Face Image Manipulation Detection

To cite this article: Lilong Wen and Dan Xu 2019 IOP Conf. Ser.: Mater. Sci. Eng. 533 012054

View the article online for updates and enhancements.

You may also like

- Competition between surface relaxation and ballistic deposition models in scale-free networks
  C. E. La Rocca, P. A. Macri and L. A. Braunstein

- Computed tomography super-resolution using deep convolutional neural network
  Junyoung Park, Donghwi Hwang, Kyeong Yun Kim et al.

- Identification of cavitation intensity for high-speed aviation hydraulic pumps using 2D convolutional neural networks with an input of RGB-based vibration data
  Qun Chao, Jianfeng Tao, Xiaoliang Wei et al.

Recent citations

- Khushi Chandani and Monika Arora
Face Image Manipulation Detection

Lilong Wen¹, Dan Xu¹,a

¹School of Information Science and Engineering of Yunnan University
danxu@ynu.edu.cn

Abstract. This paper proposes a CNN-based (Convolutional Neural Network based) network to detect altered face picture, which can cover the most common face swap methods. The network uses an autoencoder which is pre-trained on the original images to reconstruct the input images. The reconstructed one and the input image are then processed by the SRM filter which can extract the noise distribution of images. We then feed the minus result of two processed results into a CNN architecture to predict whether the input image is original or tampered. The model was trained and evaluated in FaceForensics dataset and state-of-art face swap method. Experimental results demonstrate the effectiveness of our network.

1. Introduction

Images are important media that commonly used for both daily communication and crucial situation like political or commercial cases. But with the great progress achieved in computer vision and image process, changing the content of original images without detection have become much easier. With some user-friendly software like Photoshop and Meitu, everyone can do some manipulation to the original picture without authorization and post to the Internet. Distinguishing the original picture with tampered one become much more important than ever. Especially images involve face regions. There are many advanced applications can automatically swap two faces or manipulate one’s facial features benefited by the face detection and recognition techniques. Plenty of pictures on the Internet that are well tampered with fake face for fun or some other reasons which can always cause misleading and confusions in some situations. And with recently rising deep learning technology, many advanced image-to-image translating methods have been proposed to change the human face based on generative model. The result can be hard to distinguish for human even after close observation.

In order to tell the difference between real and fake pictures, there has already been a great achievement in this particular field. Some traditional methods such as error level analysis (ELA) [1], local noise features [2] and camera filter array (CFA) [3] have been developed. Recently, neural network has been widely used in picture forensics and have obtained the promised results. However most of the current approaches concentrate on general image tamper techniques such as copy-move, splicing, removal et al. and cannot totally combat the well shopped fake face pictures on the Internet.

Since unlike other image manipulation detection, photos with altered faces usually are well tampered and has no semantic difference with the origin one. The detection should pay more attention on the facial feature differences together with the differences between the face and the rest of the picture, which suggests more fine-grained features to learn.
2. Related Work

Research on image forensics has achieved great progress in recent years and plenty of detecting methods have been proposed. Prior works include some low-level feature analysis such as error level analysis (ELA) that rely on the compression artifacts caused by the lossy compression of the JPEG image, and some improved method based on the double image compression and Discrete Cosine Transform (DTC) coefficient analysis [4,5] could achieve a better result. There are also some methods based on the fingerprints left by internal camera processors such as Camera Response Function (CRF) [6] and photo response non-uniformity (PRNU) noise pattern [7]. And the work of Yerushalmy I and HelOr H. [8] directly rely on the physical structure and characteristics of digital cameras. What’s more, Color Filter Array (CFA) based methods has also been proved to be efficient such as the method proposed by Popescu AC and Farid H [9], which quantify the specific correlations introduced by CFA interpolation to automatically detect image tampering. And there are also many pixel level statistic methods such as using possibility map [10] and statistical properties of natural photographic images [11].

