Pattern Anal Applic
DOI 10.1007/s10044-016-0587-2

THEORETICAL ADVANCES

Gaze stability for liveness detection

Asad Ali[1] · **Sanaul Hoque**[1] · **Farzin Deravi**[1]

Received: 26 November 2015 / Accepted: 20 October 2016

Abstract Spoofing attacks on biometric systems are one of the major impediments to their use for secure unattended applications. This paper explores features for face liveness detection based on tracking the gaze of the user. In the proposed approach, a visual stimulus is placed on the display screen, at apparently random locations, which the user is required to follow while their gaze is measured. This visual stimulus appears in such a way that it repeatedly directs the gaze of the user to specific positions on the screen. Features extracted from sets of collinear and colocated points are used to estimate the liveness of the user. Data are collected from genuine users tracking the stimulus with natural head/eye movements and impostors holding a photograph, looking through a 2D mask or replaying the video of a genuine user. The choice of stimulus and features are based on the assumption that natural head/eye coordination for directing gaze results in a greater accuracy and thus can be used to effectively differentiate between genuine and spoofing attempts. Tests are performed to assess the effectiveness of the system with these features in isolation as well as in combination with each other using score fusion techniques. The results from the experiments indicate the effectiveness of the proposed gaze-based features in detecting such presentation attacks.

Keywords Biometrics · Liveness · Spoofing · Fusion · Presentation attacks · Feature extraction

✉ Sanaul Hoque
S.Hoque@kent.ac.uk

[1] School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NT, UK

1 Introduction

Despite the widespread adoption of biometric recognition systems in recent decades, there still remain vulnerabilities to increasingly sophisticated spoofing attacks that can undermine the trust in such systems. The artefacts used for such attacks may be created from the biometric information of genuine users and presented at the system sensor(s). An impostor can present a fake biometric sample of a genuine user to a biometric recognition system to gain access to unauthorized data or premises. This type of spoofing is a direct attack on the sensor (also known as "presentation attack"); the impostor does not require any prior knowledge about the internal operation of the biometric system. To prevent such sensor-level attacks, biometric systems need to establish the "liveness" of the source of an acquired sample. In the context of biometric counter-spoofing, liveness detection refers to such situations where the attacker uses an artefact presented at the sensor to subvert the system. In this sense, there may still be a live human operator manipulating the artefact that mimics some attribute of the "live" subject whose identity is being compromised.

Amongst biometric modalities, face recognition has emerged as being widely adopted, accurate and convenient and is, therefore, used for a variety of security applications. But face recognition systems are more vulnerable to abuse compared to other biometric modalities, because a simple photograph or video of a genuine user can be used to deceive such systems [1]. Therefore, by introducing a liveness detection mechanism, the security of such systems can be substantially improved.

Photographs, masks and video replay are some of the means for spoofing that may be used for attacks at sensor level. Photograph spoofing can be prevented by detecting

Published online: 10 November 2016

motion, smiles, eye blinks, etc. Such techniques can be deceived by presenting a video of the genuine user to the face recognition system. However, the subtle differences between a photograph (or video) of an individual and the live person can be used to establish liveness of the presentation at the sensor.

Another potential source of liveness information could be the nature of user interactions with the system, which can be captured and analysed in real time. Vision is an active process where the viewer seeks out task-relevant visual information by actively controlling their gaze using eye, head and body movements. Similarly, sophisticated hand/eye coordination is essential for performing sports, handwriting, etc. All animals with developed visual systems learn, practice and improve such coordination acts throughout their lifetime to reach a level of subconscious spontaneity. Such spontaneity is usually absent when a task demands coordination of body parts which has not been practiced naturally for a long time. This observation is exploited in the work reported here to ascertain liveness.

In this paper, we present a novel challenge/response mechanism for face recognition systems by tracking the gaze of a user in response to a moving visual stimulus or target using a standard webcam. The stimulus is designed to facilitate the acquisition of distinguishing features from collinear and colocated sets of points along the gaze trajectory.

This paper provides a unified and formal framework for bringing together the authors' previous work that dealt with features based on gaze stability [2, 3]. The novel contributions of this work include a mathematical generalization of the originally proposed features to incorporate more complex stimulus trajectories used as a challenge and extension of the experimental work to include more test subjects and presentation attack scenarios. Additionally, fusion of liveness information from different gaze-based features is also explored in the present work, resulting in enhanced performance.

The paper is organized as follows. In Sect. 2, a brief overview of the state of the art is presented. Section 3 describes the proposed techniques while Sect. 4 reports on their experimental evaluation. Finally Sect. 5 provides conclusions and offers suggestions for further work.

2 Related work

Various approaches have been presented in the literature to establish liveness and to detect presentation attacks. Liveness detection approaches can be grouped into two broad categories: active and passive. Active approaches require user engagement to enable the biometric system to establish the liveness of the source through the sample captured at the sensor. Passive approaches do not require user cooperation or even user awareness but exploit involuntary physical movements, such as spontaneous eye blinks, and 3D properties of the image source.

2.1 Passive techniques

Passive anti-spoofing techniques are usually based on the detection of signs of life, e.g. eye blink and facial expression. For example, Pan et al. [4] proposed a liveness detection method by extracting the temporal information from the process of the eye blink. Conditional random fields were used to model and detect eye blinks over a sequence of images. Jee et al. [5] proposed a method that uses an ordinary camera and analyses sequences of images captured. The centres of both eyes in the facial image are located, and if the variance of each eye region is larger than a preset threshold, the image is considered to be live, and if not, the image is classified as a photographic artefact. Wang et al. [6] presented a liveness detection method in which physiological motion is detected by estimating the eye blink with an eye contour extraction algorithm. They use active shape models with a random forest classifier trained to recognize the local appearance around each landmark. They also showed that if any motion in the face region is detected, the sample is considered to be captured from an impostor.

Kollreider et al. [7–9] combined facial components (e.g. nose, ears) detection and optical flow estimation to determine a liveness score. They assumed that a 3D face produces a special 2D motion. This motion is higher at the central facial parts (e.g. nose) compared to the outer regions (e.g. ears); the parts nearer to the camera move differently to those which are further away in a live face. A photograph, by contrast, generates constant motion at various face regions. They also proposed a method, which uses lip motion (without audio information) to assess liveness [9].

Some anti-spoofing techniques are based on the analysis of skin reflectance, texture, noise signature, etc. Li et al. [10] explored a technique based on the analysis of 2D Fourier spectra of the face image. Their work is based on the principles that the size of a photograph is smaller than the real image and the photograph is flat. It therefore has fewer high-frequency components than real face images.

Kim et al. [11] proposed a multi-classifier method for detecting fake attempts by combining frequency information from the power spectrum and texture information obtained using local binary pattern (LBP) features. They utilized support vector machine (SVM) classifiers to train separate liveness detectors using the two types of feature vectors extracted. The decision values of these two SVM classifiers were then used as 2D feature vectors for the subsequent trainable fusion stage.

Komulainen et al. [12] explored the use of dynamic texture information for spoofing detection. They argued that masks and 3D head models are rigid, whereas real faces are non-rigid with contractions of facial muscles resulting in temporal deformation of facial features, such as moving eyelids and lips. The structure and dynamics of the micro-textures that characterize real faces were used in their proposed approach to spoof detection. They used spatiotemporal (dynamic texture) extensions of the local binary pattern in this approach. Komulainen et al. [13] further extended their work and explored the fusion of micro-texture with motion. The motion-based technique measures the correlation between the head movement and background scene. They also explored the potential of the fusion of different visual cues and showed that the performance of each method can be improved by performing score-level fusion.

