Deception Detection by 2D-to-3D Face Reconstruction from Videos

Minh L. Ngô1,4 Sezer Karaoglu1,4
Burak Mandira2 Henri Bouma3
Selim Frat Yılmaz2 Hamdi Dibeklioglu2
Ward Heij1,3 Theo Gevers1,4

1 University of Amsterdam 2 Bilkent University 3 TNO 4 3DUniverse
{l.m.ngo, th.gevers}@uva.nl, {burak.mandira, firat.yilmaz}@ug.bilkent.edu.tr
{ward.heij, henri.bouma}@tno.nl, s.karaoglu@3duniverse.com, dibeklioglu@cs.bilkent.edu.tr

Abstract

Lies and deception are common phenomena 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 jointly extracts reliable low- and high-level facial features namely, 3D facial geometry, skin reflectance, expression, head pose, and scene illumination in a video sequence. Then these features are modeled using a Recurrent Neural Network to learn temporal characteristics of deceptive and honest behavior. We evaluate the proposed method on the Real-Life Trial (RLT) dataset that contains high-stake deceptive and honest videos recorded in courtrooms. Our results show that the proposed method (with an accuracy of 72.8%) improves the state of the art as well as outperforming the use of manually coded facial attributes (67.6%) in deception detection.

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 are able to correctly classify only 47% of lies and 61% of truthful statements [5]. 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 investigation. However, the ubiquitous polygraph, the most commonly known lie detection mechanism, has been shown to be unreliable [12].

Invasive approaches such as PET (positron emission tomography) and fMRI (function magnetic resonance imaging) based methods perform better but they are neither fully reliable nor practical in many situations where compactness or portability is required. Besides, the invasive nature of such mechanisms leaves them to be easily tricked by skilled deceivers [12]. 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 lies are barely evident in behaviour [17].

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 (sub-conscious) leakage of behavioral cues to deception [17]. These cues are often weak, very fast or subjective, making them hard to interpret by humans. Recent studies on automated deception detection [23] exploits different behavioral modalities such as facial actions/expressions, head pose/movement, gaze, hand gestures, and even vocal features in the analysis [23, 2]. In contrast, our work focuses solely on facial cues (including head pose/movement), yet providing a better accuracy.

High-level visual features used in the literature [23] such as facial action units, are prone to errors due to challenging environmental conditions (i.e. illumination, view point, occlusion etc.). Thus, such features can introduce significant amount of noise in the analysis. In this paper, to cope with such issues, we propose to exploit face reconstruction to obtain an effective low level representation for a more reliable deceit detection. Face reconstruction aims at decomposing a face image into its components such as 3D facial geometry, expression, skin reflectance, head pose, and illumination parameters, which are expected to carry important information for deceit detection. While the illumination parameters sound like unrelated to be used in this task, in combination with geometry it reveals subtle changes in expression-related skin deformations. Furthermore, prediction of these parameters, in our method, are constrained...
by an image formation model that relies on joint parametric modeling of facial cues, head pose, and illumination. Therefore, it minimizes the possible negative influences of varying environmental conditions. Once such components are extracted, they are fed to a Recurrent Neural Network to model temporal characteristics of deceptive and honest behavior in videos.

Although, a successful decomposition has been a backbone for many face-related computer vision tasks (e.g. face recognition, emotional expression recognition, head pose estimation, etc.), this work is the first one that exploits face reconstruction for deceit detection. Furthermore, we propose a fully unsupervised end-to-end deep architecture for face reconstruction (including 3D facial geometry, expression, reflectance, head pose and illumination) from videos. Our results show that the proposed novel method for deception detection improves the state of the art, as well as outperforming the use of manually annotated facial attributes (e.g. facial actions/expressions, gaze, and head movement) for this task.

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 [17]. One can divide observable cues in physiological cues, body language cues and facial cues. One of the problems about 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 [8], smiling and pupil dilation and facial rigidity [24]. However, the studies often find contradictory results [6]. In addition, performance is highly dependent on the data used for training and validation, with some datasets being noticeably easier than others [31]. Secondly, the circumstances under which the lies were elicited are influential: multiple studies indicate that deceptive cues increase in magnitude with increased cognitive load [30]. Hence, the final application and training data should have comparable cognitive load during data recording.

Micro-expressions pose another viable source of information [32], 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 [23].

One of the most recent methods on automated deceit detection is proposed by Morales et al. [23]. 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 [3] and later fused with verbal and acoustic features. Fusion occurs through concatenation of statistical functional vectors, after which random forests and decision trees are used for deception classification. Differently, [25] 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.

