Fake Face Image Detection Using Deep Learning-Based Local and Global Matching

Margarita Favorskaya and Anton Yakimchuk

Reshetnev Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy ave., Krasnoyarsk, 660037 Russia

Abstract

The widespread adoption of face recognition systems in practice has provoked multiple attempts to fail these systems in order to impersonate another person. The range of such fake attacks is wide, and methods which can be used to compensate for one type of attacks are not adapted against other attacks. In this study, we propose a method for detecting fake face images based on local and global matching provided by deep neural networks. Also we do not discard the background analysis as a pre-processing stage. The idea is to assess the depth of the face in a still image as one of the main features of liveliness, which is not an easy task. The proposed method is directed against presentation attacks and attacks of adversarial perturbations. The experiments were conducted with and without deep neural networks. The use of deep learning increased the true accept rate and significantly reduced the error values.

Keywords

Fake face detection, presentation attacks, attacks of adversarial perturbations, deep learning, local matching, global matchin

1. Introduction

Face recognition is one of the most famous biometric methods of identity authentication, which is widely used in the field of security of organizations and enterprises, safety in public places such as airport terminals, train stations, stadiums, outdoor surveillance, etc. Research in this area began in the 1990s with the traditional machine learning methods (principal component analysis, Bayesian classification and metric models), methods for detecting local features (Gabor filters and Local Binary Patterns (LBPs)) and methods for detecting generalized features and advanced to deep learning techniques. Currently, the accuracy of deep learning-based face recognition has achieved 99.80%. At the same time, it is believed that human vision shows an accuracy of 97.53% [1].

Since it is quite easy to replace a face image or present a short video impersonating another person, face recognition systems must include a fake face detection module. This fake face detection module is usually introduced after the face detection and alignment module, but before the visual processing module and the recognition module. It worth noting that fake face detection and face recognition have different target functions. Detection of forgery is associated with the search for artifacts of the "liveliness" of the face. Therefore, lighting, shadows, glare, scene depth, etc. are of great importance. At the same time, face recognition involves minimizing the listed above artifacts and extracting features that are invariant to lighting, posture, emotions, overlapping objects, etc. The aim of our study is to develop a method for detecting fake faces using a single photograph. Our objective is to develop an approach which takes into account the background analysis of an image and extraction of pseudo-depth parameters from a single photograph using local and global matching provided by deep neural networks. Of course, accurate depth parameters can be estimated with additional expensive

SibDATA 2021: The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021, June 25, 2021, Krasnoyarsk, Russia
EMAIL: favorskaya@sibsau.ru (M. Favorskaya); yakimchuk_aa@sibsau.ru (A. Yakimchuk)
ORCID: 0000-0002-2181-0454 (M. Favorskaya); 0000-0002-6654-9122 (A. Yakimchuk)

devices requiring a fusion of visual, thermal and/or depth information. Our method aims at applying algorithmic solutions to complex cases such as fake face recognition.

The structure of the paper is the following. A short literature review is given in Section 2. Section 3 describes the proposed method for detecting fake faces in the images based on local and global matching. The results of the conducted experiments are discussed in Section 4. Section 5 concludes the paper.

2. Related work

Currently, there are two types of widespread attacks in face recognition systems, referred to as presentation attacks or spoofing attacks and attacks of adversarial perturbations [2]. Presentation attacks include presenting fake printed images, smartphone images or short video sequences to a facial recognition camera or disguising a person using cosmetics, makeup or a 3D mask. Masking is the most complicated case for recognizing presentation attacks. Attacks with the 3D mask are nearly impossible to identify without additional modalities. Since the 2010s, most countermeasures for presentation attacks have relied on deep neural networks (earlier, features were manually extracted). Thus, Yang et al. [3] trained a convolutional neural network (CNN) ImageNet to distinguish fake faces from genuine ones using both one frame and five scaled frames. This algorithm required preliminary image alignment using biomarkers. Binary classification (spoof/genuine) was performed on the CNN output using a support vector machine (SVM). In [4], a two-stream CNN was proposed, where one stream analyzed local fragments of the face, assigning spoofing estimates, and another stream was trained to estimate the depth of the scene using 3D samples. Li et al. [5] proposed CNN with a more complex architecture called deep part features from CNN. The features partially extracted by the first VGG (Visual Geometry Group) CNN were applied to the second fine-tuned VGG CNN for classification. An original way to decompose an image into a genuine face and spoofing noise using CNN was proposed in [6]. In this work, the classification of genuine images was implemented using noise.

The analysis of video sequences provides better detection of fake face images since in this case, artifacts of the “liveliness” of the face are available, for example, blinking [7], simple movements of the head, and so on. Note that CNN with the LSTM layers are traditionally utilized for the analysis of spatio-temporal structures. Such an architecture is applied to recognize genuine video sequences in [8]. Some research is aimed at detecting 3D masks [9-10].

