordings of 143 males and 169 females interviewees

Figure 4. Sample video frames from our newly collected LSD dataset. The dataset contains video of the similar lighting conditions, pose with a single person in the scene.

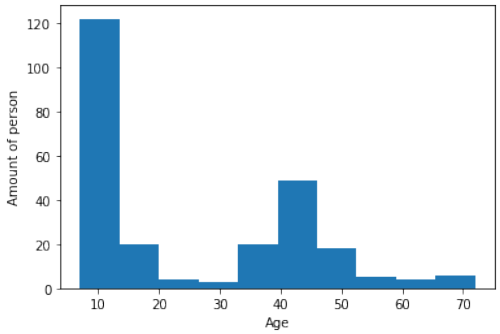

Figure 5. Age distribution of interviewees in our novel LSD dataset. Distribution is heavily tilted to 10 since it has been collected in the Science Museum popular among children.

under a controlled environment. Data collection was carried out as a part of Science Live, the innovative research programme of Science Center NEMO[1]. To our knowledge, our dataset is the first visual dataset available for studying low-stakes deceit in the literature. The age of participants varies between 7 to 72 years (Fig. 5). Among them, 209 participants speak Dutch and 103 participants speak English. Participants are facing the camera frontally and answer the interviewer's questions. The environmental conditions (*e.g.* illumination, background) are remained the same during whole recording sessions.

The interviewees are asked to describe two abstract scenes: one on the visual card provided to the interviewee beforehand, and another which s/he did not see in advance. We define the first description as truth and the second as a lie. As a result, we have collected 2 recordings for every 312 identities with positive and negative labels. *Since the experiment setting doesn't imply a punishment for the contrived answer, the collected recordings can be used to study low-stakes lie.* Samples of our novel LSD dataset are shown in Fig. 4. We asked interviewees to judge peer recordings and used this information to measure the human accuracy on this dataset.

[1]Science Center NEMO, Amsterdam, http://www.e-nemo.nl.

5. Implementation details

We train our 2D-to-3D face reconstruction network for 200K iterations on 300VW [8] and CelebA datasets [26] using a batch size of 5 and Adam optimizer [23] with learning rate of 10^{-5}. Loss weights are set to be $w_{vert} = 1.92$, $w_{land} = 0.0019$, $w_{\alpha} = 2.9 \times 10^{-5}$, $w_{\beta} = 4.93 \times 10^{-8}$, $w_{\delta} = 2.32 \times 10^{-5}$.

For training the Deceptive Prediction network we use RLT dataset for high-stakes lies and our newly collected LSD dataset for low-stakes lies. Models are trained on the batch size of 8 for 100 epochs. Early stopping is performed based on the validation score. We use Adam optimizer with a learning rate of 10^{-3}.

300VW contains video sequences with annotated 68 landmarks for each frame. We crop faces based on a bounding box on ground truth landmarks with 10% expansion. We process CelebA using dlib [22] for face detection and FAN [6] for landmark detection. In total, we have collected 94K images from 300VW coming from 49 videos and 200K images from CelebA. Images from RLT and LSD datasets are processed in the same manner.

For each video sequence of 300VW we randomly select a cropped face as an input for the Identity-CNN. We randomly sample a sequence of 3 crop faces with a random step size from 1 to 5 frames as an input for the Temporal-CNN. For CelebA we assume that we have a 1-frame video sequence for each image. Images are randomly flipped to augment the dataset size. We train the model alternating CelebA and 300VW batches.

MobileNetV2 backbones are pretrained using ImageNet. We add offset to the 0-th band SH coefficient and z-translation to make sure the initial 3D face model has a plausible initial illumination and is centered in the middle of the screen. Basel Face Model 2017 [16] is used for 3D face geometry, albedo and expression.

6. Experiments

In this section, we provide the details and results of conducted experiments. We start with a comparison with other methods on the high-stakes lies task. Next, we evaluate how methods designed for the low-stakes lies task performs in the low-stakes settings. Last, we provide additional analysis of age and gender effects. We considered lie as *positive* and truth as *negative* throughout the experiments when calculating accuracy, precision, and recall.