2.2. Monocular Face Reconstruction

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 infinite number of combinations can produce the same 2D image [4]. In general, we can categorize face reconstruction methods into two groups, namely, iterative [4, 27, 14, 28] and deep learning based [26]. Iterative approaches try to optimize parameters by minimizing the error between projected (reconstructed face) and original image in an iterative (analysis-by-synthesis) manner [4]. The energy functions are mostly non-convex. 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 therefore are not preferred in this paper.

Deep learning based methods to reconstruct face from a single monocular image typically uses either data augmentation techniques to regress prediction to be close to the ground truth [18, 15] or applies the similar analysis-by-synthesis approach to train the neural network using a physically plausible image formation model [26, 15]. 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 is quite important.

Some works have been proposed to overcome such issues by using RGB videos instead of single monocular images [27, 14, 28]. However, these works are based on iterative approach. Convolutional Neural Network (CNN) architectures are rarely 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 extracts identity related parameters (i.e. 3D facial geometry and reflectance) for a full video sequence whereas temporally dependent parameters (i.e. expression and illumination) for every single frame. The proposed method prohibits leakages of expression parameters to 3D facial geometry by temporal constraints which improves the preciseness of facial expression capture. Furthermore, using a Recurrent Neural Network (RNN), we temporally constrain the expression so that we preserve the consistency between expression through full video.

3. Methodology

3.1. Network architecture

Convolutional Neural Network is used to predict intrinsic inverse rendering parameters $P \in \mathbb{R}^{257}$ (code vector) from a set of RGB face images $\{I_i\} \in \mathbb{R}^{W \times H \times 3}$, from which a reconstructed image can be recovered:

$$P = \{\alpha, \beta, \delta, \gamma, \omega, t\}, \quad (1)$$

where $\alpha = \{\alpha_i\}, \beta = \{\beta_i\} \in \mathbb{R}^{80}$ and $\delta = \{\delta_i\} \in \mathbb{R}^{64}$ are parameters corresponded to 3D face geometry, reflectance and expression; $\gamma \in \mathbb{R}^{9 \times 3}$ represents scene illumination parameters; $\omega \in SO(3)$ and $t \in \mathbb{R}^3$ represent rotation and translation parameters.

Figure 1 shows an overview of our face reconstruction architecture. Our model consists of two AlexNet [21] backbones with shared weights, one (Identity CNN) to extract person identity 3D facial geometry and reflectance features related to $\alpha, \beta$ from a collection of images $I_1..I_M$ and another (Framewise CNN) to extract frame-dependent facial features $\delta, \gamma, \omega, t$ from particular frame $X$. For our purpose we are using all layers of AlexNet except the last FC8 layer. Those features are fused using recurrent units with 100 hidden parameters and fully connected layers without non-linearity to predict single set of identity parameters $\alpha, \beta$ defined by a set of cropped face $I_1..I_M$, expression parameters $\delta$ conditioned on the previous temporal state, illumination, rotation and translation parameters.

Identity CNN is followed by recurrent unit of 100 hidden parameters and fully connected layer without non-linearity to produce identity parameters $\alpha, \beta$. Features from a recurrent unit is concatenated with Framewise CNN. This representation is followed by another recurrent unit of 100 hidden parameters and fully connected layer to produce blend shape parameters, and just fully connected layer without recurrent unit for other parameters.

3.2. Physics-based image formation

3D facial geometry and reflectance. We parametrize 3D face geometry using a multi-linear PCA model [16] separately for neutral face and face expression (Eq. 2). 3D face geometry is represented as a point cloud $X = (x, y, z)^T \in \mathbb{R}^{N \times 3}$ in the Euclidean space.

$$X = A_{id} + \sum_{i=1}^{80} \alpha_i \sigma_{id,i} P_{id,i} + A_{exp} + \sum_{i=1}^{64} \delta_i \sigma_{exp,i} P_{exp,i} \quad (2)$$

where we denote $A_{id}, A_{exp} \in \mathbb{R}^{N \times 3}$ as an average neutral face and an average expression geometries, $P_{id,i}, P_{exp,i} \in \mathbb{R}^{N \times 3}$ as their principal components sorted by standard deviations $\sigma_{id,i}, \sigma_{exp,i} \in \mathbb{R}$ respectively. Face reflectance is modelled using a separate PCA model:

$$B = A_{tex} + \sum_{i=1}^{80} \beta_i \sigma_{tex,i} P_{tex,i} \quad (3)$$

where $A_{tex} \in \mathbb{R}^{N \times 3}$ is an average face reflectance, $P_{tex,i} \in \mathbb{R}^{N \times 3}$ are principal components sorted by standard deviations $\sigma_{tex,i} \in \mathbb{R}$.

