Abstract

The face recognition community has finally started paying more attention to the long-neglected problem of spoofing attacks and the number of countermeasures is gradually increasing. Fairly good results have been reported on the publicly available databases but it is reasonable to assume that there exists no superior anti-spoofing technique due to the varying nature of attack scenarios and acquisition conditions. Therefore, we propose to approach the problem of face spoofing as a set of attack-specific subproblems that are solvable with a proper combination of complementary countermeasures. Inspired by how we humans can perform reliable spoofing detection only based on the available scene and context information, this work provides the first investigation in research literature that attempts to detect the presence of spoofing medium in the observed scene. We experiment with two publicly available databases consisting of several fake face attacks of different nature under varying conditions and imaging qualities. The experiments show excellent results beyond the state of the art. More importantly, our cross-database evaluation depicts that the proposed approach has promising generalization capabilities.

1. Introduction

Because of its natural and non-intrusive interaction, identity verification and recognition using facial information is among the most active and challenging areas in computer vision research. Despite the remarkable progress in the face recognition technology in the recent decades, wide range of viewpoints, ageing of subjects and complex outdoor lighting are still research challenges. While there is a significant number of works addressing these issues, research on face biometric systems under spoofing attacks has mostly been overlooked although face recognition systems are known, since long time ago, to respond weakly to direct attacks.

A spoofing attack occurs when a person tries to masquerade as someone else by falsifying data and thereby gaining illegitimate access. Because of the urgent need for enhancing the security and robustness of face biometrics, a variety of spoofing detection schemes have been proposed to tackle the problem of spoofing attacks. Furthermore, the recently organized IJCB 2011 competition on countermeasures to 2D facial spoofing attacks [4] was an important collective effort for finding effective practices for non-intrusive spoofing detection. While challenge-response approach [7, 13, 16], multi-modal analysis [5, 12, 16] and multi-spectral imaging [22, 24, 29] provide efficient means for discriminating real faces from fake ones, they are also rather impractical due to interaction or unconventional imaging requirements. Therefore, it would be rather appealing to use anti-spoofing techniques requiring no user-cooperation and using conventional imaging systems within the existing face authentication systems. Another advantage is that from challenge-response based countermeasures it is rather easy to deduce which liveness clues need to be fooled. For instance, the request for uttering words suggests that analysis of synchronized lip movement and lip reading is utilized, whereas rotating head in a certain direction reveals that the 3D geometry of the head is measured. For non-intrusive approaches, it is usually not known which countermeasures are used, thus the system might be harder to deceive [20].

Typical non-intrusive 2D face anti-spoofing technique is liveness detection that aims at detecting physiological signs of life, such as eye blinking [21], facial expression changes [16, 17] and mouth movements [5, 16]. In general, motion analysis is a commonly used countermeasure since it can be assumed that the movement of planar objects, e.g. video displays and photographs, and rigid 3D face masks differs significantly from that of real human faces which are complex non-rigid 3D objects [3, 7, 15, 27]. If a background scene is incorporated with the face spoof (scenic fake face), overall motion correlation between the face and the background regions is likely to be observed [1, 27]. Even though motion is an important visual cue, vitality and non-rigid mo-
tion detectors are powerless under video-replay attacks and the lack of motion may lead to high number of authentication failures if interaction is not employed.