6.1. Baselines

In this section, we describe baselines for our experiments. **Morales *et al.* [27]** is tested with a decision tree (DT) and random forest (RF) classifiers with default parameters as in the papers. OpenFace [2] is used to extract facial features in default output (i.e. basics, gaze, pose, 2D and 3D

Type	Model	Feature	Accuracy	Precision	Recall
Manual	Perez-Rosas *et al*. [30] DT*	Hand-labeled features	0.67	0.64	0.74
	Perez-Rosas *et al*. [30] RF*	Hand-labeled features	0.71	0.70	0.70
Automatic	Morales *et al*. [27] DT*	OpenFace features	0.50	0.48	0.38
	Morales *et al*. [27] RF*	OpenFace features	0.56	0.57	0.40
	3D-ResNext [17]	CNN features	0.59	0.57	0.63
	DARE [40] RF	Motion Features	0.54	0.55	0.42
	Time-CNN [12, 42]	LSTM features	0.47	0.44	0.26
	Ours	LSTM features	0.68	0.66	0.72

Table 1. State-of-the-art comparison on the high-stakes lies task using RLT dataset.
*: only facial features are used

Model	Feature	Accuracy (EN/NL)	Precision (EN/NL)	Recall (EN/NL)
Human	Visual + Audio	0.516	-	-
Morales *et al*. [27] DT	OpenFace features	0.55 / 0.52	0.55 / 0.52	0.57 / 0.53
Morales *et al*. [27] RF	OpenFace features	0.55 / 0.50	0.54 / 0.50	0.57 / 0.45
3D-ResNext [17]	CNN features	0.53 / 0.54	0.53 / 0.55	0.53 / 0.52
Time-CNN [12, 42]	LSTM features	0.47 / 0.51	0.47 / 0.52	0.51 / 0.44
Ours	LSTM features	0.54 / 0.52	0.53 / 0.52	0.64 / 0.65

Table 2. State-of-the-art comparison on the low-stakes lies task using LSD dataset.

facial landmark locations, rigid and non-rigid shape parameters, action units) and apply some statistical metrics (max, min, mean, median, std, kurtosis, skewness, etc.) to create one feature vector per video. **Perez-Rosas *et al*. [30]**, which is the basis for Morales *et al*. [27], is also implemented with a decision tree (DT) and random forest (RF) classifiers with default parameters as mentioned in their papers. They use manually labeled features. Since our system focuses only on facial features, we excluded hand-related features from their experimental setup to obtain comparable results. **3D-ResNext [17]** is pretrained on Kinetics dataset [20] and finetuned starting from the third block. During training, a random temporal sampling of 30 frames is used. In inference, we use a non-overlapping sliding window of size 30 and take the mean scores of windows as the final score per video. **Time-CNN [42]** is a CNN for time series classification. This method reveals time series patterns through 1D convolutions on the temporal vector of each feature dimension. **DARE [40]** is a multimodal deception method. For our experiments we use a model with motion features only provided by authors.

6.2. High-stakes deceit

We perform a comparison with other methods on the high-stakes deceit settings using the Real-Life Trial dataset. Results are reported in the Table 1. Leave-one-person-out (LOPO) validation is used to solve the dataset's flaw: the imbalanced amount of videos per subject (Fig. 3) which causes one subject to appear in both training and test splits when using K-Fold or leave-one-out validation. Subjects who have either too few (1) or too many videos (20% of the remaining videos) are always kept in the training set. 15% to 20% of the remaining videos are randomly separated as the validation. We try to get as much balanced as possible training and validation splits in terms of classes. To have a balanced training set, we randomly downsampled the majority class in terms of quantity to have an equal number of instances from each class.

Morales *et al*. [27] mentioned 71.07% and 73.55% accuracy results for their visual model with DT and RF classifiers, respectively. However, they obtained these figures erroneously by applying *leave-one-out* validation which causes subject overlaps between the test and train dataset. In this experiment, the results of both Morales *et al*. and Perez-Rosas *et al*. are reported under LOPO settings instead.

The last row of Table 1 shows the performance of our proposed deception detection method. Our method performs on par with manual deceit methods that rely on hand-labeled features and achieves the best performance among automatic methods. Note that hand-labeled features are not possible in a real-life scenario. A significant improvement over other automatic facial feature extraction based methods shows that our method can extract more reliable facial features under challenging conditions since RLT dataset consists of varying illumination conditions and subjects are recorded under various viewing angles at various distances

to the camera.

