Face anti-spoofing based on projective invariants

Alexander Naitsat and Yehoshua Y. Zeevi
Viterbi Faculty of Electrical Engineering
Technion, Israel Institute of Technology
Haifa, Israel

***Abstract*—The most common security authentication systems rely on automatic face recognition, which is particularly vulnerable to various spoofing attacks. Often these attacks include attempts to deceive a system by using a photo or video recording of a legitimate user. Recent approaches to this problem are based on pure machine learning techniques that require large training datasets and generalize or scale, poorly.**

By contrast, we present a geometric approach for detecting spoofing attacks in face recognition based authentication systems. By locating planar regions around facial landmarks, our method distinguishes between genuine user recordings and recordings of spoofed images such as printed photos and video replays.

The proposed algorithm is based on projective invariant relationships that are independent of the camera parameters and lighting conditions. Unlike previous geometric approaches, the input to our system is a stream of two RGB cameras. Comparing with methods implemented by a single RGB camera, our approach is significantly more accurate and is completely automatic, since we do not require head movements and other user interactions. While, on the other hand, our method does not employ expensive devices, such as depth or thermal cameras, and it operates both in indoor and outdoor settings.

I. Introduction

Projective invariants are certain geometric properties preserved under perspective and orthographic projections. In the context of face anti-spoofing application, we are interested in projective invariance of certain sets of facial landmark points. In particular, we employed the so called *five point cross-ratio*, referred for short as to *cross-ratio*. This quantity is the projective invariant of five coplanar points.

Assume that $p_1, p_2, \cdots, p_5$ are five points located on a plane Φ in $\mathbb{R}^3$, and let $p'_1, p'_2, \cdots, p'_5$, be the corresponding projection of these points using the pin-camera model, as illustrated in Fig. 1. Then, the five point cross-ratio quantity is defined up to a permutation of point indices as the ratio of triangle areas

$$\gamma(p_1, p_2, p_3, p_4, p_5) = \frac{A'_{514} A'_{523}}{A'_{513} A'_{524}}, \tag{1}$$

where A'_{ijk} denotes the signed area of the projected triangle $\triangle(p'_i, p'_j, p'_k)$. According to the theory of projective geometry, $\gamma(p_1, p_2, p_3, p_4, p_5)$ is independent of camera orientation as long as the camera direction ray is not contained in the objective plane Φ. Using state-of-the-art methods for detecting facial landmarks in images, we can measure γ with sufficiently high precision.

Assume γ_1 and γ_2 are the cross-ratio measurements obtained from two cameras for a set of non-coplanar facial landmarks $\boldsymbol{p} = (p_1, ..., p_5)$, as illustrated in Fig. 2, we expect a significant deviation in γ_1 and γ_2 values for genuine recordings of a three-dimensional face, while spoofing attacks, such as demonstrated by video replays or printed images, will result in nearly identical cross-ratio values, since in these cases landmarks are projected from a flat surface. Consider the cross-ratio difference

$$\Delta_\gamma(\boldsymbol{p}) = |\gamma_1(\boldsymbol{p}) - \gamma_2(\boldsymbol{p})|, \tag{2}$$

and assume that an anti-spoofing algorithm detects point coplanarity if $\Delta_\gamma(\boldsymbol{p})$ is bellow some constant threshold value ε. According to our simulations on synthetic data, this simple approach works well if 2D landmarks are detected with high precision. In such cases ε can be set very close to zero. However, processing of real data in varying lighting conditions and camera settings inevitably introduces substantial noise in detected landmark positions and, thus, $\Delta_\gamma(\boldsymbol{p})$ measurements may differ significantly along the same video recording. Consequently, analysis of real data requires a per-frame calibration of the threshold ε, based on geometric considerations and scene configurations.

In view of the above observations, we propose a heuristic approach based on comparison of the average $\Delta_\gamma(\boldsymbol{p})$ values with a predicted model of facial landmarks.

II. Related work

Invariant relationships of the projective and differential geometry were extensively employed during the past decade in computer vision applications for object analysis and recognition. However, early applications, such as [9], [4] and [3], led to poor results, since these methods employed low resolution sensors and their computations were based on old generation of vision algorithms for feature point detection.

