Facial Behavior Analysis using 4D Curvature Statistics for Presentation Attack Detection

Martin Thümmel and Sven Sickert and Joachim Denzler
Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany

Abstract—The uniqueness, complexity, and diversity of facial shapes and expressions led to success of facial biometric systems. Regardless of the accuracy of current facial recognition methods, most of them are vulnerable against the presentation of sophisticated masks. In the highly monitored application scenario at airports and banks, fraudsters probably do not wear masks. However, a deception will become more probable due to the increase of unsupervised authentication using kiosks, eGates and mobile phones in self-service. To robustly detect elastic 3D masks, one of the ultimate goals is to automatically analyze the plausibility of the facial behavior based on a sequence of 3D face scans. Most importantly, such a method would also detect all less advanced presentation attacks using static 3D masks, bent photographs with eyeholes, and replay attacks using monitors. Our proposed method achieves this goal by comparing the temporal curvature change between presentation attacks and genuine faces. For evaluation purposes, we recorded a challenging database containing replay attacks, static and elastic 3D masks using a high-quality 3D sensor. Based on the proposed representation, we found a clear separation between the low facial expressiveness of presentation attacks and the plausible behavior of genuine faces.

I. INTRODUCTION

The digitalization of organizational processes to improve their efficiency in terms of time and resources is progressing extremely fast. The demand for robust, unsupervised authentication methods is getting equally stronger. However, current authentication methods are not able to detect advanced spoofing attacks or identity thefts. Organizational processes which should be robust against fraud are for example automated border control, the opening of a bank account, financial transfer, and mobile payments using self-service eGates, kiosks, and mobile phones. In the following, a presentation attack refers to all cases in which biometric copies are presented.

Biometric methods focus on the shape of the human body as physiological characteristics of e.g. the finger-, palm print, and the iris [31]. The promising field of behavioral biometrics contains the analysis of e.g. the voice [8], walking gait [7] and keystroke dynamics [22]. By definition, these methods are more robust against presentation attacks than microscopic methods that rely on static appearance. However, individual characteristics can be either replayed or imitated by another person [1]. Instead of developing an even more robust behavioral biometric method, we focus on the task of Presentation Attack Detection (PAD) as a mandatory security check for face authentication systems.

We analyze the plausibility of the individual facial trait of a person based on temporal sequences of 3D face scans which we call 4D face scans. At first, we apply an unsupervised preprocessing step for face extraction and pose normalization. Afterward, we account for the huge amount of redundant information in high-resolution 4D face scans by analyzing the plausibility of curvature changes at subsampled radial stripes. Both handcrafted steps are appropriate in this case since it is not possible to create a large and unbiased database with all different presentation attacks to perform classification using e.g. deep learning.

II. RELATED WORK

A. Presentation Attack Detection

Similarly to the production of counterfeit money, imposters of an authentication system will always try to create better biometric copies. Most current 2D face authentication methods can be attacked by presenting photographs and 3D masks of another person’s face [3]. Therefore, several methods like [17] exist for distinguishing an image of a face photograph from an image of a genuine face. Since these methods are not able to detect attacks using high-quality photographs, challenge-response authentication methods were developed. For authentication, a user is asked to read words aloud [19], move its head [10], blink its eyes [28], or show facial expressions. However, these methods can be attacked by presenting a video on a monitor where the face behaves the same way as so-called replay attacks. Again, a solution to detect these attacks is to analyze if the 3D facial structure is plausible based on 3D landmark locations [30] or the mean curvature [20]. Such methods are robust against the presentation of bent photographs, but can still be attacked using 3D masks [26]. Even the 3D facial recognition system Face ID [25] analyzes only the static facial appearance in 3D to planar presentation attacks. To detect even elastic 3D masks, we have developed a method that combines both, the temporal analysis of 3D face scans and a challenge-response protocol.

Recent deep learning methods [23] can detect many kinds of 3D masks from 2D color videos of subjects with a neutral facial expression. They do not take the depth information into account and focus on subconscious facial movements like blinking. Hence, they are still vulnerable to replay attacks and unseen or partial 3D masks. Deep learning methods for PAD based on sequences of 2.5D depth images [14] are also able to detect 3D masks. Even though they achieve impressive results, their training databases are biased due to a
few different masks and genuine faces. Even if a challenging training database containing all state-of-the-art mask types would be available, a novel mask type or adversarial example [27] could still be used to attack the system. Thus, these methods are also not robust against unseen presentation attacks. Furthermore, these methods are not invariant under Euclidean transformations as they do not directly extract features from the 3D facial surface. In contrast, our method requires no costly training on such biased databases and only obtains the lower limit of the plausible amount of facial expression changes from recordings of genuine faces.