With recent success of deep learning approaches in computer vision and image process filed, researchers apply deep learning in photo forensic field, which have achieved a promised result. Bondi L et.al [12] use convolutional neural network (CNN) to detect different image patches under the camera model theory that assuming all the pixels from the same picture should have same character. Chen Jiansheng, et al.[13] apply convolutional neural network in median filtering image forensics, and feed the median filtering residual into the later network. Other deep learning architecture such as autoencoder and Long short-term memory (LSTM) network also perform well in image anomaly detection task except using CNN. Zhang et al. [14] use a staked autoencoder to learn complex features for each individual image patches, and then integrate the contextual information of each patch to perform the final results. Cozzolino D and Verdoliva L [15] use autoencoder to generate implicit information from the noise data to detect the anomaly. Jawadul H and Amit K. et.al employ a hybrid CNN-LSTM model to discriminate the features between altered and original areas.

Recently, local noise feature-based method like steganalysis rich model (SRM) [16] has been widely used in image forensic. Cozzolino D and Verdoliva L extract local noise features from image using SRM filters and then feed into convolutional neural network architecture. Peng Zhou et al. [17] also use a steganalysis feature extractor to pre-process images before feed into the patch triplet stream. Their later work [18] also adopt the same structure but instead of using all 30 SRM filter kernels, they only choose 3 kernels which can still achieve a decent performance with lower computational cost.

![Figure 1. Architecture of Proposed Network](image-url)
Inspired by those prior work, we designed a network that adopt both autoencoder and SRM filters as well as face classification stream to jointly detect the tampered face images.

3. Proposed Method

Inspired by the recent work and newly released dataset, this paper proposed a CNN-based architecture to detect the manipulated face images. The proposed network using a pre-trained autoencoder to reconstruct image from input ones, which works kind of like a filter that can rebuilt the input image that close to the input one when the input is untampered, but the similarity would be much less when the input is tampered. The model then combined the steganalysis rich model (SRM), which has been proved to be effective in image manipulation detection, to extract the noise distribution of two images. We use the minus result of those two processed images and then feed into the fine designed VGG16 [19] network to get the final prediction result. The full architecture is shown in figure 1.

To train and evaluate our network, we adopt a large-scale dataset for forgery detection in human face called FaceForensics [20], which is a video dataset contains 500,000 frames from 1004 videos collected from Youtube and the tampered data was created by applying state-of-the-art Face-to-Face approach [21] on the original data.

3.1. Autoencoder

An autoencoder is a type of artificial neural network used to learn the data representation through iterative forward and backward training. It includes two part, the encoder and decoder, the encoder encodes the data into low level representation and the decoder is designed to perform an opposite operation that reconstruct the data to produce a result as close as possible to input data. The encoder and decoder can be presented in (1) and (2), while the (3) shows the loss function training the network. And the (4) shows the parameters updated by gradient descent.

\[
\hat{x}_j = \phi^{-1}(h_j) = f(x_j)
\]

\[
J(f) = \frac{1}{n} \sum_{i=1}^{n} || f(x_i) - x_i ||^2_2
\]

\[
g = \frac{dJ}{df} = \frac{1}{n} \sum_{i=1}^{n} g_i = \frac{2}{n} \sum_{i=1}^{n} (f(x_i) - x_i)
\]

When update the parameters of autoencoder using gradient descent as (4) shows, the gradient is averaged over all training data, so an autoencoder dose not try to reduce the error of single datum but the overall error. We can measure the error reduction in single datum by using effective gradient magnitude using (5)

\[
geffect_i = \frac{\langle g_i, \bar{g} \rangle}{|| g_i ||} = || \bar{g} || \cos \theta(g_i, \bar{g})
\]

Where \( \theta(g_i, \bar{g}) \) is the angle between \( g_i \) and \( \bar{g} \). According to the research of Xia Y et.al [22], the angle for a positive datum is often smaller than it is for an outlier, leading to larger effective gradient magnitude. In another words, the reconstruction errors of positive data are always smaller than that of outliers.

Based on this idea, the network is first pre-trained off-line on original picture to fully learn the positive data distribution, after that, it could reconstruct the positive image with good accuracy while negative ones with large errors, which means the original images and the altered ones would get a quite
different reconstructed result through the network. And this kind of difference could be used for later distinguish work. After pre-training the CAE, we freeze all parameters of this autoencoder and does not involve this into training. It will only be used as an image filter in the final network.