Five and four point cross-ratios were employed in number of face recognition and authentication applications [5] and [3]. The work of [5] suggested to measure $\gamma(\boldsymbol{p}_i)$ for landmarks obtained in n frames, where changes of pose where detected. If $\mathrm{Var}\left(\gamma(\boldsymbol{p}_i)\right) < \epsilon$, then the system identifies point planarity and spoofing attack is detected. According to our tests, the proposed method is unreliable in practice due to the following drawbacks and non-realistic assumptions:

1) Often there are no detectable relative movements between the user and the camera, and thus the system expects intentional user movements, including number of head rotations.

978-1-5386-6378-3/18/$31.00 ©2018 IEEE

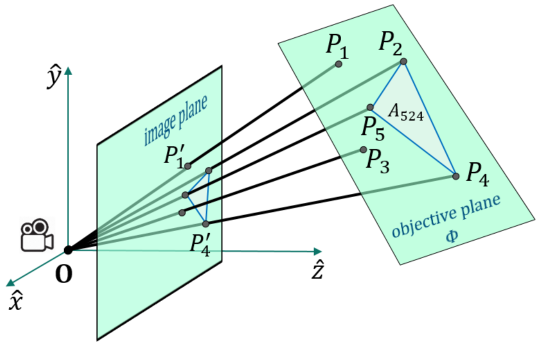

Fig. 1. Projection of five coplanar points from objective plane Φ into the image plane via the camera placed at the origin.

2) Facial landmarks set is not rigid, since mutual point distances vary from one facial expression to another (e.g., distances between eye and mouth corners).
3) This method can protect only from static images (print attacks), but not from spoofing videos, since in video clips 2D landmarks can move non-rigidly on the screen.

Algorithm 1 Algorithm for face anti-spoofing

```
Input:
  • Image sequence $I_{1j}, ..., I_{Nj}$ for $j = 1, 2$, where $I_{ij}$ is the $i^{th}$ frame image of the $j^{th}$ camera.
  • Parameters $W = (w_1, ..., w_5)$, where $w_i$ is the predicted coordinates of the $i^{\text{th}}$ landmark point in NHCS.
  • Internal model parameters $\delta_1 < \delta_2$.

Initialize: $\Sigma_P \mapsfrom 0$, $\Sigma_Q \mapsfrom 0$.

for $i = 1, ..., N$ do
    for $j = 1, 2$ do
        $p_1, p_2, ..., p_5 \mapsfrom$ projection of non-coplanar facial landmarks detected in $I_{ij}$.
        $B \mapsfrom$ estimates of the head bounding box in 3D.
        $q_1, q_2, ..., q_5 \mapsfrom$ predicted landmark positions in 2D obtained by $W$ and $B$.
        $\gamma_{ij}(p) \mapsfrom \gamma(p_1, p_2, p_3, p_4, p_5)$.
        $\gamma_{ij}(q) \mapsfrom \gamma(q_1, q_2, q_3, q_4, q_5)$.
    end for
    $\Sigma_P \mapsfrom \Sigma_P + |\gamma_{i1}(p) - \gamma_{i2}(p)|$.
    $\Sigma_Q \mapsfrom \Sigma_Q + |\gamma_{i1}(q) - \gamma_{i2}(q)|$.
end for

if $\Sigma_P < \delta_1 \Sigma_Q$ , then
    Classify as "spoofed".
else if $\Sigma_P > \delta_2 \Sigma_Q$ , then
    Classify as "genuine".
else
    Classify as "undecided".
end if
```

III. The method

Revisiting the previous methods for face authentication and anti-spoofing, we suggest the following approaches for the efficiency and performance improvement:

1) Our system employs two web-cameras to attain a stereo recording. This approach enables to simultaneously capture the scene from two different viewpoints and releases users from mandatory interactions.
2) Our algorithm is based on a new generation of facial landmark detectors.
3) By continuously tracking of the head pose, we predict 3D position of chosen landmark points and project them using camera parameters into pixels. Then, cross-ratios measurements of detected and predicted points are compared to distinguish between real human face and spoofed images.