Adversarial perturbation attacks are based on deep learning models, and, therefore, have appeared relatively recently. Adversarial perturbation is reduced to a slight distortion of the input image, such as brightness, in such a way that this perturbation is not identified by human vision, but leads to the fact that the deep network gives an incorrect classification. Goswami et al. [11] suggested detecting such masked attacks by analyzing the responses of filters in hidden layers and eliminating the most problematic filters. The SmartBox software tool for testing the performance of algorithms for detecting and mitigating adversarial attacks in face recognition systems is presented in [12]. The SmartBox software tool supports several algorithms, for example, DeepFool, Elastic-Net and utilities against gradient attacks and L2 attacks. Despite some success in confronting this type of attacks, adversarial perturbation attacks are constantly becoming more complex and they require further improvement of the algorithms. Other, more specific types of attacks can be noted, namely, stealing deep templates of faces for the purpose of manipulation by third persons. The deconvolutional neural network NbNet was proposed to confront such attacks [13]. The matter is that digital manipulation attacks using generative adversarial networks can generate fully or partially modified photorealistic facial images by altering an emotional expression, manipulating attributes or completely synthesizing a face. Thus, adversarial perturbation attacks are directed against deep neural networks which have proved to be good in the face recognition problem. The necessity to protect deep neural networks and deep patterns remains a major challenge in face recognition systems.

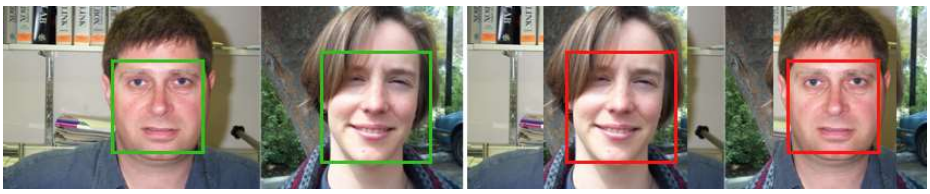
3. The proposed method

The proposed method is based on several verifications due to the fact that the impact of different attacks leads to different consequences. The method is based on two stages of the face image entering the input of the recognition system. Note that the task of verifying the genuineness of a face image is more difficult than using a short video.

The background analysis and local and global matching are described in Sections 3.1-3.2, respectively.

3.1. Background analysis

Background analysis is required to assess the correspondence of the global brightness and color parameters of a face image to the entire scene or their divergence from it. It is difficult to cut out the face image without the background in a photograph. Looking for another background for the face is a good reason to conduct a more detailed analysis for genuineness. For this, a sufficiently large fragment of the scene is segmented, where the face image occupies no more than 25-30%. The assumption is based on the fact that while it is quite simple to change the parameters of the face image, it is difficult to change the parameters of the scene background, taking into account the geometric binding of the camera, unknown to the attacker. Figure 1 depicts examples of capturing faces in the background of the scene. In Figure 1b, the background near the face does not match the background of the scene.



a b

Figure 1: Capturing face images considering the background of the scene: a) without artifacts, b) with artifacts

CCTV cameras are usually installed stationary. Therefore, for constructing the scene background model, we can use the Gaussian mixture model (GMM) with its adaptation to changes in lighting and shadows, as well as to temporal/seasonal/meteorological characteristics [14].

In the GMM model, the pixel intensity is determined by a mixture of K Gaussian distributions, where K is a small number. Each Gaussian distribution is associated with its own weight. The GMM parameters are updated recursively with every incoming sample. The pixel probability $P(X_t)$ is estimated by Eq. 1, where X_t is the pixel value at time t, K is the number of the Gaussian distributions taken into account, $w_{j,t}$ is the weight value, $\mu_{j,t}$ is the mean value, $\sum_{j,t}$ is the covariance matrix of the jth Gaussian at time t, η is the Gaussian probability density function (PDF).

$$P(X_t) = \sum_{j=1}^{K} w_{j,t} * \eta\left(X_t, \mu_{j,t}, \sum_{j,t}\right) \quad (1)$$

The probability density function η is defined by Eq. 2, where n is the dimensionality of X_t.

$$\eta\left(X_t, \mu_{j,t}, \sum_{j,t}\right) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\sum_{j,t}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(X_t - \mu_{j,t}\right)^T \left(\sum_{j,t}\right)^{-1} \left(X_t - \mu_{j,t}\right)} \quad (2)$$

For simplicity, the covariance matrix $\sum_{j,t}$ is defined as $\sigma_{j,t}^2 \mathbf{I}$ for the jth component, where $\mathbf{I}$ is the identity matrix under assumption that the X_t components (red, green and blue) are independent and have the same deviations.