Lagorio et al. [14] proposed a liveness detection method, based on the 3D structure of the face, to identify an impostor presenting a 2D image of a genuine user to spoof a face recognition system. The method computed the 3D features of the captured facial image data to detect whether a human face has been presented to the acquisition camera. They collected a 3D face database using a stereo camera system for performance evaluation. Skin reflectance models, based on non-thermal hyperspectral imagery, have been used to develop skin/face detection and classification algorithms [15, 16] which can be used for face liveness detection.

Replay of pre-recorded video can be used to spoof facial liveness detection measures. Many of the algorithms used for detecting photograph spoofing attacks are likely to be susceptible to such video-based attacks. Video spoofing thus presents an even greater challenge. Pinto et al. [17] investigated a method for detecting video-based face spoofing. They used the noise signatures generated by the recaptured video to discriminate between live and fake attempts. They used the Fourier spectrum, computation of the visual rhythm and extraction of the grey-level co-occurrence matrices as feature descriptors. These were classified using a support vector machine (SVM) and partial least squares regression to detect liveness.

2.2 Active techniques

Systems based on the challenge–response approach belong to the active category, where the user is asked to perform specific activities to ascertain liveness such as uttering digits or changing their head pose. For instance, Frischholz et al. [18] investigated a challenge–response approach to enhance the security of a face recognition system. The users were required to look in certain directions (challenge), which were chosen by the system randomly and the head pose (response) is estimated and compared in real time to establish liveness. Sharma [19] presented a similar technique in which the user was asked to perform some activities such as chewing or smiling. The camera captured sequences of images and extracted the features from the facial images using a correlation coefficient and image extension feature. They calculated skin elasticity, using a discriminant analysis method. Then the output was compared with the stored database to discriminate between fake and real images.

The liveness detection technique presented here is based on gaze tracking, estimated by measuring the movement of the pupil centre. Pupil centres can be easily extracted with limited computational effort. Pupil centre positions, while not indicating the true direction of gaze, are strongly correlated with it and provide a useful indicator of gaze, especially on platforms where computational resources are limited (e.g. mobile devices). The underlying hypothesis is that gaze stability and consistency should be greater in genuine user attempts when compared with spoofing attacks. This phenomenon is then exploited to differentiate between such presentations. Clearly, using additional facial landmarks may help improve the accuracy and robustness of the system. However, the aim of this paper is to indicate the general principles involved and pave the way for future explorations.

2.3 Gaze stability

The algorithms proposed in this paper are based on the assumption that the spatial and temporal coordination of the movements of eye, head and hand involved in the task of following of a visual stimulus is significantly different when a genuine attempt is made compared with certain types of spoof attempts. The task requires head/eye fixations on a simple target that appears on a screen in front of the user, and in the case of a photograph spoofing attack, visually guided hand movements are also required to orientate the photographic artefact to point in the correct direction towards the challenge item on the screen.

It is expected that the head pose and direction of gaze will be different when photograph spoofing is attempted as coordination may only be maintained by delaying the hand movements until the eye is available for guiding the movement [20]. The introduction of hand movements is also likely to change the relationship between head and eye movements, as the coordination of eye and head in gaze changes is usually a consequence of synergistic linkage rather than an obligatory one [20–22]. Therefore, it is assumed that accurately directing the photograph to a particular orientation indicated by the visual stimulus on the screen is likely to be less repeatable than merely looking at the stimulus. Hence, the variances in measured

gaze parameters can be used to distinguish genuine from fake attempts as described in the rest of the paper.

Although the proposed approach may require additional resources compared to simpler techniques such as blink detection, it provides protection for a wider range of attack scenarios than possible with such techniques. The proposed technique may also provide an effective basis for liveness detection on devices where a display screen and image sensor are inherently available. Clearly any liveness detection approach presents a trade-off between convenience and security and it is expected that the present contribution will further enrich the available options to system designers.

3 Liveness detection through gaze tracking

The scenario considered in this paper is that of a face recognition system using an ordinary camera (webcam). The spoofing attack is by means of an impostor attempting authentication by holding a photograph or a photograph mask or playing a recorded video of a genuine client to the camera. A typical setting is depicted in Fig. 1. A visual stimulus appears on the display which the client is asked to follow and the camera (sensor) captures facial images at various positions of the stimulus on the screen. A control mechanism is used to ensure the placement of the target and the image acquisition are synchronized. The system extracts facial landmarks in the captured frames, computes various features from these landmarks, which are then used to classify the attempt as either genuine or fake.

In Fig. 2a, a genuine user is seen to be tracking the challenge to establish liveness, while the impostor is responding to the challenge by carefully moving a high-quality printed photograph in Fig. 2b, holding a mask in Fig. 2c or replaying a video in Fig. 2d to gain access to the system.

3.1 Visual stimulus and user response acquisition

A small shape is randomly presented, one after another, at D distinct locations on the screen. A simple cross sign was used as the challenge stimulus as shown in Fig. 3. In this figure, the dots indicate the chosen locations in which the cross sign may randomly appear. The cross sign is chosen as it is commonly used to direct attention to a specific point. One could use other symbols, shapes or words as the stimulus to direct the user gaze to certain locations on the screen. Let C be a set of these coordinates.

$$C = \{c_1, c_2, \ldots, c_d, \ldots, c_D\} \quad \text{where,} \quad c_d = (x, y); \quad d = 1, \ldots, D \tag{1}$$

It is not necessary to space these locations uniformly, but ideally these should not be too close to one another to encourage greater head/eye movements. It is so arranged that some of these locations are visited by the stimulus several times during a challenge session. Let P be the sequence of M such presentations.

$$P = \{p_1, p_2, \ldots, p_m, \ldots, p_M\} \quad \text{where,} \quad p_m \in C; m = 1, \ldots, M \tag{2}$$

The stimulus appears in a random sequence to prevent predictive video attacks. Face images are then captured at each presentation of the stimulus.

3.2 Facial landmark detection and feature extraction

The images thus captured during the challenge–response operation were processed using STASM [23] in order to extract facial landmark points. STASM returns 68 different landmarks on the face region using an active shape model algorithm. The coordinates of some of these landmarks were used for feature extraction in the proposed scheme. Feature extraction methods proposed here are based on collinearity and colocation properties of the presented stimulus during the challenge.

3.3 Collinearity features

A set of points lying on a straight line is referred to here as a collinear set of points, and this property of this set of points is hereby referred to as collinearity. Collinearity features are, therefore, extracted from sets of images