6.3. Low-stakes deceit

We compare methods, which are designed for high-stakes settings, on the low-stakes deceit detection task using our newly collected LSD dataset. To our knowledge, we are the first to evaluate automatic deception detection methods on low-stakes deceit detection. Methods are evaluated separately on subsets with Dutch and English speakers. We applied X-Fold validation and made sure the same subject didn't occur simultaneously in training/validation/testing splits. Results are reported in the Table 2.

Automatic methods in general works as well as human evaluators (51.6% accuracy) on our benchmark, in spite of using visual-only features versus visual and audio for humans. In the case of our method, it's constrained to facial expression and pose related properties alone. This constraint prevents the model from biases toward subject identity and environment condition (an important property for deceit detection systems), however, simultaneously creates more challenges for deceptive behavior prediction. In addition, our dataset is collected under controlled settings (*e.g.* subjects are frontally facing the camera, subjects are sitting at a certain distance from the camera, faces are well lit). Such a controlled setting eases the problem of reliable facial feature extraction which explains why all automatic facial feature extraction based deceit detection methods achieve similar accuracy (54%) in our dataset.

Low-stakes deceit detection is a very challenging problem since people in low-stakes situations tend to behave less nervous, and hence showing less behavioral changes. In more than 20 human behavior studies on low-stakes deceit conducted by other researchers an average accuracy of 55% has been achieved by *professional experts* in comparison with high-stakes deceit studies with an average accuracy of 67% [28]. Thus, our deceit detection method performs with similar accuracy to that of professional experts in the low-stakes deceit detection task on our benchmark.

6.4. Influence of age

Since our LSD dataset provides age labels, we have clustered results into age classes to evaluate the correlation between age and deceit detection accuracy (Table 3). We separated samples into 3 categories: children, young adult, middle age, and above. We have observed higher accuracy on lie detection for children in comparison to adults in the English language split. This might be explained by children being more expressive with their expression. However, results require further research for a definitive conclusion.

6.5. Influence of gender

We investigate the effect of gender on RLT and our LSD datasets. RLT dataset has been manually annotated with gender labels. The results are summarized in Table 4. High precision and recall values of females may suggest that the feature extraction of males is more challenging and has high variation. However, this can also be related to the number of samples as we have female subjects almost as twice as males subjects in the RLT dataset. For the low-stakes settings, we have observed better accuracy on the female split for English speakers with less conclusive results for Dutch.

Age	Lang.	Acc.	Prec.	Rec.	# samples
< 18	EN	0.56	0.56	0.58	62
	NL	0.51	0.51	0.64	228
$\geq 18, < 45$	EN	0.53	0.52	0.66	106
	NL	0.53	0.52	0.70	134
≥ 45	EN	0.50	0.50	0.64	28
	NL	0.54	0.53	0.57	56

Table 3. Clustering low-stakes results by age.

Dataset	Gender	Acc.	Prec.	Rec.	# samples
RLT	Male	0.65	0.38	0.50	46
	Female	0.70	0.76	0.77	70
LS EN	Male	0.53	0.53	0.58	106
	Female	0.55	0.54	0.69	98
LS NL	Male	0.52	0.51	0.67	176
	Female	0.52	0.52	0.64	242

Table 4. Gender-specific deceit detection results on the RLT and LSD datasets.

7. Conclusion

We have presented a novel method for deception detection based on reliable facial expression and head pose related features. Those properties have been disentangled using a 2D-to-3D face reconstruction technique which simultaneously learns (a) face identity, environment parameters, and (b) facial expression and head pose using separate convolutional neural networks, and hence achieves their separation. Our pipeline models deceit detection as a Multiple Instance Learning problems conditioned on reconstruction features. It's real-time and (with an accuracy of 68%) improves the state-of-the-art as well as providing on par results with the use of manually coded facial attributes (71%) in the high-stakes deception detection on the challenging RLT dataset. We have collected a new low-stake deceit detection dataset. To our knowledge, we are the first to evaluate automatic visual-based high-stake deceit detection methods on low-stakes deceit detection tasks. In the low-stakes lies deception detection task it has achieved results on par with professional experts however there is still room for improvement. We hope that the newly collected dataset will allow further research to be done on low-stake deceit detection.