Another category of anti-spoofing methods are based on the analysis of skin properties such as reflectance and texture. One can assume that photographs are usually smaller in size or would contain fewer high frequency components compared to real faces, thus countermeasures based on analysing the high frequency content have been proposed [18, 25]. Such approaches may work well for down-sampled photos but are likely to fail for higher-quality face spoofs. Alternatively, it is likely that real faces and fake ones present different texture patterns because of facial texture quality degradation due to recapturing process and differences in surface and reflectance properties. Therefore, micro-texture analysis has been utilized for capturing these differences [2, 19]. Recently, texture analysis based spoofing detection has been extended into spatiotemporal domain [17, 23]. In addition to analysing the structure of facial micro-textures, spatiotemporal texture analysis is applied for describing specific dynamic events, e.g. facial motion and sudden characteristic reflections of planar spoofing media [17] and dynamic video noise signatures of display devices [23]. The major drawback of texture analysis is that rather high resolution input images are required in order to extract the fine details needed for discriminating genuine faces from spoofing media. While lower imaging quality might be enough for detecting the most crude attack attempts, such as small mobile phone displays and prints with strong artefacts, the grid structure of a display device or facial pores can be captured only in high-definition close-up images. On the other hand, also high false rejection rate might be an issue if acquisition quality is not good enough. Furthermore, the nature of texture patterns varies a lot due to different acquisition conditions and spoofing media, thus diverse datasets are needed for training the micro-texture based methods, especially at conventional webcam image quality.

Indeed, many directions for non-intrusive spoofing detection have been already explored and impressive results have been reported on individual databases. However, the varying nature of spoofing attacks and acquisition conditions makes it impossible to predict how single anti-spoofing techniques, e.g. facial texture analysis, can generalize the problem in real-world applications. Moreover, we cannot foresee all possible attack scenarios and cover them in databases because the imagination of the human mind always finds out new tricks to fool existing systems. It is reasonable to assume that no single superior technique is able to detect all known, let alone unseen, spoofing attacks. Therefore, the problem of spoofing attacks should be broken down into attack-specific subproblems that are solvable if a proper combination of complementary countermeasures is used. In this manner, a network of attack-specific spoofing detectors could be used to construct a flexible anti-spoofing framework in which new techniques can be easily integrated to patch the existing vulnerabilities in no time when new countermeasures appear.

Scene information has been neglected in the previous works, i.e. the existing countermeasures are mainly based on different facial texture and motion analysis techniques. The background information has been exploited only for measuring the overall motion correlation between face and background [1, 27] or for checking if the background scene of a stationary face recognition system suddenly changes due to (scenic) fake face attack [20]. Face images captured from face spoofs may visually look very similar to the images captured from live faces, thus face spoofing detection is rather difficult to perform based on only single face image or a relatively short video sequence. Furthermore, countermeasures based on facial motion analysis can be easily deceived using video recordings or animated faces that have become very realistic nowadays. Depending on the imaging and fake face quality, even for us humans it is nearly impossible to tell the difference between a genuine face and a fake one without any scene information or any unnatural motion or facial texture patterns. However, we can immediately notice if there is something suspicious going on in the view, e.g. if someone is holding a video display or a photograph in front of the camera.

In this paper, we follow the aforementioned principle of attack-specific spoofing detection and tackle face spoofing scenarios in which scene information can be exploited. In other words, we are trying to detect whether someone is presenting a fake face on a spoofing medium to the camera in the provided view. As we humans rely mainly on scene and context information when performing spoofing detection, our algorithm tries to mimic human behaviour and exploits scenic cues for determining whether spoofing medium is present in the observed scene. The proposed approach consists of a cascade of an upper-body (UB) and a spoofing medium (SM) detectors that are based on histogram of oriented gradients (HOG) descriptors [6] and linear support vector machines (SVM). The method can operate either on a single video frame or video sequences. Temporal processing of longer video sequences may naturally lead to better detection results but at the cost of more computational time which could be an issue in real-life applications.

To the best of our knowledge, this is the first work that attempts to detect the presence of display medium itself in the observed scene. We introduce below our novel approach and provide extensive experimental analysis on the publicly available CASIA Face Anti-Spoofing Database [28] consisting of several fake face attacks of different natures and under varying conditions and imaging qualities, showing excellent results beyond the state of the art. Further-
more, we perform a cross-database evaluation and show that the proposed approach is able to generalize the problem of face spoofing detection on NUAA Photograph Imposter Database [25] when trained and tuned solely on the CASIA Face Anti-Spoofing Database [28].