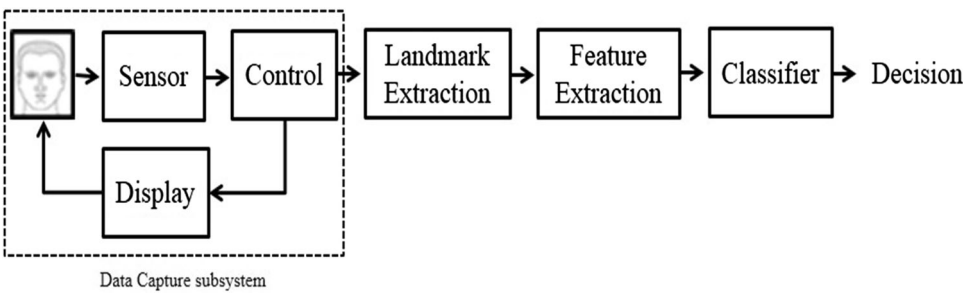

Fig. 1 Proposed system block diagram

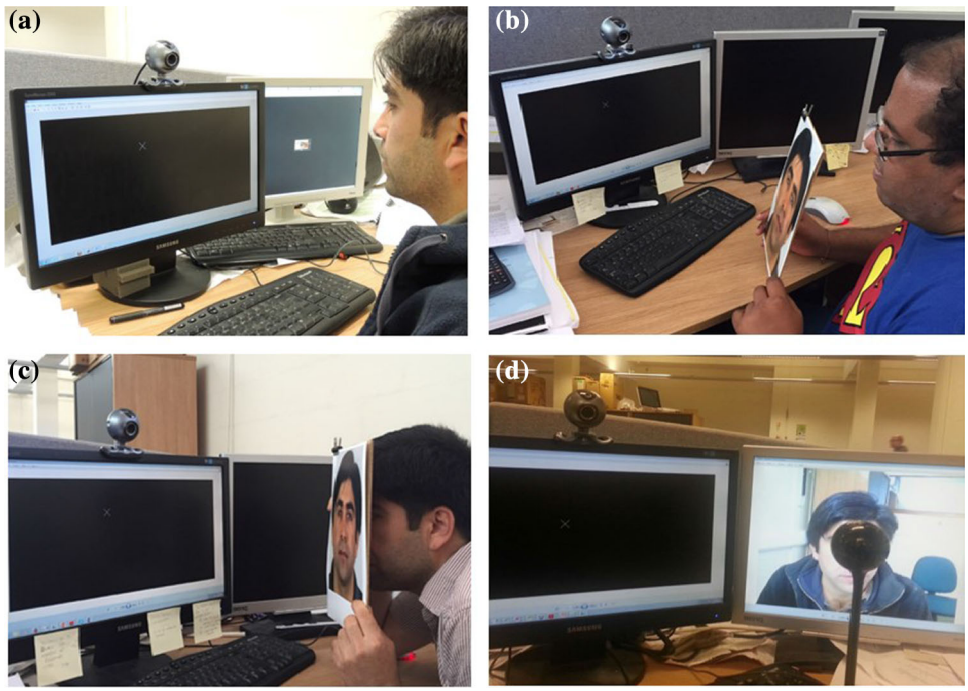

Fig. 2 Example of **a** genuine attempt, **b** photograph spoof attempt, **c** 2D mask spoof attempt and **d** video spoof attempt

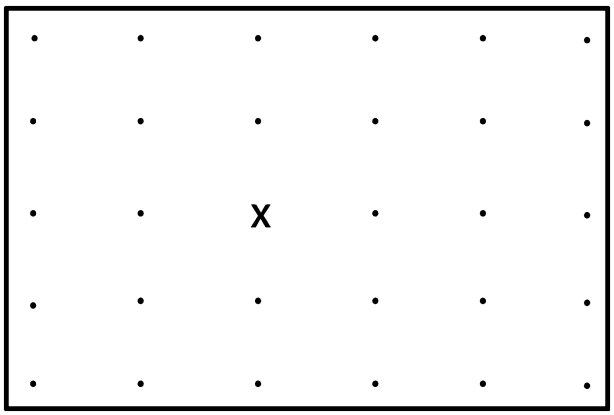

Fig. 3 Stimulus shape and selected display positions

captured when the stimulus is on a given line. In the investigations reported here, only horizontal or vertical collinearity cases were studied.

Let S_l be a collinear subset of C, where the stimuli are horizontally aligned. $S_l \subseteq C, l = 1, \ldots, L$ where L is the number of horizontally aligned sets of stimulus locations. For $(x, y) \in S_l, y = a_l$ where a_l is constant. Let R be the set of landmark locations in the captured images. For a given landmark k (e.g. centre of the left eye)

$$R = \{r_{p_1}, r_{p_2}, \ldots, r_{p_i}, \ldots, r_{p_M}\} \tag{3}$$

where, $r_{p_i} = \{(u_{ik}, v_{ik})\}\ 1 \leq i \leq M, 1 \leq k \leq K$ and (u, v) are the pixel positions in the image coordinate system and K is the total number of such landmarks. Individual subjects moved their eyes and heads by different amounts in response to the movement of the stimulus. They may also be sitting in different positions relative to the screen and camera in each session. So in order to remove these user- and session-dependent factors in estimating gaze-based features, the data were normalized. The spatial coordinates of the landmarks for each session were normalized using the Min–Max normalization technique [24] prior to feature extraction. Min–Max algorithm was used in this application due to its simplicity and the absence of outliers in the genuine attempts. The (u, v) coordinates used in this paper refer to these normalized values.

For each S_l there is a corresponding subset of R. Let this be denoted by T_l.

$$T_l \subseteq R, l = 1, \ldots, L \tag{4}$$

For any given landmark, k, let $v_{ik} = f(u_{ik})$ denote the trajectory of the facial landmark in response to the challenge. Since the trajectory of the challenge S_l is horizontal, a horizontal response can be assumed and this may be approximated by the equation of a horizontal line.

$$\hat{v}_k = b_k \quad \text{where } b_k \text{ is a constant} \tag{5}$$

The particular value of b_k depends on the system set-up. Let, e_{ik} denote the deviation between the estimated $\hat{v}_{ik}$ and observed v_{ik} (see Fig. 4), i.e.

$$e_{ik} = v_{ik} - \hat{v}_{ik} \tag{6}$$

For simple horizontal collinearity, $\hat{v}_{ik}$ is calculated as the mean of the observed v_{ik}. So, the mean square error (MSE) for T_l will be

$$E_{lk} = \frac{1}{N}\sum_i e_{ik}^2 = \frac{1}{N}\sum_i (v_{ik} - \hat{v}_{ik})^2 \tag{7}$$

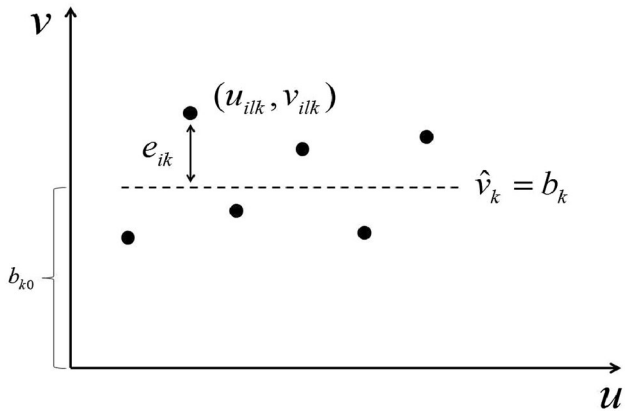

Fig. 4 Observed locations (*bullet symbols*) and expected locus of the landmark positions (*dash symbols*)

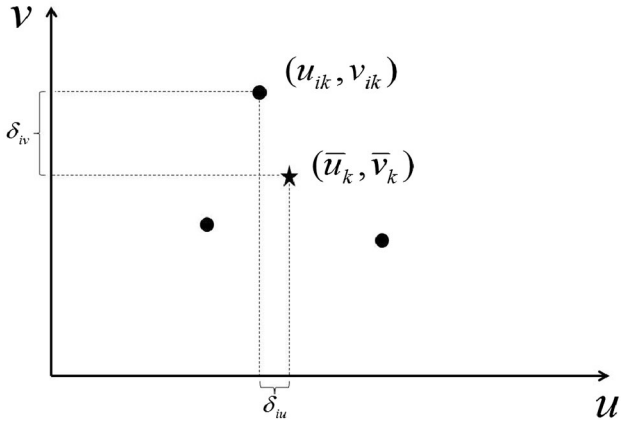

Fig. 5 Observed (*bullet symbols*) and expected (*asterisk*) landmark positions

where N is the cardinality of T_l. A similar expression can be derived when the challenge is vertically aligned. A generalized form of the expression for collinearity feature along any straight line is given in "Appendix". As there are multiple face landmarks as well as several stimulus challenge trajectories, a feature vector, F_{colin}, can be constructed from the concatenation of these MSE values (and optionally other feature values) and used for liveness detection.

$$F_{\text{colin}} = [E_{11}, E_{12}, \ldots, E_{1K}, E_{21}, \ldots, E_{ik}, \ldots, E_{LK}] \tag{8}$$

3.4 Colocation features

The colocation features are extracted from the images acquired when the stimulus is presented at a given location several times. This stimulus can be considered as a special case of collinear trajectory where the line is reduced to a single point. Since the coordinates of the stimulus are identical, it can, therefore, be expected that the coordinates of the facial landmarks in the corresponding frames should also be closely spaced if not coincident. This should result in a significantly smaller variance in the observed landmark coordinates in genuine attempts than that in fake attempts.