Since we do not employ depth cameras, our system can measure directly only 2D positions of points projected form 3D into camera image planes. Nevertheless, as presented in Algorithm 1, our technique integrates both the direct measurements of two dimensional data and our estimates of related 3D features.

The first step of our method is the detection of two ordered sets of facial landmark points in $\mathbb{R}^2$, denoted by $\boldsymbol{p'}$ and $\boldsymbol{w'}$, respectively. The 5-tuple $\boldsymbol{p'}$, used for the cross-ratio computation, is the projection of chosen five non-coplanar points $\boldsymbol{p}$; while $\boldsymbol{w'}$ is a larger set of rigid landmarks employed for head pose estimation.

The head pose is represented by the transformation $(R, \boldsymbol{t})$ in the Camera Coordinate System (CCS), where R is the rotation matrix and $\boldsymbol{t}$ is the translation vector. In particular, $\boldsymbol{t}$ and R define the origin and the coordinate axes of the Head Coordinate System (HCS), respectively. Furthermore, camera parameters and mutual distances between points of $\boldsymbol{s'}$ are used for approximating dimensions $\boldsymbol{h} = (h_x, h_y, h_z)$ of the head bounding box in the HCS. Together, the triplet $(R, \boldsymbol{t}, \boldsymbol{h})$ forms the HCS-to-CCS transformation T. As depicted in Fig 3, this leads to the definition of the Normalized Head Coordinate System (NHCS), where positions of the right-top-forward and left-bottom-back corners of the head bounding box are set to $(1, 1, 1)$ and $(-1, -1, -1)$, respectively.

The input parameters of Algorithm 1 include mean positions $W = (w_1, .., w_5)$ in NHCS of the landmarks $\boldsymbol{p} = (p_1, ..., p_5)$. The values of W are evaluated only once in the offline process of minimizing distances between camera projections $T(w_1), .., T(w_5)$ and landmarks $p_1', .., p_5'$ detected over a large set of recorded images. During the online stage, $(w_1, .., w_5)$ are transformed to CCS and then projected using camera parameters to predict 2D position $q = (q_1, ..., q_5)$ of the landmarks.

Finally, cross-ratio differences $\Delta_\gamma(p)$ and $\Delta_\gamma(q)$ are compared for evaluating image authenticity. By employing additional parameters δ_1 and δ_2, Algorithm 1 classifies image sequences into one of the following categories: *spoofed*, *gen-*

uine and *undecided* images. The flowchart of Algorithm 1 and related techniques is presented in Fig. 5.

By testing various configurations of points $\boldsymbol{p'}$ chosen to compute cross-ratios, the best results are considered to be achieved for the following landmark selection: middle nose, eye and moth corners (see Fig. 2 and 3).

Note that in order to complete the authentication process the system should identify person form the list of authorized users using a face recognition block, as depicted in Fig 5.

IV. Implementation

We have written our algorithm in C++ using OpenCV, Dlib and OpenFace [2] facial analysis toolkit. The code is based on HOG SVM face detector and CLNF model [1] for facial landmark localization. Our algorithm can be run in single and multi-face modes and it was tested on number of RGB cameras with 30 FPS recording rates.

During the testing, we sample each 8th frame of the first and second camera into blocks of eight images passed to Algorithm 1 as the input parameters $I_{1j}, ...I_{8j}$.

The offline stage, including data analysis and evaluation of optimal algorithm parameters, was implemented in MATLAB. For testing synthetic data, we employed Blender modeling software along with freely available 3D face scans.

V. Results

We first experiment with synthetic face data, where $\boldsymbol{p'}$ are computed directly from 3D via the camera projection matrix. In this scenario, we achieve almost 100% accuracy by taking $\varepsilon \approx 10^{-6}$ as the threshold that divides $\Delta_\gamma(\boldsymbol{p'})$ values into planar and three-dimensional face categories. Example of two synthetic models that represent a genuine face and a spoofed image are shown in Fig. 2.

Next, we test our method on several online video sessions that include number of participants and different print and video spoofing attacks. Under proper positions of cameras and with well tuned parameters, the system can detect face three-dimensionality with about 85% accuracy. Fig. 4 demonstrates our results obtained from few indoor video sessions.