3.2. Noise Distribution of Image
Visual features are not quite enough to cover all the face tamper operations. Pictures that well altered with sophisticated methods can reduce any kinds of splicing boundary and color inconsistent cannot be detected by only using the visual features. So we adopt this noise classification stream to utilize the local noise distribution of images to get more information about the tampered face area. According to the work of [18], We use the same 3 SRM kernels to reduce the computational cost and speed up the process, which is showed in (6).

\[
\begin{array}{cccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 2 & -2 & 2 & -1 & 0 \\
0 & -2 & -4 & 2 & 0 & 1 \\
0 & -1 & 2 & -1 & 0 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

(6)

And to demonstrate the effectiveness of the SRM filter, we pick an original image and it’s corresponding tampered one from dataset to perform minus operation before and after SRM filter, which is shown in figure 2.

![Figure 2. The Noise Feature of both Original and Altered Images Extracted by SRM Filters. The Third Column Shows the Subtraction Result of Previous Two Columns. (The Subtraction Value is Round to be Integer)](image)

3.3. Classification
After processed by the SRM filter, we perform minus operation between two images to get an intermediate result. Figure 3 shows the minus results between original untampered input and tampered input. To speed up the training process, we adopt VGG16 and perform a transfer learning process in this part.

4. Experiments
In this section, we demonstrate our architecture on FaceForensics dataset and compared with other state-of-the-art classification method on this dataset. What’s more, to prove the effectiveness beyond the dataset, we also test our model in several state-of-art open-sourced face-swap method found in github.

4.1. Setting Up Datasets
FaceForensics dataset is a large-scale dataset created by using automated version of Face2Face approach. All videos are downloaded from Youtube and cut down to short continuous clips that
contain mostly frontal faces. The dataset is divided into three parts, the training set, the validation set and the test set. All sets include number-balanced original and altered images. The training set, the test set and the validation set include 736270, 155490 and 151052 images, respectively. The figure 3 shows some examples of original and altered images from dataset.

![Figure 3. Some Examples of Original and Altered Images from Faceforensics Dataset](image)

4.2. Training Model
Firstly, we pre-train the CAE on the positive original data until the reconstruction loss is converged into an acceptable level that the reconstructed image is very close to the input one, then save the weights of the network. In addition, to prevent the over-fitting to the dataset, we also used some public available face dataset [23, 24, 25] while training the CAE.

Then, we perform an end-to-end training process using the whole dataset including the untampered original images the tampered ones.

4.3. Testing and Compared with Other Methods
We evaluated the model using the test set of the dataset which have 77745*2 pictures. And compared with traditional classification architecture VGG16 and Resnet [26]. We use the H.264 compressed data with quantization parameter 35 since the uncompressed data is too easy to classify. The test result is shown in table 1.

What’s more, to prove the effectiveness of our model, we also trained three face-swap models found in github which all using state-of-art methods to produce tampered face images for testing. The face-swap models we used are PRNet [27, 28], DeepFaceLab [29] and face2face-demo [30]. All those three models are trained using video and image data from FaceForensics dataset and use the same swap paradigm that swap the same person’s face in the same scenes, which make the result image looks more real and only difference would be the face region. The test result in those tampered images produced by above three model is shown in table 2, and those images are using the same compression method as well.

| method   | accuracy |
|----------|----------|
| VGG16    | 78.13    |
| ResNet   | 82.45    |
| Ours     | 90.76    |

| method                | accuracy |
|-----------------------|----------|
| PRNet                 | 90.80    |
| DeepFaceLab           | 94.87    |
| face2face-demo        | 93.14    |

5. Conclusions
We propose a CNN-based architecture for tampered face detection. The experimental result shows that our approach outperforms other prior works. What’s more, since this paper has proved that combined with a pre-trained autoencoder, the negative input could have much difference with the reconstructed
one compared to the positive input. This method could be used in all kinds of binary classification problem and hopefully could produce a better performance compared to the previous methods.