2. Proposed spoofing detection using scenic cues and local shape analysis

To follow the principle of attack-specific spoofing detection and to find out well-generalizing countermeasures to particular attack scenarios, we first need to define and categorize clearly the spoofing cases that we are dealing with. Face spoofing attacks can be categorized in several ways. For instance, an obvious way is based on the used display medium type, such as photograph, video screen or mask. We humans can easily detect spoofing attacks if we are able to notice something suspicious in the observed scene, e.g. if someone is holding a video display or a photograph in front of the camera. Therefore, we suggest a two-part categorization based on whether the whole spoofing medium is visible in the view or not. In both attack scenarios, common and, more importantly, their own distinctive visual cues can be exploited in spoofing detection schemes.

There are two ways for hiding the display medium outside the view, either by incorporating background scene in the face spoof and placing the resulting scenic face spoof very near to the sensor. Fortunately, the proximity between the spoofing medium and the camera might cause the recaptured face image to be out-of-focus and reveal also other facial texture quality issues if the imaging quality is good enough to capture the differences in fine details of surface properties between a human face and spoofing medium [2, 17]. Furthermore, for stationary systems, it should be possible to observe high correlation between the overall motion of the face and the background regions [1, 27] or check if the background scene suddenly changes [20].

In the other scenario, in which the whole spoofing medium is visible, the attacks is usually performed using a close-up fake face that describes mainly the facial area which is presented to the sensor. The main weakness with the tightly cropped face spoofs is that the boundaries of the spoofing medium, e.g. video screen frame or photograph edges, or the attacker’s hands are usually visible during the attack, thus can be detected in the scene. Moreover, if the fake face is not well aligned with the upper half of the torso of the imposter, a natural upper-body profile cannot be observed.

In this work, we propose to detect close-up fake faces by describing the aforementioned scenic cues with a cascade of two HOG descriptor based detectors. The alignment of the face and the upper half of the torso is analysed using an upper-body detector and the presence of the display medium is determined with a specific detector that is trained on actual face spoofing examples.

2.1. Upper-body detection

For determining whether the head-and-shoulder region is properly aligned, we considered the upper-body detector that is a component of the human pose estimation pipeline proposed in [8]. The detector is based on the deformable part-based hierarchical model improvement of [6] proposed in [10], thus HOG descriptors and latent SVM are used to form a sliding window detector followed by non-maximum suppression. The publicly available source code¹ of the upper-body detector contains a model trained on the near-frontal upper-bodies collected within the work of [11].

As no actual fake faces are introduced in the training phase, upper-body detector cannot be considered directly as a spoofing detector. However, it can still be useful in anti-spoofing solutions if utilized within proper context. We apply it as a preprocessing stage for filtering out the most crude attack attempts in which the spoofing medium is poorly positioned or there are strong discontinuities between the face-and-shoulder region, e.g. thick screen frame of a video display.

2.2. Spoofing medium detection

The boundaries of the used spoofing medium, e.g. video screen frame or photograph edges, can be usually easily ob-

¹http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/
served as very distinctive discontinuities around the face. Thus, the target face does not blend in the background scene as well as a genuine face and looks also a bit disconnected of the upper-body. Fig 1 presents example images of genuine face and two close-up fake faces, a photo and a video attack, and the corresponding HOG descriptors computed around the detected face. As it can be seen, the local shape features can capture well the continuous edges of the used display medium that form a closed frame around the face, especially in the case of a video attack. Therefore, we chose to use the HOG descriptors also for performing more thorough scene abnormality analysis as it seems to provide a discriminative feature space for refining the upper-body description and for performing the actual spoofing detection.

Our proposed method for spoofing medium detection computes the HOG descriptors from a bounding box that contains roughly the upper-body area and is symmetrically expanded also above the detected face. As Fig 1 depicts, all aforementioned scenic cues can be seen inside the detection window. Once the features are computed, we use a linear SVM classifier for determining whether the presence of spoofing medium or other abnormalities, such as attacker’s hands, is detected in the view.