The background distributions have a higher probability and lower standard deviations because the background colors remain the same for longer time than the foreground objects. This observation makes the GMM model updated when an incoming pixel is checked against the existing GMM components. If the pixel value is within 2.5 of the standard deviation of some weighted Gaussian distribution, then the distribution is updated. Otherwise, the distribution with the minimum weight is replaced by a new distribution with high initial variance and low prior weight.

3.2. Detecting local and global matching

The analysis of local areas near the face is close to the approach used in [4], but in contrast to it, we use the grid representation of the face image with the size of 3×3 elements. We get 9 patches which can be analyzed by 9 sub-streams in the form of the simplest CNN. At the output of such CNNs, the values of entropy and loss functions are estimated for each of 9 patches, forming the general assessment of the genuineness of the face image. Such local matching is a countermeasure to gradient attacks, which are usually local in their nature, and partly to attacks of adversarial perturbations. The global matching performs the global assessment of the entire face image. Its purpose is to identify 3D features. To do this, one can use different hardware and software solutions. Hardware solutions include the use of a 3D scanner (for example, Microsoft Kinect) or a stereo camera which is not always possible for practical application. Therefore, it is better to focus on software solutions, in particular, on using CNN trained to classify the depth of the scene.

The local and global matching is performed if the image passed the first stage (as the roughest fake). Moreover, this stage can be presented into a single network with two global streams. Presentation attacks usually distort image details. Therefore, special attention should be paid to the areas around the eyes, because these areas contain the most detailed information. Our approach of local matching is close to [15] and is based on the fully convolutional network (FCN), which was proposed by Long et al. in 2014 [16]. FCN is widely used in semantic image segmentation and differs from the traditional CNNs by convolutional layers instead of fully connected layers. Such an architecture tunes the network output into a heat map. The loss function has the form of Eq. 3, where $p_{i,j}(k) \in \{0, 1\}$ is the prior probability, $q_{i,j}(k)$ is the prediction probability, k is the true class (0 or 1, genuine or fake image).

$$L_{i,j} = -\sum_{k=0}^{1} p_{i,j}(k) \log q_{i,j}(k) \tag{3}$$

The general loss function is defined as the sum of the local loss functions on the grid. CNN builds a $2 \times n \times n$ probability map, and after summing the values of each $n \times n$ map, a 1×2 vector is formed to predict the class. In this case, the decision is made taking into account the predictions of each local region rather than on the basis of any dominant region.

The global matching is the assessment of the entire face image, which partly serves to validate the previous decision. Various representations of the input image are allowed, for example, representation in the YCbCr color space, in the form of LBP, high-frequency components, training on 3D models, etc. The experiments have shown good results for the models based on the transition to the YCbCr color model and analysis of high-frequency components of genuine and fake images. For the global matching, FCN with 6 convolutional layers and 2 pooling layers is also used, and SVM serves as a classifier. Then, the results of two streams are combined, and the final decision on the genuineness of the face image is made.

4. Experimental results

For the experiments, the OULU-NPU dataset [17] and own dataset were used. The OULU-NPU dataset contains 4950 videos received from 6 smartphones. The own dataset includes around 420 short videos with real faces, printed face images and videos from the tablet. The presentation attacks are of two types: print attacks and replay attacks. For experiments, print attacks were simulated. The

dataset was divided into a training set and a test set with the ratio of 70% to 30%. The proposed method showed the robustness to the presentation attacks and even to the attacks based on adversarial examples. According to ISO/IEC 30107-3:2017 [18], we calculated the following metrics: true accept rate (TAR), attack presentation classification error rate (APCER) as the false accept rate (FAR) and bona-fide presentation classification error rate (BPCER) as the false reject rate (FRR) (in terms of face recognition) provided by Eqs. 4-5, where *TP* is the true positive, *FP* is the false positive, *TN* is the true negative, *FN* is the false negative.

$$APCER = FP/(FP + TN) \tag{4}$$

$$BPCER = FN/(FN + TP) \tag{5}$$

Table 1 includes the estimates without and with deep learning approach with significant difference.

Table 1
Estimates of the fake image detection

Types of attacks	TAR, %	APCER, %	BPCER, %
	Without CNN		
Print attacks	59.3-65.1	10.4-15.7	8.4-9.2
	With CNN		
Print attacks	82.4-89.1	3.6-7.1	1.9-3.5
Attacks of adversarial perturbations	69.5-75.2	7.5-8.7	4.7-6.2

The experiments show that the accuracy of detecting fake face images reached 82.4-89.1% and 69.5-75.2% for the presentation attacks (print attacks) and attacks of adversarial perturbations, respectively.