Liveness detection methods analyze the differences between genuine facial skin and masks or partial face modifications. One example of such a method is the estimation of the heart pulse rate from a regular camera [21]. However, emitting a green flashing light onto a mask with a plausible frequency can fool these methods since the light green channel varies the most with the pulse [9]. Furthermore, these methods can be attacked by presenting thin/partial masks and require almost static facial expression, pose, and illumination.

Alternatively, presentation attacks could be detected as all cases where the resulting shape and expression parameters of a fitted 3D morphable model (3DMM) [4] exceed their expected range. However, 3DMMs cannot adapt to 3D scans of photographs/monitors and would result in underfitting. Thus, the shape and expression parameters would still remain in the plausible range of genuine faces because of their inherent, statistical bounds of all faces during training.

B. Curvature Analysis of Radial Stripes

As emphasized by Katina et al. [15], 3D anatomical curves provide a much richer characterization of faces than landmarks, which are just individual points along anatomical curves. The method from Vittert et al. [29] localizes a complete set of anatomical curves along ridges and valleys of the facial shape. The overall curvature along a curve is iteratively maximized over points with a positive or negative shape index, respectively. A complete facial model is built from these anatomical curves, subsampled intermediate curves, and manually annotated landmarks. However, it was applied for statistics of the neutral face appearance. In the case of facial expressions, the shape index and 1D curvature can change drastically and would require many heuristics to still obtain robust anatomical curves.

Instead of relying on anatomical curves, the representation form Berretti et al. [2] approximates the facial shape with a set of geodesics. A set of radial curves is created by intersecting the facial surface with planes through the nose tip. Starting with the anatomical mid-sagittal plane, it is repeatedly rotated around the roll axis by a fraction of the full angle. This representation has proven to result in superior performances in 4D facial expression recognition [33], face recognition [12], and even body part analysis in general [5]. However, the method from Zhen et al. [33] would have a too high computational effort, if applied to facial behavior analysis on high-quality 4D scans due to the costly alignment procedure. All consecutive face scans are roughly aligned using the iterative closest point algorithm, following a fine-alignment of all consecutive radial curves using dynamic programming.

For feature extraction, all mentioned methods rely on the first feature extraction, all mentioned methods rely on the first D surface curvature [24] which takes the neighborhood on a 2D surface into account and is more robust against overfitting through noise, outliers, and holes. For registration purposes, the mentioned methods fit 1D curves and enforce a homogeneous geodesic distance by equidistant sampling along the arc-length. However, it is assumed that the complexity of the underlying surface must be similar for all surfaces and all expected shapes as the degrees of freedom of the curves are fixed to a certain number. Due to the huge shape variety of faces, 3D masks, and planar presentation attacks, this is not given in case of presentation attacks. Curves along the nasal ridge and through the cheeks have a lower curvature than curves through the eyes and mouth. In the extreme case of presentation attacks using photographs and monitors, all curvature values should be small. Depending on whether the degrees of freedom are adjusted to more or less complex surfaces, it results in over- or underfitting in the other cases, respectively.

III. METHODOLOGY

The calculation of our representation is illustrated in the pipeline in Fig. 1. Given two synchronized sequences of color and depth images, two preprocessing steps are performed. First, the anatomical landmarks are localized in the color images, transformed to the depth images and 3D reconstructed (Section III-A.1). Second, the face is extracted and the pose is normalized for each frame, to obtain a local coordinate system which is centered at the nose tip (Section III-A.2).

Afterward, we are proposing a representation of 4D face scans based on the following two steps: Equidistant surface curvatures are extracted at equiangular radial stripes to subsample the point cloud (Section III-B.1). The curvatures of consecutive frames are correlated over time to locally measure the curvature change for each radial stripe (Section III-B.2). The graph in Fig. 1 shows the maximum cross-correlation for each radial stripe over time. For example, the temporal changes of the two inner peaks relate to the eye movements and the two outer peaks to the mouth movements. We found that the standard deviation of the temporal changes in the mouth region allows for PAD (Section IV-A).

A. Preprocessing Steps

1) 3D Landmark Localization: The starting point for many methods which analyze human faces is the extraction of anthropometrical landmarks around the eyebrows, eyes, mouth, and nose (see Fig. 1, left). Since 2D facial landmark localization methods usually achieve superior performances than 3D methods due to larger training databases, we use the method from Kazemi et al. [16]. The even higher robustness of more recent deep learning methods [6] is not required as
we are focusing on the cooperative self-service scenario with an almost frontal pose and little occlusion.