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

Illumination model. Illumination changes are modelled using first 3 bands of spherical harmonics basis function $H_j$ assuming face is a Lambertian surface [29]. Intensity of the i-th vertex $c_i$ is defined as a product of vertex reflectance $b_i$ and a shading components.

$$c_i = b_i \sum_{j=1}^{3^2} \gamma_j H_j(R(\omega) n_i), \ i = 1..N, \quad (4)$$

where $n_i$ is a vertex normal of the i-th vertex. We define illumination parameters $\gamma_j$ separately for each $R, G, B$ channels and therefore have 27 parameters in total. Vertex normal is estimated based on 1-ring triangle neighbours. Triangle topology is known from the face morphable model.

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

$$\begin{bmatrix} x \\ y \\ z \\ d \end{bmatrix} = \begin{bmatrix} V \times [P] \times [R(\omega) t] \times [\Pi] \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \quad (5)$$

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

3.3. Training loss

We employ the energy minimization strategy of Tewari et al. [26]. In total our loss contains of 3 main components: landmark loss $E_{land}$, vertex-wise photometric loss $E_{vert}$ and regularization term $E_{reg}$ (Eq. 6).

$$\mathcal{L} = w_{land}E_{land} + w_{vert}E_{vert} + E_{reg}$$

(6)

**Landmark loss.** $L_2$ 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 regions.

$$E_{land} = \frac{1}{|\mathcal{F}|} \sum_{j \in \mathcal{F}} \| p_{k_j} - l_j \|^2$$

(7)

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

**Vertex-wise photometric loss.** 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 bilinear interpolation.

$$E_{vert} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \| c_i - X_{u_i, v_i} \|^2,$$

(8)

We filter out vertices which contributes to the photometric loss based on normal direction, $|\mathcal{V}|$ is an amount of those vertices.

**Statistical regularization.** We use Tikhonov regularization [29] to enforce predicted parameters to be in the plausible range.

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

(9)

3.4. Modeling Deceptive Behavior

Once the facial representation is obtained, we classify videos as deceptive or honest. We employ recurrent neural network (RNN) to capture temporal relations between facial representation vectors of frames $P \in \mathbb{R}^{257}$ for each video. We use the loss function

$$\mathcal{J} = \frac{1}{\sum_{j=0}} (y_{i,actual} - y_{i,pred})^2$$

(10)

where $y_{i,actual}$ is the label of the video $i$, and $y_{i,pred}$ is the predicted probability of video $i$ for class $j$. Deceptive and honest behaviors correspond to 1 and 0, respectively.

We employ single layer RNN of 128 units followed by 2 output neurons with sigmoid activation function. Each output neuron corresponds to a class, deception or truth. At evaluation stage, the class that corresponds to the output neuron with maximum probability is the final prediction.
4. Implementation details

We train our 2D-to-3D face reconstruction network for 200K iterations on 300VW [9] and CelebA datasets [22] using a batch size of 5 and Adam optimizer [20] with learning rate of $10^{-5}$. Loss weights are set to be $w_{vert} = 1.92$, $w_{land} = 0.0019$, $w_{\alpha} = 2.9 \times 10^{-5}$, $w_{\beta} = 4.93 \times 10^{-8}$, $w_{\delta} = 2.32 \times 10^{-5}$.

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 [19] for face detection and FAN [7] for landmark detection. In total we have collected 94K images from 300VW coming from 49 videos and 200K images from CelebA.

For each video sequence of 300VW we randomly select 3 crop faces in random order 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 Framewise-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.

AlexNet backbones are initialized using a pretrained model on ImageNet. We add additional offset to the 0-th band SH coefficient and z-translation to make sure 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, reflectance and expression. We take first 80 principal components for $\alpha$ and $\beta$ and 64 for $\delta$. We implement $SO(3)$ gradients in the compact form based on the work of Gallego et al. [13]. Our implementation is written in Tensorflow [1]. We ran our experiments on NVIDIA GTX 1080.

We train our deception modeling RNN for 10 epochs on Real-Life Trial (RLT) dataset using a batch size of 16 and Adam optimizer [20] with learning rate $10^{-3}$.

5. Dataset

In this study, we employ the Real-Life Trial dataset [25] 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. [25] enforce some visual constrains for videos such as the defendant or witness and his or her face should be clearly identified during most of the footage as well as some vocal enforcements which are not relevant within our context. We discard 40 of these videos from the dataset due to technical errors: 1) failure in facial landmark detection using [19] 2) some videos do not display the target subject and instead show something else such as courtroom while having the voice of target subject. Thereby, a subset of 81 videos (39 deceptive, 42 truthful) from this dataset is used which is constructed from 28 male and 53 female subjects.