Figure 5 illustrates the observed coordinates (u_{ik}, v_{ik}) of a given landmark k in response to the stimulus presented at the same location at different times. To quantify the deviation from perfect colocation, the variances in the observed landmarks are calculated.

Let Q_w be a subset of P where the stimuli appeared at the same location c_w on the screen at different times.

$Q_w \subseteq P$, $w = 1, \ldots, W$ where W is the number of such colocation sets used in the challenge. Let T_w be the corresponding subset of R.

$$T_w \subseteq R, w = 1, \ldots, W \tag{9}$$

Let σ_{uk}^2 and σ_{vk}^2 denote the variances of the observed landmarks along u the v and directions, respectively.

$$\begin{aligned} \sigma_{uk}^2 &= \frac{1}{N}\sum_i \delta_{iu}^2 = \frac{1}{N}\sum_i (u_{ik} - \bar{u}_k)^2 \\ \sigma_{vk}^2 &= \frac{1}{N}\sum_i \delta_{iv}^2 = \frac{1}{N}\sum_i (v_{ik} - \bar{v}_k)^2 \end{aligned} \tag{10}$$

where $(u_{ik}, v_{ik}) \in T_w$, $(\bar{u}_k, \bar{v}_k)$ is the mean of the observed landmark locations and N is the cardinality of T_w.

Let $\Gamma_{wk} = [\sigma_{uk}^2, \sigma_{vk}^2]$. As there are K different landmarks as well as W colocation subsets, a colocation feature vector, F_{coloc}, can be constructed from the concatenation of these values and used for liveness detection.

$$F_{\text{coloc}} = [\Gamma_{11}, \Gamma_{12}, \ldots, \Gamma_{1K}, \Gamma_{21}, \ldots, \Gamma_{wk}, \ldots, \Gamma_{WK}] \tag{11}$$

Many other features can be extracted from these facial landmarks. All these can be combined into a global feature vector,

$$F = [F_{\text{colin}}, F_{\text{coloc}}, F_{\text{other}}, \ldots]. \tag{12}$$

In order to spoof the system the attacker may hold the photograph still (without moving the photograph in response to the stimuli) to generate near-perfect collinearity and colocation features. However, such an attack is easily detected by measuring the overall spread of landmark locations in the captured images during the entire presentation session, R, and check that this value is above a certain threshold to detect such presentation attacks [2]. In fact, two thresholds are used in the operation of the proposed system. One movement threshold is used to check if the attacker is trying to subvert the liveness detection system by minimizing movements of the artefact in response to the stimulus. The other threshold is used to detect if the movements of the artefact are resulting in repeatable positioning of the eyes in response to the stimulus. Therefore, if a skilled attacker intentionally makes micro-movements to defeat the system they will be caught by the first detection system and if they do not use micro-movements, they will be caught by the second detection

system. There thresholds were empirically set for the given test configuration. However, it is relatively easy to adjust these to match any application scenario.

4 Experiments

The data acquisition system set-up was similar to the one shown in Fig. 6. It consists of a webcam, a PC and a display monitor. The distance between the camera and the user was approximately 750 mm. This distance was not a tight constraint but had to be such that the facial features could be clearly acquired by the camera.

Data were collected from 30 volunteers of both genders aged between 20–45 years. Three potential presentation attack scenarios were studied: photograph attack, mask attack and video replay attack. Each subject provided data for genuine attempts as well as for the three attack scenarios, thus creating 30 sets of data for each scenario. For hand-held photograph spoofing attacks, a high-quality colour photograph of a genuine user was held in front of the camera while the volunteer attempted to follow the stimulus. In the case of photograph mask spoofing attacks, a high-quality colour photograph of a genuine user with holes made in the place of the pupils was held by the user in front of the eyes as a mask and used to follow the stimulus.

Photographs of both male and female subjects were chosen for the hand-held photograph and the photograph mask spoofing trials. The photographs were printed on A4 matt paper, which bends easily. These photographs were from two subjects. All volunteers used one of these two subjects for the spoofing attempt. Photographs from more subjects could have been used, but as in this study only spoofing attack detection (and not face recognition) was the focus there was no need to include a wide range of faces in the construction of attack artefacts. Hard cardboard was attached to the back of the photograph to attempt to minimize any unintended deformation of the paper. For the photograph mask attempt, three different photograph sizes (small, medium and large) with different pupillary distance (PD) were printed. The reason for producing a set of photographs with pupillary holes at different distances was to better fit the facial dimensions of the attackers with different PDs. Before the mask was given to the attacker the pupillary distance was measured, using a pupillary distance ruler. The photograph with the PD closest to the attackers PD was used for the attempt. The diameter of the hole in pupil centre was 4 mm. The 4-mm hole was large enough to see through to follow the challenge. A bigger hole could have made the task of gaze direction easier for the attackers but may have exposed other biometric indicators of the attacker (e.g. iris) that would have undermined their spoofing attempt [25]. During a real attempt, the video of the genuine user was recorded and used for subsequent replay attacks. This database is comparable in size to other databases used for evaluation of liveness detection algorithms such as replay attack database which has 50 clients [26].

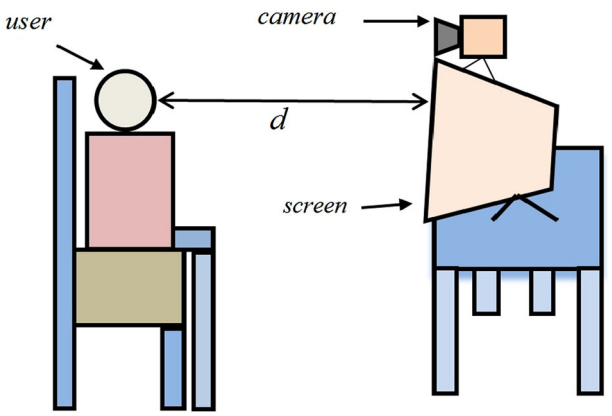

Fig. 6 Data acquisition set-up

In this implementation, the stimulus was displayed on the screen at 30 distinct locations (i.e. $D = 30$), as shown in Fig. 3, in a random order visiting each position 3 times (thus, $M = 90$). Typically 225–275 ms is needed for gaze fixation in reading tasks [22]. In this work, a 1 s delay between each presentation is used to provide ample time for the users to fixate their gaze. Total duration of the challenge was about 2 min. The challenge duration for the data collection sessions used in these experiments is relatively long for most practical applications. However, the initial experiments used a large number of points covering the whole screen to explore the sensitivity of the algorithm to the various challenge locations. Nevertheless, the liveness check can be achieved using a much smaller number of locations to reduce the duration of the challenge. Results supporting this assertion are given in Sect. 4.C.