2.3. Proposed spoofing detection pipeline

The block diagram of the proposed spoofing detection scheme is shown in Fig 2. Our cascade structure of detectors first tries to find out whether there is a valid upper-body, i.e. a face with a properly aligned upper half of the torso, in front of the camera. Unlike Eichner et al. [8], we do not use the detected face location for regressing secondary upper-body detections but for validating whether it is within the bounding box of the detected upper-body in order to discard spurious upper-body detections. The purpose of the first decision stage is to ignore the most crude attack attempts and decide whether there is a need to perform more detailed scene analysis. Therefore, the operating point of the upper-body detection should be fixed to a very low false rejection rate. If an upper-body is found in the scene, the next step is apply the actual spoofing detector that complements the upper-body detector using a more specific upper-body model trained for discriminating genuine faces from close-up fake faces.

The proposed method can operate either on a single video frame or video sequences. If several video frames are utilized, e.g. in order to achieve more reliable detection performance, majority voting is first performed on the output of the first decision stage. If there are enough valid upper-body detections, the final label for the captured sequence is determined by averaging the scores of the second decision stage.

3. Experimental analysis

To assess of the effectiveness of our proposed anti-spoofing technique, we performed a set of experiments on the CASIA Face Anti-Spoofing Database\(^2\). We also applied cross-database testing to see how well our algorithm is able to generalize the problem of face spoofing detection on NUA Photograph Imposter Database\(^3\) when trained and tuned solely on the CASIA Face Anti-Spoofing Database. On both datasets, the face locations are retrieved using Modified Census Transform (MCT) based face detector [14].

In our experiments, we compared the spoofing detection performance of the upper-body and spoofing medium detectors separately and jointly using the proposed cascade structure. In other words, the detection threshold of the upper-body detector varies when its performance is evaluated. When combined with the spoofing medium detector in the proposed cascade structure, the operating point is fixed to 0% false rejection rate (FRR) using the training set of the CASIA Face Anti-Spoofing Database. The spoofing medium detection window is computed based on the detected face location. More specifically, the height and width of the face bounding box are expanded to 4\(d\) and 3.5\(d\) where \(d\) represents the side length of the face bounding rect-

---

\(^2\)http://www.cbsr.ia.ac.cn/english/FaceAntiSpoofDatabases.asp
\(^3\)http://parnec.nuaa.edu.cn/xtan/data/NuaaImposterdb.html
angle (see Fig. 1). The detection window is then normalized into $80 \times 70$ pixels and mirroring and rotation ($\pm 3$ degrees) was applied in order to generate more training samples like in [11]. The HOG descriptions are computed with VLFeat [26] using nine orientations and cell size of ten pixels. The final classification is performed using a linear SVM implementation of LIBLINEAR [9].

### 3.1. Evaluation on the CASIA Face Anti-Spoofing Database

We started by conducting experiments on the CASIA Face Anti-Spoofing Database [28] and comparing our results against those which are provided along with the database and have been recently published in [17]. The dataset includes significant improvements compared to previous databases, since it provides more variations in the collected data. The data set contains 50 real clients and the corresponding fake faces are captured with high quality from the original ones. The variety is achieved by introducing three imaging qualities (low, normal and high) and three fake face attacks which include warped photo, cut photo (eyeblink) and video attacks. Altogether the database consists of 600 video clips and the subjects are divided into subsets for training and testing (240 and 360, respectively). Results of a baseline system are also provided along the database for comparison. The baseline system is inspired by the work of Tan et al. [25], thus considers the high frequency information in the facial region using multiple Difference of Gaussian (DoG) features and SVM classifier.

Since the main purpose of the database is to investigate the possible effects of different fake face types and imaging qualities, the test protocol consists of seven scenarios in which particular train and test samples are to be used. The quality test considers the three imaging qualities separately, low (1), normal (2) and high quality (3), and evaluates the overall spoofing detection performance under variety of attacks at the given imaging quality. Similarly, the fake face test assesses how robust the anti-spoofing technique is to specific fake face attacks, warped photo (4), cut photo (5) and video attacks (6), regardless of the imaging quality. In the overall test (7), all data is used to give a more general evaluation. The results of each scenario are reported as Detection-Error Trade-off (DET) curves and equal error rates (EER), which is the point where false acceptance rate (FAR) equals false rejection rate (FRR) on the DET curve.