6. Experiments

In this section, we explain all the experiments that are conducted in detail. We considered lie as positive and truth as negative throughout the experiments when calculating accuracy, precision and recall.

6.1. Comparison with monocular 2D-to-3D methods

Our pipeline constraints prediction of identity parameters $\alpha$, $\beta$ by making use of randomly selected multiple frames. In this experiment, we evaluate how sensitive the proposed face reconstruction is to the choice of frames for identity estimation. We compare our network accepting random single monocular image against to the work of Tewari et al. which does not perform any conditioning on identity. The evaluation is performed on the 300VW valida-
tation split. It contains 3 video subsets with different scene complexity. Set 1 contains 31 videos, set 2 contains 19 videos with difficult illumination conditions, set 3 contains 14 videos with occlusion, extreme pose and expressions. Results are reported in the Table 6.1. Results have shown that our method produces more consistent predictions for albedo and shape parameters in comparison to Tewari et al. This shows that the proposed method avoid leakages between expression and geometry. Consequently, proposed method predicts more precise expression and geometry.

Table 1. Identity parameters ($\alpha$, $\beta$) variation comparison on 300VW validation sets. Our network is accepting single monocular image during the testing time. Lower - better.

|                | Set 1 | Set 2 | Set 3 |
|----------------|-------|-------|-------|
| Tewari et al. [26] ($\alpha$ std) | 0.240 | 0.191 | 0.299 |
| Tewari et al. [26] ($\beta$ std)  | 0.299 | 0.159 | 0.174 |
| Ours ($\alpha$ std)              | 0.051 | 0.052 | 0.064 |
| Ours ($\beta$ std)               | 0.065 | 0.052 | 0.082 |

Table 1. Identity parameters ($\alpha$, $\beta$) variation comparison on 300VW validation sets. Our network is accepting single monocular image during the testing time. Lower - better.

6.2. Face Reconstruction Visual Results

We show visuals of our reconstruction pipeline in the Figure 3. Our method successfully recover intrinsic properties of a face such as shading, normals and color intensity preserving facial identity over video frames.

6.3. Gender Effect

In this experiment, we investigate the effect of gender to our results. The dataset does not provide gender annotation, therefore we annotate the dataset with gender information. Then dataset samples are grouped based on their gender to analyze results for each gender. The results are summarized in Table 2. High precision and recall values of females may suggest that feature extraction of females 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.

6.4. Comparison to Other Models

First, we start our comparisons with reproducing the baseline models. The model of Morales et al. [23] is tested with decision tree (DT) and random forest (RF) classifiers with default parameters as mentioned in their papers. Morales et al. use OpenFace [3] to extract facial features in default output (i.e. basics, gaze, pose, 2D and 3D facial landmark locations, rigid and non-rigid shape parameters, action units) and apply some statistical functionals (i.e. maximum, minimum, mean, median, standard deviation, variation, kurtosis, skewness, 25% percentile, 50% percentile, and 75% percentile) in order to create one feature vector per video.

The model of Perez-Rosas et al. [25], which is the basis for Morales et al. [23], is also implemented with decision tree (DT) and random forest (RF) classifiers with default parameters as mentioned in their papers. They use manually crafted features (i.e. smile, laughter, scowl, gaze, lips, openness and closeness of eyes and mouth, position of eyebrows like frowning and raising, head movements, hand trajectory and movements). Thus, the accuracy results of their work indeed show the performance of human annotators. Note that, since our system focuses only on facial features, we excluded hand-related features from their experiment setup to obtain comparable results.

Morales et al. [23] mention 71.07% and 73.55% accuracy results for their visual model with DT and RF classifiers, respectively. However, they obtain these figures erroneously by applying leave-one-out cross-validation which causes to subject overlaps between test and train dataset. Thereby, in our experiments for both [23] and [25], we applied leave-one-person-out (LOPO) cross-validation (where videos of a single subject are separated as test set and from the remaining videos five of them are randomly sampled as validation set and the remaining videos are taken as training set for each test fold) to reproduce their corrected accuracy results, given in Table 6.3. Note that, in order to have balanced training dataset at the end, we randomly downsampled majority class in terms of quantity to have equal number of instances from each class. In Table 3, each DT and RF result is obtained by taking the average of 20 iterations.