In this particular challenge, the locations are so arranged that there are 33 collinear sets and 30 colocation sets (i.e. $L = 33$, $W = 30$). For each presentation of the stimulus, the camera acquires a facial image. The image resolution was 352×288 pixels. This resolution provided adequate picture quality for locating the facial landmarks. Using higher-resolution images with STASM may not improve landmark detection but will increase the processing time [23]. The maximum expected gaze deviation from the normal to the screen is approximately 15 degrees for the experimental set-up. If the subject turns their head beyond this pose angle, they are not following the instructions for using the system. In such a case, landmark detection may be compromised. Such frames are excluded in the feature extraction phase. If this occurs 5 times or more in a single presentation attempt, the whole attempt is excluded from

the experiment, and the user is asked to try again in a new attempt. The choice of this number is determined by the number of points and their placement in the stimulus. Such attempts are considered to be cases of failure to detect liveness. If the number of such frames is 5 or less, then the missing landmark values were substituted by estimated data from the remaining landmarks. The presentation attempts where there are more than 5 frames in which face (or facial landmarks) cannot be detected by the system were not used in subsequent experiments. There were 21 out of 120 attempts where facial landmarks were not detected in more than 5 frames. Only 3 of them were genuine attempts. If the system policy was changed so that all such as attempts were classified as impostor attacks, the overall performance of the proposed system would be further improved. However, in this work only attempts with good landmark detection were considered in order to focus on the evaluation of the proposed features.

For the experiments reported here, a subset of the database composed of 92 presentation attempts comprising of genuine attempts and attacks using hand-held photographs, photograph masks and video replays were used. 65% of the data were used for training and the remaining was used for testing. Therefore, for each individual attack scenario (photograph, mask, video) 15 genuine and 15 impostor attempts were randomly selected for training. However, for scenarios where different attack artefacts are combined in the evaluation 15 genuine and 45 impostor attempts were randomly selected for training. These data are available to other researchers upon request.

4.1 Facial liveness detection system

In the works reported here, the effectiveness of the two proposed features, collinearity and colocation, in detecting liveness were investigated. Subsequently, the use of both these sources of information in combination with each other was investigated. Several schemes were set up to explore the gain in accuracy achieved by combining features extracted from both eyes in a multi-classifier configuration [27].

Fusion of information from multiple sources can be achieved in a number of ways. In the first scheme investigated, feature vectors from right and left eyes were concatenated as shown in Fig. 7. Various classifiers were then used to obtain classification results for the fused feature set.

Alternatively, using score fusion, classifiers were independently trained to obtain the individual classification score for each eye. The score from the primary classifiers were combined at the fusion stage for liveness detection. The scheme is illustrated in Fig. 8.

4.2 Liveness detection performance measures

Face liveness detection is a two-class problem. There are four possible outcomes of the classification process hereby referred to as: true positive, true negative, false negative and false positive, with "positive" indicating a live/genuine detection decision. When a genuine (live/non-spoof) attempt is classified as genuine and a false (fake/spoof) attempt is classified as genuine, these are termed true positive (TP) and false positive (FP) classifications, respectively. Similarly, when a genuine attempt is classified as a fake and fake attempt is classified as fake these are called false negative (FN) and true negative (TN), respectively. FP and FN are the error outcomes of the process and the likelihoods of their occurrence are reported as false positive rate (FPR) and false negative rate (FNR) in this report in order to facilitate the assessment and comparison of system performance. The term true positive rate (TPR) is also used and is equal to (1-FNR). The term true negative rate (TNR) is equal to (1-FPR) [28]. The total error rate (TER) is also used to quantify the overall performance of the system at a particular operating point and is defined in Eq. 13.

$$\mathrm{TER} = \frac{(\mathrm{FP} + \mathrm{FN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{13}$$

4.3 Experimental results

Error rates were calculated for a range of system parameters and are reported in this section. Each experiment was run 400 times with random partitioning of the data into disjoint training and test sets and the average performances from these runs were reported in the paper. True positive

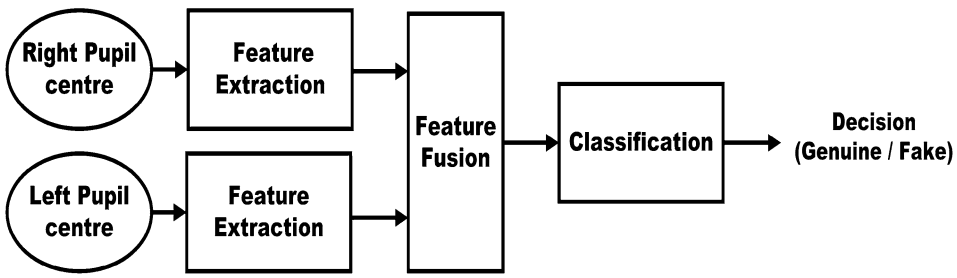

Fig. 7 Proposed liveness detection scheme using feature fusion

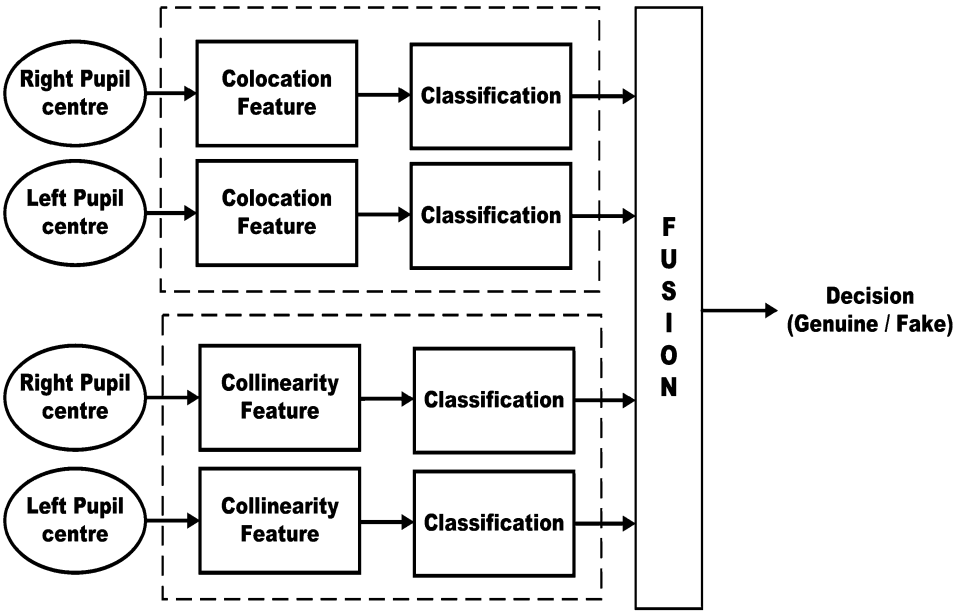

Fig. 8 Proposed liveness detection scheme using score fusion

Table 1 Comparison of feature fusion performance for different classifiers (TPR at FPR = 0.10)

Feature	Attack type	Classifier		
		k-NN	SVM	LDC
Collinearity	Photograph	0.53	0.37	0.12
	Mask	0.60	0.44	0.19
Colocation	Photograph	0.25	0.25	0.20
	Mask	0.18	0.24	0.15

Table 2 TPR at FPR = 0.10 using the entire feature set

Feature sets	Photograph	Mask	Video replay
Collinearity	0.55	0.71	0.99
Colocation	0.43	0.25	0.83
Collinearity and colocation	0.58	0.69	0.99

rates at a set of pre-defined FPR values were obtained and used for comparison.