The highest imaging quality scenario differs from the others as its purpose is to capture the appearance of the target faces at high precision. Thus, it consists of close-up videos of faces when the scenic cues are usually missing in the view, i.e. the upper-body of the subject is not visible and the used display medium can be easily hidden outside the view. However, the measurement of facial texture quality seems to provide sufficient means to reveal whether degradation due to recapturing process is observed as the imaging quality is good enough to capture the differences in fine details of surface properties between a human face and spoofing medium [17]. On the other hand, the performance or generalization capabilities of texture based methods can be questioned at lower imaging qualities. Therefore, we consider only the access attempts at the two lowest imaging qualities for our experiments. As this dataset consists of video sequences, we follow the same principle that was applied for the provided DoG baseline [28] and process each access attempt by randomly selecting 30 video frames for tuning the upper-body detector and training the spoofing medium detector. Also the final label of each test video sequence is based on 30 randomly selected frames over which the upper-body votes and the average of spoofing detection scores are determined.

First, we conduct the fake face test without highest imaging quality in order to find out how well the proposed approach performs under different attack types consisting of...
### 3.2. Cross-database evaluation

In real-world applications, face anti-spoofing techniques must be working well under varying attack scenarios in different acquisition conditions. Therefore, we experiment now if our proposed approach is able to generalize its good performance beyond CASIA Face Anti-Spoofing Database.

For our cross-database evaluation, we considered the publicly available NUAA Photograph Imposter Database [25] that consists of images of both real client accesses and high-quality photo attacks which were recorded using conventional webcams at 20fps with resolution of $640 \times 480$ pixels. The face images of live humans and their photographs were collected in three sessions at intervals of about two weeks. In addition, during each session, the environmental and illumination conditions are changing.

The database is originally divided into separate training and test sets and the evaluation is based on analysing performance on still images. In our cross-database experiment, we maximize the amount of test samples by combining the two original subsets when also the number of real subjects increases significantly. Furthermore, we restore the original real client accesses and photo attacks by mapping the series of single images into corresponding video sequences. In this manner, we can follow the access-attempt (video) based evaluation of the CASIA Face Anti-Spoofing Database to test if the proposed approach is able to detect the three categories of photo-attacks simulated in the NUAA Photograph Imposter Database.

The proposed countermeasure was tuned and trained solely on the CASIA Face Anti-Spoofing Database, i.e. the operating point of 0% FRR was set for the upper-body detector and the spoofing medium detector was trained using warped and cut photo attacks of both train and test sets.

### Table 1. EER (%) comparison on CASIA Face Anti-Spoofing Database (*scenarios in the official test protocol).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>1*</th>
<th>2*</th>
<th>photo</th>
<th>video</th>
<th>both</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoG [28]</td>
<td>13.3</td>
<td>13.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LBP [17]</td>
<td>4.4</td>
<td>10.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LBP-TOP [17]</td>
<td>3.3</td>
<td>3.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UB</td>
<td>17.8</td>
<td>17.8</td>
<td>21.7</td>
<td>8.3</td>
<td>17.8</td>
</tr>
<tr>
<td>Medium</td>
<td>2.2</td>
<td>1.1</td>
<td>3.3</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td>UB+medium</td>
<td>2.2</td>
<td>1.1</td>
<td>3.3</td>
<td>0.0</td>
<td>3.3</td>
</tr>
</tbody>
</table>

The results of these experiments are presented in Fig 3 and Table 1. As it can be seen, the upper-body detection cannot ignore the different face spoofing attacks well enough as it can be deceived by aligning the fake face with the upper-body of the imposter which is neatly performed on the CASIA dataset (see Fig 5). However, the spoofing medium detector is able to capture the nature of both attack types well because it is particularly trained for finding abnormalities around the detected face. As shown in Section 2.2 and in Fig 1, the HOG features can easily describe the black screen frames of the video attacks, thus leading to perfect detection performance. Also the HOG feature based upper-body detector performs significantly better under video replay-attacks as the video screen frame causes strong discontinuities between face and upper-body. The performance of the cascade of both detectors is the same as for the spoofing medium detector alone, i.e. the upper-body detector fails rule out any spoofing attacks that our primary spoofing medium detector is not able to handle. On the other hand, the upper-body detector does not reject any real clients either, thus its usefulness is evaluated again in the cross-database evaluation.