References

[1] Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 58–65. ACM, 2014.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[4] Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

[5] Henri Bouma, Gertjan Burghouts, Richard den Hollander, Sophie Van Der Zee, Jan Baan, Johan-Martijn ten Hove, Sjaak van Diepen, Paul van den Haak, and Jeroen van Rest. Measuring cues for stand-off deception detection based on full-body nonverbal features in body-worn cameras. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII*, volume 9995, page 99950N. International Society for Optics and Photonics, 2016.

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[7] Judee K Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and Alessandro Vinciarelli. *Social signal processing*. Cambridge University Press, 2017.

[8] Grigorios G. Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, 126(2-4):198–232, 2018.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[10] Nicole M Lawless DesJardins and Sara D Hodges. Reading between the lies: Empathic accuracy and deception detection. *Social Psychological and Personality Science*, 6(7):781–787, 2015.

[11] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R1-pca: Rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 281–288, New York, NY, USA, 2006. ACM.

[12] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

[13] Klaus Fiedler, Jeannette Schmid, and Teresa Stahl. What is the current truth about polygraph lie detection? *Basic and Applied Social Psychology*, 24(4):313–324, 2002.

[14] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.*, 32(6):158–1, 2013.

[15] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] Thomas Gerig, Andreas Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. Morphable face models - an open framework. *CoRR*, abs/1709.08398, 2017.

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[18] Maria Hartwig and Charles F Bond Jr. Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, 28(5):661–676, 2014.

[19] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[21] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. InverseFaceNet: Deep monocular inverse face rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[24] Tatsuro Koizumi and William Alfred Peter Smith. "look ma, no landmarks!" - unsupervised, model-based dense face alignment. In *16th European Conference on Computer Vision (ECCV 2020) Proceedings*, LNCS. Springer-Verlag, July 2020. This is an author-produced version of the published paper. Uploaded in accordance with the publisher's self-archiving policy. Further copying may not be permitted; contact the publisher for details. ; 16th European Conference on Computer Vision (ECCV 2020) ; Conference date: 23-08-2020 Through 28-08-2020.

[25] Mital Lakhani and Rachel Taylor. Beliefs about the cues to deception in high- and low-stake situations. *Psychology, Crime & Law*, 9(4):357–368, 2003.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[27] Michelle Renee Morales, Stefan Scherer, and Rivka Levitan. Openmm: An open-source multimodal feature extraction tool. In *Proc. Interspeech 2017*, pages 3354–3358, 2017.

[28] Maureen O'Sullivan, Mark G Frank, Carolyn M Hurley, and Jaspreet Tiwana. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530, 2009.

[29] Steven J Pentland, Nathan W Twyman, Judee K Burgoon, Jay F Nunamaker Jr, and Christopher BR Diller. A video-based screening system for automated risk assessment using nuanced facial features. *Journal of Management Information Systems*, 34(4):970–993, 2017.

[30] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 59–66, New York, NY, USA, 2015. ACM.

[31] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018.

[32] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[33] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[34] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015.

[35] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[36] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016.

[37] Aldert Vrij, Ronald P Fisher, and Hartmut Blank. A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21, 2017.

[38] Aldert Vrij, Maria Hartwig, and Pär Anders Granhag. Reading lies: Nonverbal communication and deception. *Annual Review of Psychology*, 70(1):295–317, 2019. PMID: 30609913.

[39] Pingping Wu, Hong Liu, Chao Xu, Yuan Gao, Zheyuan Li, and Xuewu Zhang. How do you smile? towards a comprehensive smile analysis system. *Neurocomputing*, 235:245–254, 2017.

[40] Zhe Wu, Bharat Singh, Larry S. Davis, and V. S. Subrahmanian. Deception detection in videos. *CoRR*, abs/1712.04415, 2017.

[41] Wen-Jing Yan and Yu-Hsin Chen. Measuring dynamic micro-expressions via feature extraction methods. *Journal of Computational Science*, 25:318–326, 2018.

[42] Bendong Zhao, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017.