Table 1 shows the performance for various classification schemes for feature fusion using photograph and mask attacks. It is clear from the table that the k-NN classifier performs better than the alternatives explored. Hence, the k-NN classifiers were used to investigate the performance of the system for the remaining experiments.

In the score fusion schemes, only the k-NN classifiers were employed as the primary classifier to obtain the individual classification scores and then the product rule was employed for the fusion phase. Table 2 presents the TPR values for three spoofing attack detection scenarios. It is obvious that, for both the collinearity and the colocation feature-based implementations, the error rates are lower, in most cases, than what was achieved while using the feature fusion scheme. The video replay attack detection outperformed the other two types of attacks. The collinearity features were superior to the colocation features.

In the subsequent implementation, the collinearity and the colocation feature schemes themselves are combined using the product rule, as shown in Fig. 8, and the corresponding TPR values are presented in the bottom row of Table 2. The TPR for hand-held photograph attack detection further improved, whereas the photograph mask detection performance was slightly decreased, and video replay attack detection remained the same. These experiments indicate that the score-based fusion was more effective than the feature fusion scheme.

All subsequent experiments were, therefore, carried out for the score fusion scheme using the k-NN classifier only. The value of k was optimized with respect to the leave-one-out error rate on the training data. Each experiment was run 400 times with random partition of available data for training and testing, which resulted in different optimum k values for each run. The mean optimal k values were found to be 7 and 6 for collinearity and colocation schemes, respectively.

Figure 9 shows the receiver operating characteristic (ROC) curves [28] for combined collinearity and colocation features using the proposed fusion scheme. The system displayed a near-perfect performance in the case of video attack detection for a range of FPRs. The performance of the system for mask attack detection was marginally better than that achieved for photograph attack detection.

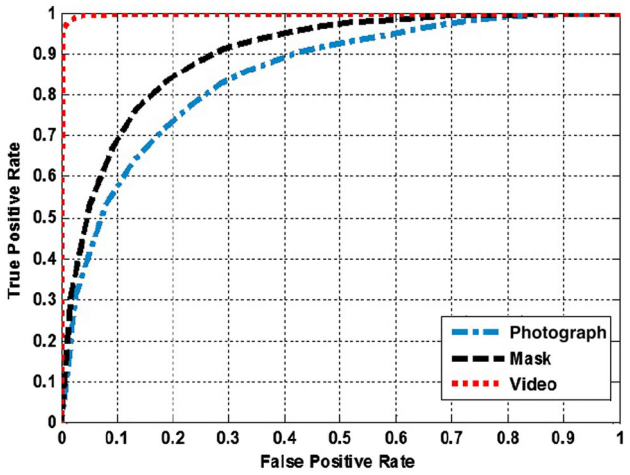

Fig. 9 ROC curve using entire feature vector

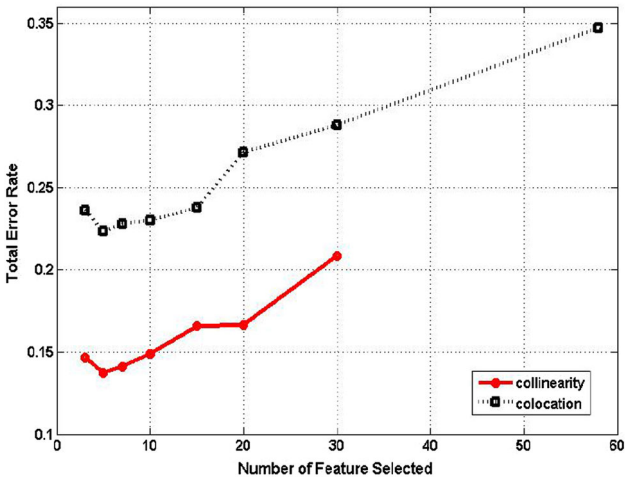

Fig. 10 Variation in total error rate with feature dimension

In order to establish the trade-off between the feature dimensionality and liveness detection, a forward feature selection method [14] was used. The feature selection method was run 400 times with random sets of data for training and testing. This resulted in different rankings of features for each run. The feature that most frequently had the first rank was assigned the first overall ranking. This procedure was repeated for all the other ranks so that the feature that appeared most frequently at rank N was given rank N in the overall ranked list. Figure 10 presents total error rates as a function of the number of features selected to find reduced feature subsets for collinearity and colocation features. In this experiment, the photograph and mask attack modalities were combined as a single attack class. The combination of these attack modalities allows the establishment of an optimal feature subset that can be used for all of these major spoofing challenges. Video attack data were excluded from this feature ranking

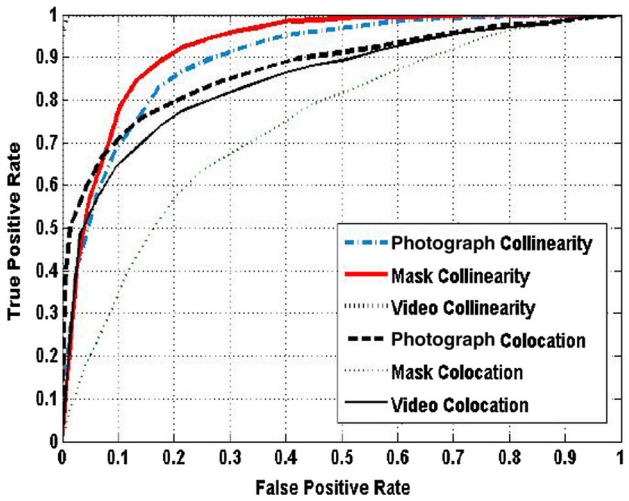

Fig. 11 ROC curve of the proposed system using reduced feature set

exercise as the system already performs very well in detecting video spoofing attacks.

As shown in Fig. 10, the lowest total error rate was observed when the feature dimension was significantly reduced. As referred to previously, a decreased number of features implies that the challenge will need to be presented at fewer locations; therefore, the time duration of the challenge can be substantially shortened. Feature reduction can, therefore, not only shorten the time duration for this approach but also improve its performance. The collinearity and colocation feature performance for photograph, mask and video spoofing attacks using this reduced feature set is illustrated in Fig. 11. Video replay attack detection gives best performance while the photograph mask attack detection ranks second in performance followed by hand-held photograph attack detection using the collinearity feature. At 10% FPR, TPR of 70, 78 and 100% are achieved for photograph, mask and video replay attacks, respectively. The colocation feature performance is much weaker compared to the collinearity performance. At 10% FPR about 70, 38 and 65% TPR are achieved for photograph, mask and video replay detection, respectively. Figure 12 shows the ROC curves for the reduced feature sets for fusion of collinearity and colocation information. The performance of the system was found to be worse when collinearity or colocation features were used separately for most scenarios as can be seen in comparison with Fig. 11. At 10% FPR, video replay performance is 100% and photograph attack TPR increased to about 90%. The mask attack detection performance marginally decreased after fusion and is lower compared to the video and photograph spoof detection performance.