VI. Discussion and Conclusions

Our lightweight approach to face anti-spoofing is based on detecting three-dimensionality of landmark points. In contrast to other methods, such as these presented in [6] and [8], we achieve satisfactory results without employing expensive sensors and heavy computations.

Our technique can be employed in alternative settings where deployment of the second camera is impractical (e.g., mobile devices). In particular, the absence of a second RGB camera may be compensated by employing a single movable camera or a single static camera with a mirror. Capturing face reflections in the properly placed mirror attains a second vantage point and, thus this approach readily fits into our framework. If FPS rate and velocity of a movable camera are sufficiently high, then two frames acquired within a short time period approximate stereo imaging. As illustrated in Fig. 5, Algorithm 1 can be integrated into any setting capable of recording the scene from two real or artificial viewpoints.

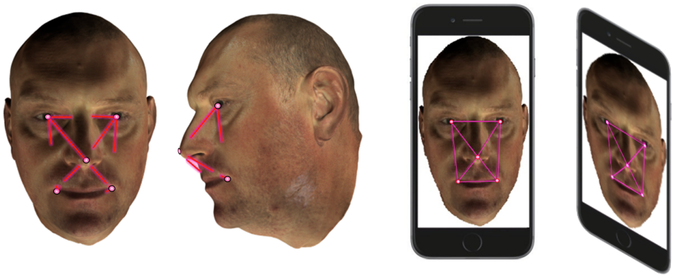

Fig. 2. Five non-coplanar facial landmark points selected to measure (1) with examples of frames taken from different cameras for synthetic 3D face (left) and its images displayed on a flat surface (right).

Our algorithm can be improved in straightforward manner by employing latest state-of-the-art methods for landmark detection and head pose estimation [7]. Employing several sets of five non-coplanar landmarks yields a more robust algorithm version, since then the cross-ratio data can be measured from one or many different facial areas, depending on the head-camera angles.

The proposed approach can be extended to a model with more internal parameters and more landmark estimates. Then, a more sophisticated classifier can be trained using standard machine learning techniques. Further, combination of our approach with device based methods can be employed for detecting hardly noticeable fraud actions such as movable mask attacks. These and other geometry-based methods for anti-spoofing are currently under investigation.

Appendix

The relationship of (1) is formalized as follows:

Theorem 1. *Let $P_1, ...P_5$ be coplanar points in 3D, such that no three points are placed on the same line. Denote by $P'_i = (x'_i, y'_i, z'_i)$ the projection of $P_i = (x_i, y_i, z_i)$ for $i = 1, .., 5$ and denote by A_{kij} and A'_{kij} the area of triangles $\triangle P_k P_i P_j$ and $\triangle P'_k P'_i P'_j$, respectively; then,*

$$\frac{A_{\sigma(5,1,4)} A_{\sigma(5,2,3)}}{A_{\sigma(5,1,3)} A_{\sigma(5,2,4)}} = \frac{A'_{\sigma(5,1,4)} A'_{\sigma(5,2,3)}}{A'_{\sigma(5,1,3)} A'_{\sigma(5,2,4)}}\,, \tag{3}$$

where $\sigma(k,i,j) = (\sigma(k), \sigma(i), \sigma(j))$ denotes a permutation of indices $(1, 2, 3, 4, 5)$.

Proof. Without loss of generality assume $\forall i : \sigma(i) = i$ as presented in equality (1). Each pair of (P_i, P'_i) points are located on the same line passing through the origin, thus

$$\forall i : \frac{x'_i}{y'_i} = \frac{x_i}{y_i}, \frac{y'_i}{z'_i} = \frac{y_i}{z_i}\,. \tag{4}$$

Writing equations of objective and image planes for points P_i and P'_i implies

$$\frac{a}{d}x_i + \frac{b}{d}y_i + \frac{c}{d}z_i + 1 = 0 = \frac{a'}{d'}x'_i + \frac{b'}{d'}y'_i + \frac{c'}{d'}z'_i + 1\,. \tag{5}$$