For stereoscopic and multi-view camera systems, the 2D landmarks can be 3D reconstructed using the registered depth image. However, in case of denser and more accurate 3D scans of a structured-light 3D sensor, the depth and color images are not registered. In this case, texture coordinates contain the pixel-wise mappings from the depth image to the color image. We store the inverse mapping from color to depth pixels during the calculation of the texture coordinates and call them depth coordinates. Finally, the closest non-zero depth coordinate for a given color coordinate points to its corresponding depth. Since the texture coordinates are always calculated, the computational overhead of this approach is negligible. Fig. 1 shows the corresponding landmark locations in the color and depth image.

Alternatively, 3D landmarks could be 3D reconstructed using lossy RGB-D images in two ways. First, a registered depth image can be created by extracting the color values for each pixel in the depth image using texture coordinates. Second, a depth value can be assigned to each color pixel after 3D reconstructing and projecting all depth pixels to the color camera. As the resolution of the depth images is usually smaller than of the color images, the first approach would result in low-resolution color images and the second approach would result in depth images with holes. Furthermore, our approach avoids the high computational demand of both common alternatives.

2) Face Extraction and Pose Normalization: In the application scenario of a static 3D sensor in an eGate or kiosk system, the field of view must be large enough to capture faces of people with different sizes and positions. As such 3D face scans mostly contain background, the preprocessing first step centers the point cloud at the nose tip landmark and extracts all points inside a sphere with \( r = 0.1 \text{ m} \).

The second preprocessing step normalizes the head pose to an upright and frontal view. For head pose estimation, we adopt the approach from Derkach et al. [11] based on facial landmarks (see Fig. 2). Finally, the corresponding transposed rotation matrix is applied for pose normalization. This allows us to also achieve robust results in case of deviations of the head pose from the frontal view.

B. Spatiotemporal Curvature Analysis

The most important subtask of a method for 4D facial behavior analysis is to deal with the extremely high spatiotemporal redundancy of the 4D face scans. The temporal difference between consecutive face scans and the spatial difference between points and its immediate neighbors are very small. Thus, a feature representation should only extract facial expression changes, which constitute the overall facial behavior. Our new representation achieves this goal by calculating the curvature of point subsets in Section III-B.2 and a correlation-based approach in Section III-B.2.

1) Curvature Analysis of Radial Stripes: After applying the preprocessing steps, we account for the common issue of holes and peaks in 3D scans by calculating the mean depth image between three consecutive 3D scans. To extract radial stripes in the next step, a reference coordinate system given by the three unit vectors \( \mathbf{n}_{1..3} \) in x-, y-, and z-direction is
Fig. 3: After normalizing the pose to an upright and frontal view, we extract a radial stripe through the face (left). We then rotate the plane repeatedly by a fraction of the full angle until we obtained \( N \) radial stripes (right, after 4 rotations).

Fig. 4: Due to overfitting of radial curves, the 1D curvature becomes large (middle) in case of noisy, planar 3D scans of monitors (left) and photographs. Instead, the 3D surface curvature (right) is more robust against noise and is almost zero everywhere in this case.

centered at the nose tip landmark. First, we intersect the face with the yz-plane as shown in Fig. 3 (exemplary, w/o cropping). A radial stripe is then extracted by taking all points for which the projection onto the x-axis \( p_i^T n_1 \) is lower than a certain threshold \( \delta \) and the projection onto the y-axis \( p_i^T n_2 \) is positive as

\[
\left\{ \begin{array}{l}
p_i = \left( \frac{x_i}{y_i} \right) : |p_i^T n_1| \leq \delta; p_i^T n_2 > 0 \\
\end{array} \right. \tag{1}
\]

among all points \( p_i \) of a single 3D face scan. Afterward, the x- and y-axis are rotated around the z-axis by \( \alpha = \frac{360^\circ}{N} \) and the process is repeated until \( N \) radial stripes are extracted.

To reduce the feature dimension, we calculate the curvature for each sampled point based on its local neighborhood in the original point cloud. The usage of the curvature is encouraged by its invariance under 3D Euclidean transformations. The curvature can be calculated from the 1D parametric arc-length representation of each curve as

\[
\kappa(s) = \frac{y''}{(1 + y'^2)^{3/2}}. \tag{2}
\]

However, the resulting 1D curvatures in case of a 3D scan of a flat surface in Fig. 4 are in the range of the curvatures of genuine face scans due to overfitting to the noisy point cloud. As a consequence, we are approximating the 3D surface curvature [24] as the surface variation from the statistics of the neighborhood for each point as

\[
\tilde{\kappa}(i) = \left| \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} \right|, \tag{3}
\]

where \( \lambda_0 < \lambda_1 < \lambda_2 \) and \( \lambda_i \) are the eigenvalues of a principal component analysis applied to a local point subset. In Fig. 5, it is shown that this approach is more robust against overfitting as a larger neighborhood on a 2D surface is taken into account. This approximation also avoids the major impact of the unbounded, unstable second derivation in (2). The surface variation is bounded to \( \frac{1}{3} \) for isotropically distributed points.