Table 3 summarizes some of the key results from Fig. 11 and Fig. 12 and presents results for each feature type separately along with the results for the combined

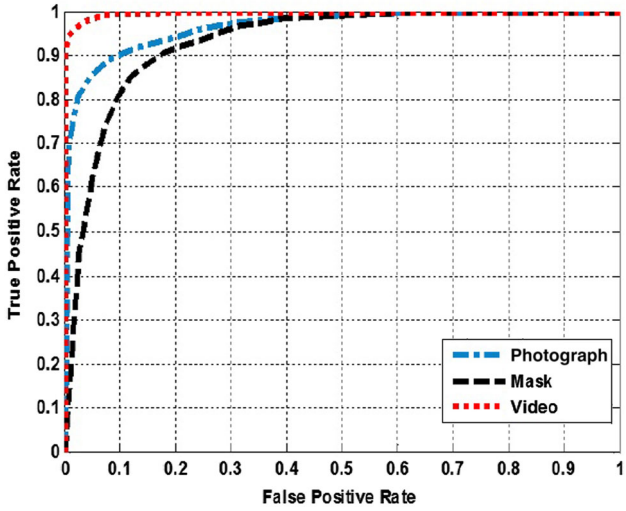

Fig. 12 Score fusion performance using reduced feature sets

Table 3 TPR at FPR = 0.10 using reduced feature sets

Feature sets	Photograph	Mask	Video replay
Collinearity	0.70	0.78	1.00
Colocation	0.70	0.38	0.65
Collinearity and colocation	0.90	0.81	1.00

collinearity and colocation features using the reduced feature sets. The video replay attack detection rate using the colocation feature has decreased when using the reduced feature set. However, this is expected as the video attack data were not used for establishing the optimum feature set. For video replay attack detection, the proposed combined system is error-free (for this data set). Using the reduced features, the TPRs of combined collinearity and colocation features increased by 32, 12 and 1% for photograph, mask and video replay attack detection, respectively, when compared to using the entire feature set.

In the following experiments, all attack types were treated as one class (fake) rather than as three separate attack scenarios. Figure 13 illustrates the ROC curves for real and fake attempts. Combination of collinearity and colocation data again gave better performance. The colocation feature performance is much weaker compared to the performance of collinearity and fusion-based schemes. The performance of collinearity features is very close to that achieved by the performance of the combined features. At 10% FPR, TPR of about 87, 63 and 91%, were achieved for collinearity, colocation and their fusion, respectively.

Table 4 shows a comparison of our experimental results (GS for gaze stability) with the performances reported for similar photograph spoofing attacks published in the literature. Although the results are based on different databases, they indicate the relative promise of the proposed methods. Given the novel nature of the challenge–response system used in this work, it has not been possible to make direct comparison with other algorithms which use different approach to liveness. The performance of our proposed approaches can be seen to compare favourably with the other methods considered and results lend support to its potential applicability in detecting spoofing attacks.

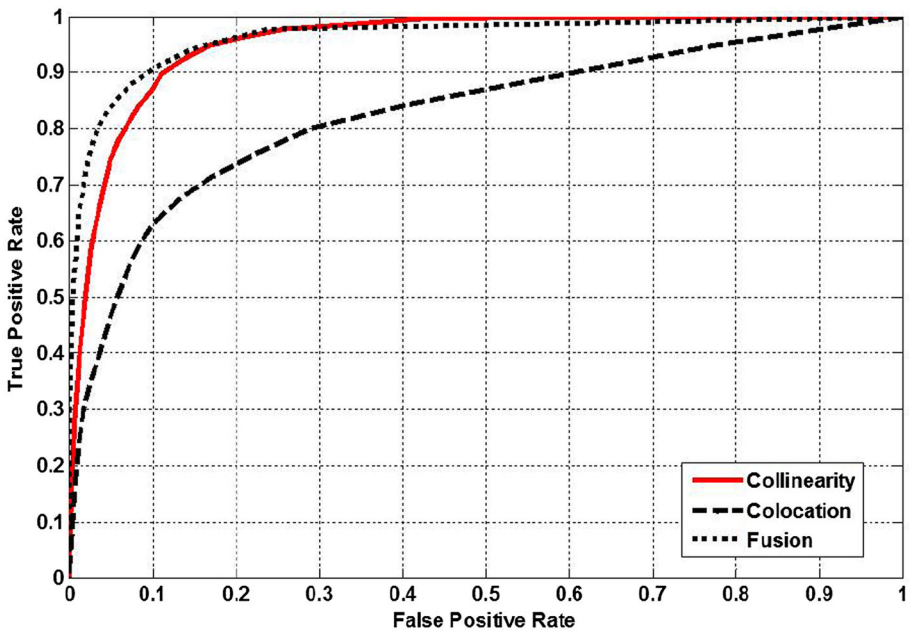

Fig. 13 Genuine versus fake (photograph, mask, video) performance using reduced feature sets

Table 4 Comparison of performance reports

Method	FPR	FNR
Kollreider et al. [29]	0.02	0.19
Tan et al. [29]	0.09	0.18
Peixoto et al. [30]	0.07	0.07
IGD [31]	0.17	0.01
MaskDown [31]	0.00	0.05
GS photograph attack	0.05	0.13
GS all attack	0.05	0.16

5 Conclusion

The work presented here explores the notion of gaze stability and features based on it for the task of detecting presentation attacks which is one of the major challenges facing the use of biometric systems. An active challenge–response approach is adopted using a visual stimulus to direct the gaze, and the system provides gaze stability measures to discriminate between genuine and fake attempts. Two gaze-based features, collinearity and colocation, have been introduced and extensively evaluated. Three attack scenarios were investigated, and data were collected to evaluate the performance of the proposed system using different combinations of features and attack modalities. Feature selection together with a multi-classifier approach, combining information from separate feature sets using score fusion, provided the best results showing the potential effectiveness and viability of this approach. In case of photograph and mask attacks, there may be a possibility to circumvent the system depending on the ability of the attacker to manipulate the artefact in way similar to natural head/eye movements. The potential for this type of "skilled" attack will be considered in future work.

Appendix: Collinearity and colocation feature extraction

The collinearity feature provided in Sect. 3.C, derived for horizontal stimulus loci, may be generalized to include any linear trajectory. Let S_l be a collinear subset of C, where the stimuli are linear. $S_l \subseteq C$, $l = 1, \ldots, L$ where L is the number of linear sets of stimulus locations. For $(x, y) \in S_l, y = a_{l1}x + a_{l0}$ where a_{l1} is constant.

Let R be the set of landmark locations in the captured images.

For each S_l there is a corresponding subset in R. Let this be denoted by T_{lk}

$$T_{lk} \subseteq R, \quad l = 1, \ldots, L \tag{14}$$

for any given facial landmark k, and let $v_{ik} = f(u_{ik})$ denote the trajectory of the landmark in response to the challenge. Since the trajectory of the challenge S_l is linear, a linear response can be assumed and this can be approximated by the equation of a line

$$\hat{v}_k = b_{k1}u_k + b_{k0} \quad \text{where} \quad b_{k1}, b_{k0} \text{ are constants.} \tag{15}$$

b_{k1} should be the same as a_{l1} (the slope of the challenge trajectory), whereas b_{k0} depends on the system set-up, user interaction, etc.

Let, e_{lk} denote the deviation between the estimated $\hat{v}_{ik}$ and observed v_{ik} (see Fig. 14), i.e.

$$e_{ik} = v_{ik} - \hat{v}_{ik} \tag{16}$$

So, the mean square error (MSE) for T_{lk} will be

$$E_{lk} = \frac{1}{N}\sum_i e_{ik}^2 = \frac{1}{N}\sum_i (v_{ik} - \hat{v}_{ik})^2 \tag{17}$$

where N is the cardinality of T_{lk}.