In order to get comparable results with previous works, we followed also two official test protocols that do not contain any video sequences recorded at the highest imaging quality. The performance comparison for the two quality tests, low and normal, can be seen in Fig 4 and Table 1. The results indicate that the proposed cascade of upper-body and spoofing medium detectors is able to improve the state of the art under the two official test protocols, especially at the normal imaging quality as nearly perfect detection performance is achieved. Thus, the scene information is indeed very important visual cue for face spoofing detection and should be considered as one building block when constructing anti-spoofing solutions.
The results of the cross-database experiment are shown in Fig. 6 as DET curves. Again, the use of only spoofing medium detector leads to better results compared to the upper-body detector. More importantly, when the cascade of both detectors is considered, a significant performance enhancement is obtained at low FRR, thus confirming the benefits of cascading the both detectors. It is also worth mentioning that the performance gap between the upper-body and spoofing medium detectors (in terms of EER 12.8% and 6.8%, respectively) is decreased substantially.

The proposed approach managed the cross-database testing well as the excellent performance on the CASIA Face Anti-Spoofing Database decreases only from 3.3% to 6.8% in terms of EER. The upper-body detector performs better on the NUAA Photograph Imposter Database because the photos are not usually aligned with the upper-body of the imposter, unlike on the CASIA Face Anti-Spoofing Database. An example image of a false detection can be seen in Fig 5. The main reason for the performance drop of the spoofing medium detector is the more challenging acquisition conditions of the NUAA dataset, i.e. the target faces are more vivid and the background scene is not static (other people moving around) that cause two problems. First, the face localization data was not as accurate and stable, thus also the predicted window locations for detecting the presence of spoofing medium were more noisy. Secondly, the complex and vivid background scenes introduce new artefacts inside the spoofing medium detection window. Although the results of the cross-database evaluation were promising, there is room for improving the appearance description and segmentation of the spoofing medium in order to solve the aforementioned problems.

4. Conclusion

It is reasonable to assume that there exists no generic spoofing detection scheme that is able to perform robustly in all known, let alone unseen, attack scenarios. Thus the problem of face anti-spoofing should be divided into attack-specific subproblems that are solvable if a proper combination of complementary countermeasures is found. In this work, we addressed this issue by proposing countermeasure to close-up fake face attacks. More specifically, we considered scene information for detecting whether someone is presenting a fake face on a spoofing medium to the camera because it is one of the few visual cues that we humans can rely on.

The proposed approach consists of a cascade of an upper-body and a spoofing medium detector that are based on histogram of oriented gradients descriptors and linear support vector machines. Our attack-specific countermeasure obtained excellent results under various fake face attacks, especially under video replay attacks. Furthermore, the generalization capability of the method tested using cross-database evaluation showed very promising results.

It is worth mentioning that a genuine output label of the proposed cascade of detectors does not rule out the possibility of a spoofing attack, since it is designed for detecting specific attack scenarios. Thus, the proposed cascade structure could be used for triggering other spoofing detection schemes if it is used as a part of a larger anti-spoofing solution. For instance, face and background motion correlation and facial texture quality measurements could be placed after our close-up face spoof detector in order to detect scenic face spoofs. As future work, the spoofing medium detector could be improved by applying custom segmentation algorithms and scene motion analysis in order to overcome the noise due to inaccurate face detection and the size limitation for the bounding box.

5. Acknowledgments

The research leading to these results has received funding from the European Community’s FP7 under grant agreements 257289 (TABULA RASA). The financial support of Academy of Finland and Infotech Oulu Doctoral Program is also gratefully acknowledged.

References