The augmentation or generation of new data based on the existing dataset makes it quite easy to expand the training set. We applied data augmentation "on-the-fly", when new distorted samples were created directly during the training process between learning epochs without increasing the amount of initial data. The augmentation was carefully implemented using slight distortions of shooting conditions, affine deformation of objects, blur and reflection. This procedure improved the quality of the model and its robustness to noise in the input data. Using augmentation without changing the network architecture, it was possible to increase the accuracy of the fake face detection by 3.4% for print attacks.

5. Conclusion

At present, fake face image detection is a necessary procedure for the normal functioning of face recognition systems. In this study, it is shown that there are different approaches to solving this problem. However, for the protection against various types of attacks, it is reasonable to use several methods. We offer a two-stage method for verifying the genuineness of a face image before its entering the face recognition system. The first stage is the background analysis, while the second stage is local and global matching. For the background estimation, a Gaussian mixture model is built, and a two-stream deep neural network is created to assess local and global features. The experiments conducted on the OULU-NPU dataset and own dataset show the accuracy for the presentation attacks and attacks of adversarial perturbations to be 82.4-89.1% and 69.5-75.2%, respectively. Using data augmentation, it was possible to increase the accuracy of detecting the presentation attacks to 85.7-92.5%. However, the temporal estimates of the recognition process do not correspond to the real time and require further refinement of the algorithms.

6. References

[1] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Columbus, OH, USA, 2014, pp. 1701–1708.

[2] M. Wang, W. Deng, Deep face recognition: A survey. Neurocomputing 429 (2021) 215–244.

[3] J. Yang, Z. Lei, S.Z. Li, Learn convolutional neural network for face antispoofing. Cornell ArXiv Print, arXiv preprint arXiv:1408.5601, 2014, pp. 1–8.

[4] Y. Atoum, Y. Liu, A. Jourabloo, X. Liu, Face anti-spoofing using patch and depth-based CNNs, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, Denver, CO, USA, 2017, pp. 319–328.

[5] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, A. Hadid, An original face antispoofing approach using partial convolutional neural network, in: the 6th International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, Oulu, Finland, 2016, pp. 1–6.

[6] A. Jourabloo, Y. Liu, X. Liu, Face de-spoofing: Anti-spoofing via noise modeling, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. LNCS, volume 11217, 2018, pp 297–315.

[7] K. Patel, H. Han, A.K. Jain, Cross-database face antispoofing with robust feature representation, in: You, Z., Zhou, J., Wang, Y., Sun, Z., Shan, S., Zheng, W., Feng, J., Zhao, Q. (eds) Biometric Recognition (CCBR 2016), LNCS, volume 9967, 2016, pp. 611–619.

[8] Z. Xu, S. Li, W.Deng, Learning temporal features using LSTM-CNN architecture for face anti-spoofing, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, Kuala Lumpur, Malaysia, 2015, pp. 141–145.

[9] R. Shao, X. Lan, P. C. Yuen, Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, Denver, CO, USA, 2017, pp. 748–755.

[10] R. Shao, X. Lan, P. C. Yuen, Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing. Transactions on Information Forensics and Security 14(4) (2019) 923–938.

[11] G. Goswami, N. Ratha, A. Agarwal, R. Singh, M. Vatsa, Unravelling robustness of deep learning based face recognition against adversarial attacks, in: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana, USA, volume. 32, 2018, pp. 6829–6836.

[12] A. Goel, A. Singh, A. Agarwal, M.Vatsa, R. Singh, Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition, in: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, Redondo Beach, CA, USA, 2018, pp. 1–7.

[13] G. Mai, K. Cao, P. C .Yuen, A .K. Jain, On the reconstruction of face images from deep face templates, Transactions on Pattern Analysis and Machine Intelligence 41(5) (2018) 1188–1202.

[14] M. N. Favorskaya, V. V. Buryachenko, Background extraction method for analysis of natural images captured by camera traps, Information and Control Systems 6 (2018) 35–45.

[15] Y. Ma, L. Wu, Z. Li, F. Liu, A novel face presentation attack detection scheme based on multi-regional convolutional neural networks, Pattern Recognition Letters 131 (2020) 261–267.

[16] J. Long, E. S helhamer, T. Darrell, Fully convolutional networks for semantic segmentation, Transactions on Pattern Analysis and Machine Intelligence 39(4), (2014) 640–651.

[17] OULU-NPU – a mobile face presentation attack database with real-world variations, URL: https://sites.google.com/site/oulunpudatabase.

[18] ISO/IEC 30107-3:2017 Information technology – Biometric presentation attack detection – Part 3: Testing and reporting, URL: https://www.iso.org/standard/67381.html.