Combining (4) and (5) together yields the matrix equation of the form

$$N'(a',b',c')X'_{ij} = N(a,b,c)X_{ij}\,, \tag{6}$$

where

$$X_{ij} = \left[\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \frac{1}{z_5}\begin{bmatrix}1\\x_5\\y_5\\z_5\end{bmatrix} \frac{1}{z_i}\begin{bmatrix}1\\x_i\\y_i\\z_i\end{bmatrix} \frac{1}{z_j}\begin{bmatrix}1\\x_j\\y_j\\z_j\end{bmatrix}\right], \tag{7}$$

$$N(a,b,c) = \begin{bmatrix}0&0&0&1\\0&1&0&0\\0&0&1&0\\1&\frac{a}{d}&\frac{b}{d}&\frac{c}{d}\end{bmatrix}, \tag{8}$$

and X'_{ij} is defined similarly to (7) with respect to coordinates of projected points P'_5, P'_i, P'_j. Equation (6) yields the following equality of the determinant products

$$|N'||X'_{ij}| = |N||X_{ij}|\,. \tag{9}$$

Furthermore, the product and ratio of (9) with indices (i,j) taken from (1) implies

$$\frac{|X_{14}|\,|X_{23}|}{|X_{13}|\,|X_{24}|} = \frac{|X'_{14}|\,|X'_{23}|}{|X'_{13}|\,|X'_{24}|}\,. \tag{10}$$

The determinant of X_{ij} can be evaluated using the volume formula for tetrahedron with height d (equals distance between the origin and the objective plane) as follows

$$|X_{ij}| = \frac{1}{z_5 z_i z_j}\det\begin{bmatrix}1&1&1&1\\0&x_5&x_i&x_j\\0&y_5&y_i&y_j\\0&z_5&z_i&z_j\end{bmatrix} \tag{11}$$

$$= \frac{2d}{z_5 z_i z_j}A_{5ij}\,. \tag{12}$$

Finally, (10) and (12) prove the theorem.

□

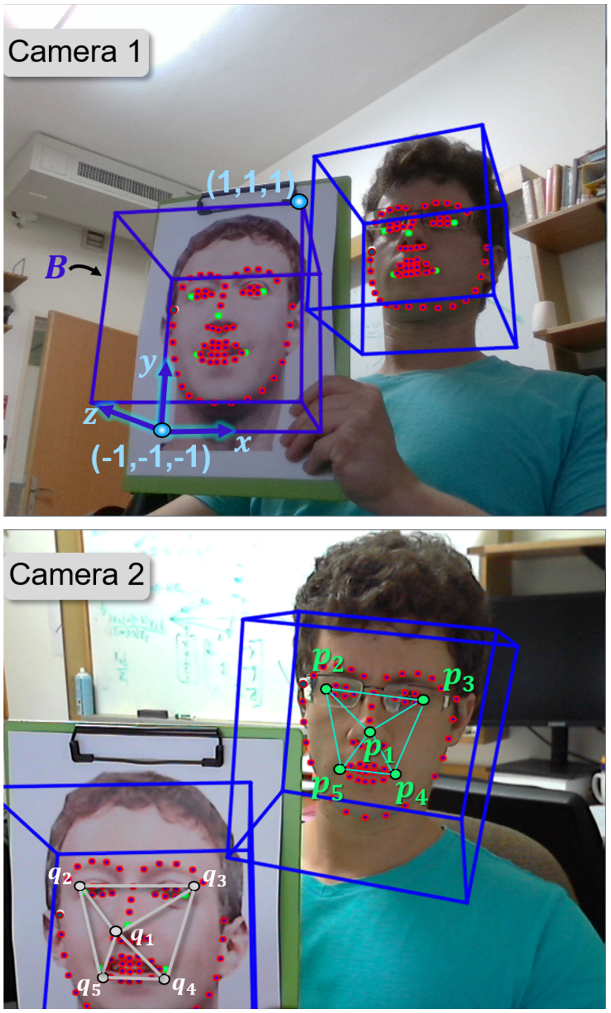

Fig. 3. Illustration of two camera frames simultaneously captured with the following annotations: axes of HCS and corner coordinates of the head bounding box B in NHCS (top); five landmark points $\boldsymbol{p}$ detected in the image and predicted landmark positions $\boldsymbol{q}$, highlighted in green and white, respectively (bottom).

Acknowledgment

Research has been supported in part by the office of the Israeli Chief Scientist–OMEK Consortium and by the Ollendorff Minerva Center for Vision and Image Sciences.