To compare consecutive radial stripes, we subsample \( M \) points along each radial stripe. We obtain an equidistant spacing between each point by sampling along the projection to the y-axis \( p_i^T n_1 \). An important advantage of the resulting representation is, that the degree of linear and angular subsampling can be varied depending on the desired accuracy and runtime.

2) Measuring the Facial Expression Change: The resulting curvatures of genuine faces would still be indistinguishable from well-shaped copies like the 3D masks from REAL-f Co. Therefore, we measure the temporal curvature change between consecutive face scans and perform PAD on the resulting time series representation.

After extracting the curvature of \( M \) points along \( N \) radial stripes, the point-wise product of the curvature values between consecutive radial stripes is calculated. Since the detected position of the nose tip varies slightly, it is necessary to align the radial stripes to each other. This is done by taking the corresponding shift at the point of maximum cross-correlation

\[
R_t = \max_j \left( \sum_{i=-M}^{M} \tilde{\kappa}_t(i) \cdot \tilde{\kappa}_{t-1}(i+j) \right) \tag{4}
\]

between consecutive radial stripes at time steps \( t \) and \( t - 1 \). The maximum cross-correlation measures how similar the consecutive curvature values are. The resulting multivariate
Fig. 6: Our new database contains 4D scans of presentation attacks using monitors (top middle/right), static 3D mask (bottom left) and elastic 3D masks (bottom right). The implemented challenge-response protocol results in the first database which allows for detecting presentation attacks using 3D behavioral analysis.

time series containing the values of $R_t$ for each radial stripe measures the individual change of the facial expression.

To compare the amount of change between presentation attacks and genuine faces, we calculate the standard deviation of each time series. The standard deviation measures the overall facial expression intensity and is comparable between all recordings if the number of facial expressions is similar. For PAD, all subjects were asked to answer the same number of questions to make the number of induced facial expressions and visemes comparable.

IV. EXPERIMENTS

A. 4D Presentation Attack Detection

Many published databases for PAD contain at most only 2D presentation attacks using bent photographs with holes [32]. The recently published 3D-MAD [13], CS-MAD [3] and WMCA [14] databases also contain static, elastic and partial 3D masks as presentation attacks. However, they were recorded using a low-quality 3D sensor and capture only the neutral face appearance without any facial expression. Since our method analyzes facial behavior, it is not possible to apply it to these databases.

Therefore, we collected an own database containing 48 4D face scans of 16 subjects and 9 presentation attacks using replay attacks, static and elastic 3D masks (see Fig. 6). Furthermore, some subjects have beards, wear classes and make-up to make it more challenging as these cases can only be reconstructed very inaccurately using active 3D sensors. We used a structured-light 3D sensor with a high accuracy of $\sigma = 100 \mu m$ and a depth resolution of 1 MP at 30 Hz over a duration of 18 s (540 frames). For each face scan, the computation takes only 162 ms on a Core-i7 CPU and the number of extracted points is around 100000. We will publish the source code and our new database.

In general, challenge-response protocols are used in case of recent behavioral biometric or PAD methods to become robust against static presentation attacks. We adopted the common practice of asking familiar small-talk questions at a bank or concerning the entry regulations in cross-border traffic. We instructed the user to answer by speaking, to induct visemes and facial expressions. In our case, the captured face scans serve as the response and are analyzed for irregularities in facial appearance and expression.

After calculating $R_t$ for all radial stripes and time series using (4), we found that the overall facial expression change of genuine faces is much larger than of presentation attacks (see Fig. 7). Fig. 8 shows the distribution of $\sigma(R_t)$ among all radial stripes and recordings. The standard deviation of the first and last few radial stripes through the mouth differ between genuine faces and presentation attacks by a large margin. For the eye regions, the standard deviation is also different, but not well-separated. Active 3D sensors are not able to accurately measure the eye region which is either reflecting or absorbing the projected stripe pattern. As shown in Fig. 9, a single radial stripe through the mouth region

1In general, visemes contain all mouth appearances during speech.
Fig. 9: We used the radial stripe which comprises the maximum overall curvature around the mouth region for PAD. A threshold at the lower limit of the plausible amount of facial expression change allows finally to separate genuine faces from presentation attacks.