References

[1] Barton K C. A picture's worth[J]. Social Education, 2001, 65(5): 278.
[2] Pan X, Zhang X, Lyu S. Exposing image splicing with inconsistent local noise variances[C]//Computational Photography (ICCP), 2012 IEEE International Conference on. IEEE, 2012: 1-10.
[3] Goljan M, Fridrich J. CFA-aware features for steganalysis of color images[C]//Media Watermarking, Security, and Forensics 2015. International Society for Optics and Photonics, 2015, 9409: 94090V.
[4] Bianchi T, De Rosa A, Piva A. Improved DCT coefficient analysis for forgery localization in JPEG images[C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 2444-2447.
[5] Galvan F, Puglisi G, Bruna A R, et al. First quantization matrix estimation from double compressed JPEG images[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(8): 1299-1310.
[6] Hsu Y F, Chang S F. Image splicing detection using camera response function consistency and automatic segmentation[C]//Multimedia and Expo, 2007 IEEE International Conference on. IEEE, 2007: 28-3.
[7] Lukáš J, Fridrich J, Goljan M. Detecting digital image forgeries using sensor pattern noise[C]//Security, Steganography, and Watermarking of Multimedia Contents VIII. International Society for Optics and Photonics, 2006, 6072: 60720Y.
[8] Yerushalmy I, Hel-Or H. Digital image forgery detection based on lens and sensor aberration[J]. International journal of computer vision, 2011, 92(1): 71-91.
[9] Popescu A C, Farid H. Exposing digital forgeries in color filter array interpolated images[J]. IEEE Transactions on Signal Processing, 2005, 53(10): 3948-3959.
[10] Li H, Luo W, Qiu X, et al. Image forgery localization via integrating tampering possibility maps[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(5): 1240-1252.
[11] Lyu S. Natural Image Statistics in Digital Image Forensics[M]//Digital Image Forensics. Springer, New York, NY, 2013: 239-256.
[12] Bondi L, Lameri S, Guera D, et al. Tampering Detection and Localization Through Clustering of Camera-Based CNN Features[C]//CVPR Workshops. 2017: 1855-1864.
[13] Chen, Jiansheng, et al. "Median filtering forensics based on convolutional neural networks." IEEE Signal Processing Letters 22.11 (2015): 1849-1853.
[14] Zhang Y, Goh J, Win L L, et al. Image Region Forgery Detection: A Deep Learning Approach[C]//SG-CRC. 2016: 1-11.
[15] Cozzolino D, Verdoliva L. Single-image splicing localization through autoencoder-based anomaly detection[C]//Information Forensics and Security (WIFS), 2016 IEEE International Workshop on. IEEE, 2016: 1-6.
[16] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
[17] Zhou P, Han X, Morariu V I, et al. Two-stream neural networks for tampered face detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2017: 1831-1839.
[18] Zhou P, Han X, Morariu V I, et al. Learning Rich Features for Image Manipulation Detection[J]. arXiv preprint arXiv:1805.04953, 2018.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[20] Rössler A, Cozzolino D, Verdoliva L, et al. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces[J]. arXiv preprint arXiv:1803.09179, 2018.
[21] Thies J, Zollhofer M, Stamminger M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2387-2395.
[22] Xia Y, Cao X, Wen F, et al. Learning discriminative reconstructions for unsupervised outlier removal[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1511-1519.
[23] Zheng T, Deng W, Hu J. Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments[J]. 2017.
[24] Kumar N. Attribute and Simile Classifiers for Face Verification[C]// IEEE International Conference on Computer Vision. IEEE, 2010.
[25] Liu Z, Luo P, Wang X, et al. Deep Learning Face Attributes in the Wild[J]. 2014.
[26] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015.
[27] Feng Y, Wu F, Shao X, et al. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network[J]. 2018.
[28] https://github.com/YadiraF/PRNet
[29] https://github.com/iperov/DeepFaceLab
[30] https://github.com/datitran/face2face-demo