References

[1] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[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] Eamon B Barrett, Paul M Payton, Nils N Haag, and Michael H Brill. General methods for determining projective invariants in imagery. *CVGIP: Image Understanding*, 53(1):46–65, 1991.

[4] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland. Multimodal person recognition using unconstrained audio and video. In *Proceedings, International Conference on Audio-and Video-Based Person Authentication*, pages 176–181. Citeseer, 1999.

[5] Maria De Marsico, Michele Nappi, Daniel Riccio, and Jean-Luc Dugelay. Moving face spoofing detection via 3d projective invariants. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 73–78. IEEE, 2012.

[6] Tejas I Dhamecha, Aastha Nigam, Richa Singh, and Mayank Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In *Biometrics (ICB), 2013 International Conference on*, pages 1–8. IEEE, 2013.

[7] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. *arXiv preprint arXiv:1803.07835*, 2018.

[8] Neslihan Kose and Jean-Luc Dugelay. On the vulnerability of face recognition systems to spoofing mask attacks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2357–2361. IEEE, 2013.

[9] Daniel Riccio and Jean-Luc Dugelay. Geometric invariants for 2d/3d face recognition. *Pattern Recognition Letters*, 28(14):1907–1914, 2007.

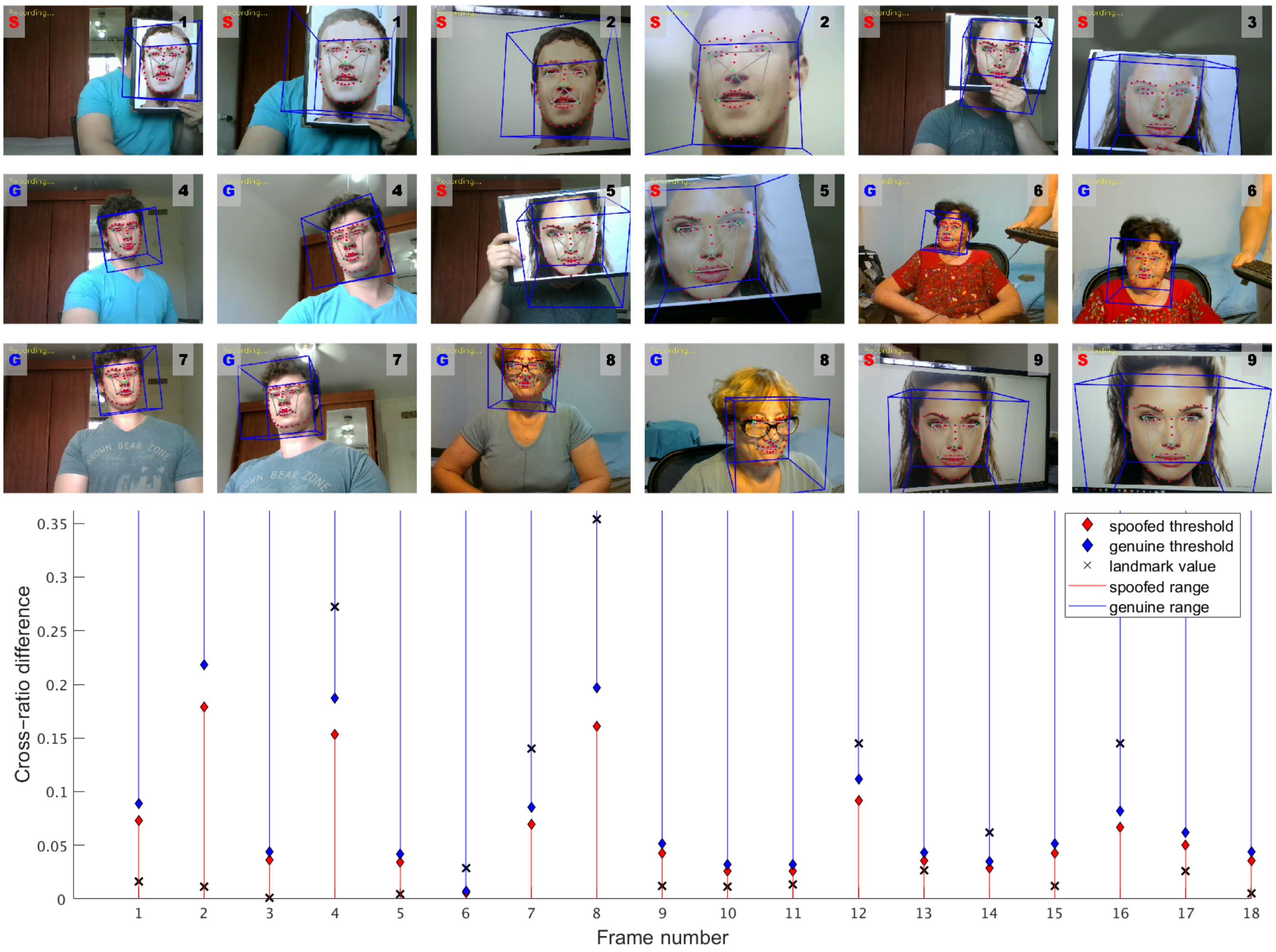

Fig. 4. Face anti-spoofing results are depicted on top by pairs of chosen stereo frames, marked by Algorithm 1 with "S" (spoofed) and "G" (genuine) labels. The bottom plot depicts for each pair of frames the following quantities, employed in the algorithm: thresholds $\delta_1\Sigma_Q$ and $\delta_2\Sigma_Q$, cross-ratio difference Σ_P measured from the landmarks, ranges of Σ_P values that belong to "spoofed" and "genuine" classes.

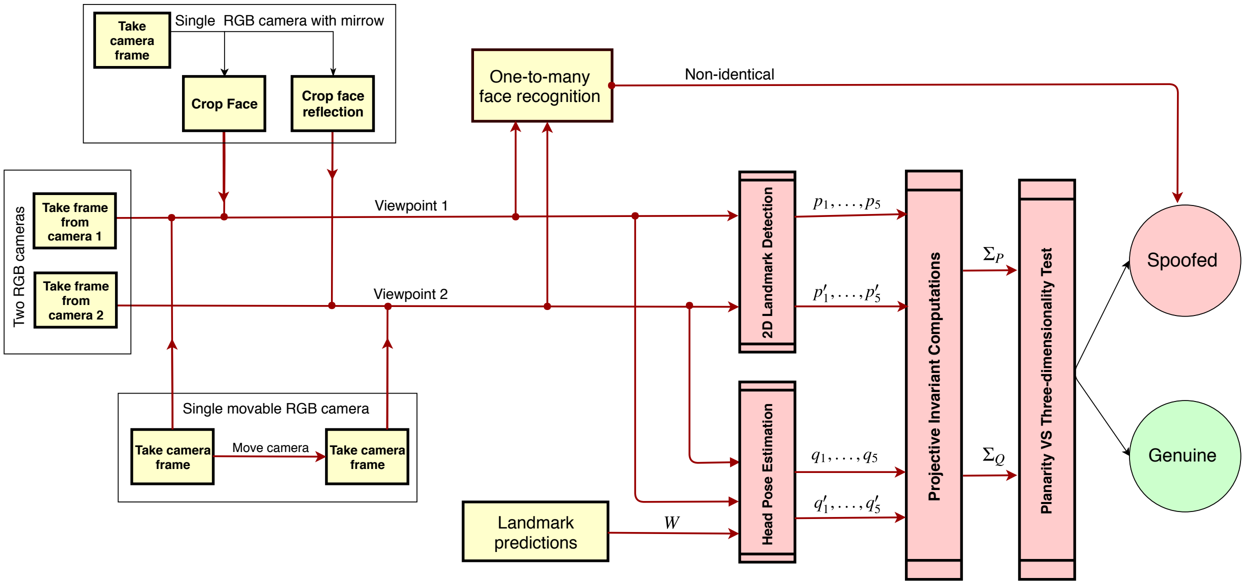

Fig. 5. Flowchart of a general authentication system based on Algorithm 1, where $\boldsymbol{p}, \boldsymbol{p'}$ and $\boldsymbol{q}, \boldsymbol{q'}$ denote landmarks of the 1st and 2nd viewpoints, respectively.