As our third baseline, we experiment with convolutional neural network for time series classification (Time-CNN) model [33], as shown in Table 3. This method reveals time series patterns through 1D convolutions on temporal vector of each individual feature dimension. This model emphasizes the strength of our proposed deception modeling RNN as it constructs a strong baseline that uses the same features (i.e. face reconstruction components (FRC) that our proposed CNN face reconstruction network extracts) with the proposed RNN.

Table 2. Gender specific deceit detection results.

| Gender | Accuracy | Precision | Recall | # samples |
|--------|----------|-----------|--------|-----------|
| Male   | 0.64     | 0.29      | 1.00   | 28        |
| Female | 0.77     | 0.87      | 0.77   | 53        |

The results are summarized in Table 6.3. The results show that the proposed RNN method improves the state of the art with 4%. In addition, we also improve (6%) [25], the method which uses manually annotated features. There might be leakage of behavioural cues to deception. This
leakage may not be an obvious behaviour which may not necessarily annotated by the human observer. This may indicate that the proposed method captures even subtle behaviours for deceit detection.

7. Conclusion

We have presented a novel method for deception detection based on reliable low- and high-level facial features obtained using 2D-to-3D face reconstruction technique. To be able to reconstruct faces (including 3D facial geometry, expression, reflectance, head pose and illumination) from videos, we propose a fully unsupervised end-to-end deep architecture. We show our method to produce consistent identity prediction in contrast to other deep learning methods which take only one monocular frame during the testing time. Our pipeline uses recurrent neural networks to learn temporal behaviour, works real time and shows state-of-the-art accuracy in the challenging Real-Life Trial (RLT) dataset. Our results show that the proposed method (with an accuracy of 72.8%) improves the state of the art as well as outperforming the use of manually coded facial attributes (67.6%) in deception detection.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo. Deception detection using a multimodal approach. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 58–65. ACM, 2014. 1
[3] T. Baltrušaitis, P. Robinson, and L.-P. 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.

[4] V. Blanz and T. 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.

[5] C. F. Bond Jr and B. M. DePaulo. Accuracy of deception judgments. Personality and social psychology Review, 10(3):214–234, 2006.

[6] H. Bouma, G. Burghouts, R. den Hollander, S. Van Der Zee, J. Baan, J.-M. ten Hove, S. van Diepen, P. van den Haak, and J. 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 Defense XII, volume 9995, page 99950N. International Society for Optics and Photonics, 2016.

[7] A. Bulat and G. 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.

[8] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli. Social signal processing. Cambridge University Press, 2017.

[9] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking “in-the-wild”. International Journal of Computer Vision, 126(2-4):198–232, 2018.

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

[11] C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: Rotational invariant 1-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] K. Fiedler, J. Schmid, and T. Stahl. What is the current truth about polygraph lie detection? Basic and Applied Social Psychology, 24(4):313–324, 2002.

[13] G. Gallego and A. J. Yezzi. A compact formula for the derivative of a 3-d rotation in exponential coordinates. CoRR, abs/1312.0788, 2013.

[14] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. ACM Trans. Graph., 32(6):158-1–1, 2013.

[15] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[16] T. Gerig, A. Forster, C. Blumier, B. Egger, M. Lüthi, S. Schömhorst, and T. Vetter. Morphable face models - an open framework. CoRR, abs/1709.08398, 2017.

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

[18] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. InverseFaceNet: Deep monocular inverse face rendering. In Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

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

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS’12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), Dec. 2015.

[23] M. R. Morales, S. Scherer, and R. Levitan. Openmm: An open-source multimodal feature extraction tool. In Proc. Interspeech 2017, pages 3354–3358, 2017.

[24] S. J. Pentland, N. W. Tywman, J. K. Burgoon, J. F. Nunemaker Jr, and C. B. Diller. A video-based screening system for automated risk assessment using nuanced facial features. Journal of Management Information Systems, 34(4):970–993, 2017.

[25] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. 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.

[26] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In The IEEE International Conference on Computer Vision (ICCV), 2017.

[27] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. ACM Trans. Graph., 34(6):183–1, 2015.

[28] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. 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.

[29] 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.

[30] A. Vrij, R. P. Fisher, and H. Blank. A cognitive approach to lie detection: A meta-analysis. Legal and Criminological Psychology, 22(1):1–21, 2017.

[31] P. Wu, H. Liu, C. Xu, Y. Gao, Z. Li, and X. Zhang. How do you smile? towards a comprehensive smile analysis system. Neurocomputing, 235:245–254, 2017.

[32] W.-J. Yan and Y.-H. Chen. Measuring dynamic micro-expressions via feature extraction methods. Journal of Computational Science, 25:318–326, 2018.
[33] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017.