By substituting Eq. (15) in Eq. (17) and replacing with b_{k1} with a_{l1}

$$\begin{aligned} E_{lk} &= \frac{1}{N}\sum_i (v_{ik} - (a_{l1}u_{ik} + b_{k0}))^2 \\ &= \frac{1}{N}\left(\sum_i (v_{ik}^2 + a_{l1}^2\sum_i u_{ik}^2 + b_{k0}^2 N \right. \\ &\quad \left. - 2a_{l1}\sum_i v_{ik}u_{ik} + 2a_{l1}b_{k0}\sum_i u_{ik} - 2b_{k0}\sum_i v_{ik}\right) \end{aligned} \tag{18}$$

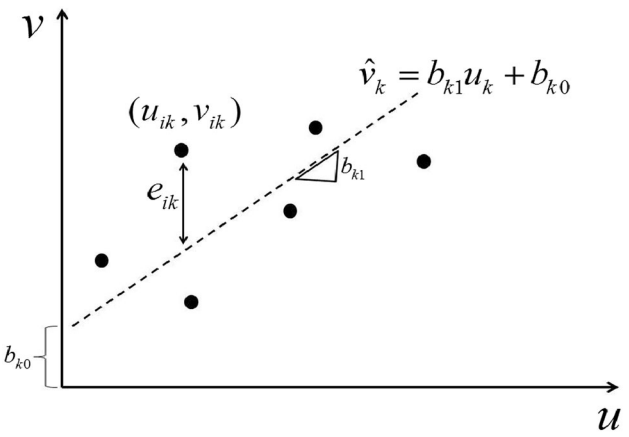

Fig. 14 Observed locations (*bullet symbols*) and expected locus of the landmark positions (*dash symbols*)

Here, b_{k0} should be such chosen that E_{lk} is a minimum. Hence,

$$\frac{\partial E_{lk}}{\partial b_{k0}} = 0$$
$$\Rightarrow 2b_{k0} + \frac{2a_{l1}}{N}\sum_i u_{ik} - \frac{2}{N}\sum_i v_{ik} = 0 \quad (19)$$
$$\Rightarrow b_{k0} = \frac{\sum_i v_{ik} - a_{l1}\sum_i u_{ik}}{N}$$

Equation 19 can be used to calculate E_{lk}. As there can be multiple face landmarks as well as several distinct linear challenge trajectories, a feature vector F_{colin} can be constructed from the concatenation of these values and used for liveness detection.

$$F_{\text{colin}} = [E_{11}, E_{12}, \ldots, E_{1K}, E_{21}, \ldots, E_{lk}, \ldots, E_{LK}] \quad (20)$$

The colocation features are extracted from the images acquired when the stimulus is presented at a given location several times. This stimulus can be considered as a special case of collinear trajectory where the line is reduced to a single point.

References

1. Tronci R, Muntoni D, Fadda G, Pili M, Sirena N, Murgia G, Ristori M, Roli F (2011) Fusion of multiple clues for photo-attack detection in face recognition systems. In: 2011 International joint conference on biometrics (IJCB), pp 1–6
2. Ali A, Deravi F, Hoque S (2012) Liveness detection using gaze collinearity. In: 2012 Third international conference on emerging security technologies (EST), pp 62–65
3. Ali A, Deravi F, Hoque S (2013) Spoofing attempt detection using gaze Colocation. In: 2013 International conference of the biometrics special interest group (BIOSIG), pp 1–12
4. Pan G, Sun L, Wu Z, Lao S (2007) Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: IEEE 11th International conference on in computer vision 2007. ICCV 2007, pp 1–8
5. Jee H-K, Jung S-U, Yoo J-H (2006) Liveness detection for embedded face recognition system. Int J Biomed Sci 1(4):235–238
6. Wang L, Ding X, Fang C (2009) Face live detection method based on physiological motion analysis. Tsinghua Sci Technol 14(6):685–690
7. Kollreider K, Fronthaler H, Bigun J (2009) Non-intrusive liveness detection by face images. Image Vis Comput 27(3):233–244
8. Kollreider K, Fronthaler H, Bigun J (2008) Verifying liveness by multiple experts in face biometrics. In: IEEE computer society conference on computer vision and pattern recognition workshops, 2008. CVPRW '08, pp 1–6
9. Kollreider K, Fronthaler H, Bigun J (2005) Evaluating liveness by face images and the structure tensor. In: Fourth IEEE workshop on automatic identification advanced technologies, pp 75–80
10. Li J, Wang Y, Tan T, Jain AK (2004) Live face detection based on the analysis of fourier spectra. In: Defense and Security. International society for optics and photonics, pp 296–303
11. Kim G, Eum S, Suhr JK, Kim DI, Park KR, Kim J (2012) Face liveness detection based on texture and frequency analyses. In: 2012 5th IAPR international conference on biometrics (ICB), pp 67–72
12. Komulainen J, Hadid A, Pietikäinen M (2013) Face spoofing detection using dynamic texture. In: 2012 Workshops computer vision-ACCV. Springer, pp 146–157
13. Komulainen J, Hadid A, Pietikainen M, Anjos A, Marcel S (2013)Complementary countermeasures for detecting scenic face spoofing attacks. In: 2013 International conference on biometrics (ICB), pp 1–7
14. Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intell 97(1):245–271
15. Nunez AS (2009) A physical model of human skin and its application for search and rescue. DTIC Document, Technical report
16. Ryer DM, Bihl TJ, Bauer KW, Rogers SK (2012) Quest hierarchy for hyperspectral face recognition. Adv Artif Intell 2012:1
17. da Silva Pinto A, Pedrini H, Schwartz W, Rocha A (2012) Video-based face spoofing detection through visual rhythm analysis. In: 2012 25th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), pp 221–228
18. Frischholz R, Werner A (2003) Avoiding replay-attacks in a face recognition system using head-pose estimation. In: IEEE international workshop on analysis and modeling of faces and gestures, 2003. AMFG 2003, pp 234–235
19. Sharma DCKN (2013) Fake face detection based on skin elasticity. Int J Adv Res Comput Sci Softw Eng 3(5):1048–1051
20. Pelz J, Hayhoe M, Loeber R (2001) The coordination of eye, head, and hand movements in a natural task. Exp Brain Res 139(3):266–277
21. Volkmann FC (1986) Human visual suppression. Vis Res 26(9):1401–1416
22. Land MF (2006) Eye movements and the control of actions in everyday life. Prog. Retin Eye Res 25(3):296–324
23. Milborrow S, Nicolls F (2008) Locating facial features with an extended active shape model. In: Computer Vision-ECCV. Springer, pp 504–513
24. Jain A, Nandakumar K, Ross A (2005) Score normalization in multimodal biometric systems. Pattern Recogn 38(12):2270–2285
25. Singh AK, Joshi P, Nandi G (2014) Face recognition with liveness detection using eye and mouth movement. In: 2014 International conference on signal propagation and computer technology (ICSPCT). IEEE, pp 592–597
26. Chingovska I, Anjos A, Marcel S (2012) On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG—Proceedings of the international conference of the biometrics special interest group (BIOSIG), pp 1–7
27. Ross A, Jain A (2003) Information fusion in biometrics. Pattern Recogn Lett 24(13):2115–2125
28. Fawcett T (2004) Roc graphs: notes and practical considerations for researchers. Mach Learn 31:1–38
29. Kollreider K, Fronthaler H, Faraj M, Bigun J (2007) Real-time face detection and motion analysis with application in liveness assessment. IEEE Trans Inf Forens Secur 2(3):548–558
30. Peixoto B, Michelassi C, Rocha A (2011) Face liveness detection under bad illumination conditions. In: 2011 18th IEEE international conference on image processing (ICIP), pp 3557–3560
31. Chingovska I, Yang J, Lei Z, Yi D, Li SZ, Kahm O, Glaser C, Darner N, Kuijper A, Nouak A et al (2013) The 2nd competition on counter measures to 2D face spoofing attacks. In: ICB, pp 1–6