Fig. 10: As a baseline, we calculated the standard deviation of the mouth openings in y- and z-direction over time as a simple measurement of the facial expression intensity. Even though the standard deviation is also larger in case of genuine faces, both classes are not well separated.

is finally sufficient to perfectly classify between genuine faces and presentation attacks. A threshold of 0.4 also leaves enough space for even more elastic 3D masks.

The key idea behind our approach is that presentation attacks show less facial expressions compared to genuine faces concerning the standard deviation of our representation. Hence, it would also be possible to track the position of any patch of the facial surface, extract local features and measure their standard deviation. Since it is not trivial to select a suited surface patch and to track its position in case of facial expressions and pose changes, we implemented a similar method based on the mouth openings as a baseline. After pose normalization, the Euclidean y- and z-distances between the labrale superior (ls) and labrale inferior (li) 3D landmarks are calculated as the mouth openings. However, the resulting standard deviations of these time series in Fig. 10 are too similar between genuine faces and elastic 3D masks.

For comparison purposes, we implemented the simple, yet powerful method from Lagorio et al. [20]. They found that the mean 3D surface curvature of genuine 3D face scans is larger than the mean curvature of (bent) photographs. To improve the robustness of this method, we also applied our preprocessing steps from Section III-A.2. Fig. 11 shows that this also holds in case of the four rightmost presentation attacks using monitors. However, since a 3D scan of a reflecting monitor is noisy, the differences are much smaller than their stated difference of an order of magnitude between photographs and genuine faces. As expected, their method cannot detect the presentation attacks using 3D masks since their mean curvature is in the same range as for genuine faces.

B. Paresis Treatment Analysis

In many cases of treatment analyses of patients, the individual progress matters and is too complex to be deduced from the average or a similar patient. As an outlook, we also applied our representation to individual treatment analysis of a patient with facial paresis. We recorded a patient every month over one year using the same 3D sensor as before. At the beginning of the study, the facial nerve of the right facial half was cut, resulting in a complete paralysis. During the treatment, the nerve grew back together (reinnervation), showed first signs of its restored functionality on 09/04/2018 and allowed for substantial, voluntary muscle movements at the end of the treatment (see Fig. 12).

For each recording, we calculated the mean of the cross-correlation $\bar{R}_t$ over the snarl facial exercise. Fig. 13 shows that $\bar{R}_t$ improved continuously over time during the reinnervation. The fluctuations on the healthy half of the face are caused by the varying motivation of the patient. Due to mass movements, i.e. stretching of the paretic muscles to the healthy half of the face, $\bar{R}_t$ is also high for the paretic half of the mouth. Since the eyes cannot be accurately reconstructed with 3D sensors in general, fluctuations of $\bar{R}_t$ occur in both eye regions.
As a single radial stripe through the mouth was sufficient to distinguish between 2D/3D presentation attacks and genuine faces. For future work, it remains a difficult task to distinguish elastic 3D masks from genuine faces which show minimal facial expressions while speaking.

To show the potential of our representation of 3D face scans for other research topics, we applied it to individual treatment analysis of patients with facial paresis. In this case, multiple radial stripes of our representation were required to highlight and localize the individual improvement in facial symmetry.

V. CONCLUSIONS

Even though current face authentication methods achieve impressive accuracies, most methods can be fooled by presenting a facial photograph of someone else instead of their own face. The ultimate goal for detecting also advances presentation attacks like bent photographs, replay attacks on monitors and elastic 3D masks, is to analyze if the behavior and the 3D shape of a face scan are plausible. To the best of our knowledge, we developed the first method which is able to robustly detect all of these presentation attacks directly based on 4D face scans. We subsampled the 3D surface curvature at equiangular radial stripes and calculated the standard deviation of the cross-correlation between consecutive stripes over time. Our proposed representation also allows for varying the degree of subsampling depending on the desired accuracy and runtime.

Many published databases for PAD contain at most only 2D presentation attacks using bent photographs with holes or the neutral face appearance in case of 3D presentation attacks. Since our method focuses on facial behavior, we collected a challenging database containing three different types of sophisticated masks and monitor replay-attacks. To induct facial expressions, we implemented a challenge-response protocol and asked the user to answer familiar questions in cross-border traffic by speaking. Our evaluation results for PAD showed the potential of our representation as a single radial stripe through the mouth was sufficient to perfectly distinguish between 2D/3D presentation attacks and genuine faces.

REFERENCES


