https://doi.org/10.2352/ISSN.2470-1173.2017.7.MWSF-329

Image Recapture Detection with Convolutional and Recurrent Neural Networks

Haoliang Li, Shiqi Wang and Alex C.Kot, Nanyang Technological University

Abstract

In this paper, we aim to address the image recapturing detection problem with the convolutional and recurrent neural networks. With the advances of image display and acquisition techniques, the recaptured images are of satisfactory quality nowadays. This has been creating an ever stronger demand for sophisticated image recapturing detection algorithms which can efficiently prevent the unauthorized image distributing and forging. In this paper, we propose a hierarchical feature learning strategy by leveraging the intra-block information and inter-block dependency for image recapturing detection. In particular, the image blocks are first employed as the input of the convolutional neural network (CNN) and subsequently recurrent neural network (RNN) is further adopted to extract block dependencies. The CNN and RNN serve as effective tools to extract discriminative and meaningful features regarding both intra- and inter-block information. As such, the inherent properties within local blocks and the correlations between non-local neighbouring blocks are all exploited to identify the recaptured images. Experimental results on three databases show that significantly better performance can be achieved with our proposed framework compared to traditional handcrafted and deep learning based approaches.

Introduction

With the technological advances of digital image acquisition and display, it becomes effortless to recapture an authorized image from printed paper and LCD monitor screen under well-controlled conditions. In general, image recapturing brings more authenticity to fool people. According to the study in [1], human beings are difficult in distinguishing the original and recaptured images. A similar problem, face spoofing, which aims at fooling the face recognition system using fake facial images/videos, can also be regarded as the recapturing problem. It refers to the attack by presenting a printed face on paper or replaying a video by mobile devices (e.g. mobile phone or tablet) to get through the face recognition system. In [2], it is reported that the state-of-the-art Commercial Off-The-Shelf face recognition systems are even vulnerable to face spoofing attack. As a result, there is considerable concern regarding the security issue of biometric identification.

The multimedia reproduction has inspired numerous algorithms proposed to address the problem of image recapturing and face anti-spoofing detection to prevent the illegal use of recaptured images. One of the earliest works was reported by Farid and Lyu [5], who used wavelet statistical features to detect image scanning. It achieves an average accuracy of 99.6% using their proposed dataset. The physical based methods [7], such as the spatial distribution of specularity and chromaticity features have also been used to distinguish between the original image and the printed version. The detection of recaptured image from LCD screen was first proposed by Cao and Kot [1]. The texture pattern caused by aliasing was detected by calculating multiple scales of Local Binary Pattern (LBP) feature, which has been proved to be very powerful in texture classification. The loss of detail, which is caused by the relatively low resolution of LCD screen, is detected by computing multi-scales wavelet decomposition where the absolute mean and standard values of Haar wavelet coefficients are used as features. Since the texture and physical artifacts may also influence the frequency spectrum of images, Yin and Fang [4] proposed to use DCT based Markov Transition Matrix to model DCT coefficients dependencies under the assumption that they may get tampered by recapturing. A recent work [10] proposed a capturing process which can get rid of the unexpected texture pattern. In particular, images captured under the specified conditions showed better quality compared with normal recapture images. On the other hand, they proposed to employ blur degree as discriminative features for image recapturing detection. More recently, a convolutional neural network (CNN) network with Laplacian filter was proposed in [9] to learn the deep representation of recapturing detection which shows promising performance even based on small image patches.

Computer Graphic Image (CGI) Detection shares a similar idea with image recapturing detection. CGI detection aims at distinguishing realistic photo with computer generated photo. A multi-scale detection approach was proposed in [11] where local fractal dimension and local patches were used as the finest scale feature. In addition, Beltrami flow vectors were adopted as the intermediate scale feature. In view of the fact that CGI detection and image recapturing detection are different to some extent since computer generated photo can still be considered as originally captured photo, Gao *et al.* [7] showed that the proposed feature in [11] can be transferred to image recapturing detection. Another similar topic is steganalysis which targets on detecting hidden message inside a given image. As stated in [7], image recapturing detection problem is similar to steganalylsis by considering both of them as recognition problem. In particular, they can be regarded as detecting the variations on statistical properties. In [12], the Markov based Transition Probability matrix is computed as the feature to detect both intra-block and inter-block statistics change for steganalysis.

In this paper, we target at learning the powerful deep representation for image recapturing detection which can extract both intra-block and inter-block information. Our motivation originates from two observations. Firstly, recapturing process introduces unexpected artifacts, which can be further categorized into three different parts including loss-of-the-detail artifact, color distortion and anti-aliasing texture [1]. As such, we expect to learn an appropriate representation with CNN network to account for these artifacts. Secondly, natural images hold a certain inter-

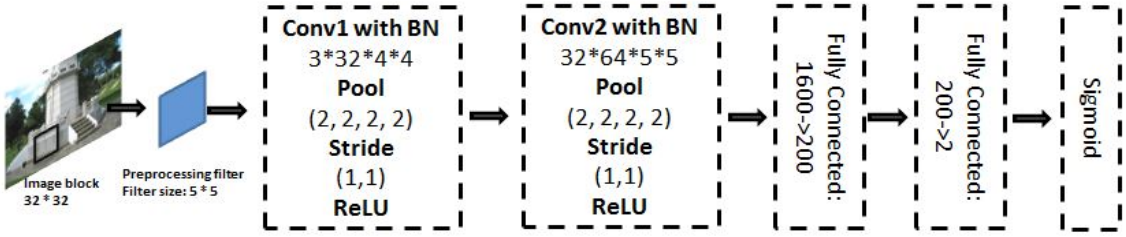


Figure 1. *CNN model for intra-block feature extraction.The input image block size is* 32×32.

block dependency during image coding, and such dependency can be distorted when a captured image goes through the coding process twice. Therefore, we employ a 2D recurrent neural network (RNN) to capture such dependency.

The main contributions of this work are as follows:

- We propose to learn a pre-processing filter for CNN model instead of directly adopting a pre-defined filter. By learning such a filter, we expect to extract useful information which covers the general artifacts appeared in the recaptured images.
- We propose to learn the dependencies of adjacent image blocks by employing 2D RNN model based on the extracted CNN feature. As such, image block dependency introduced by compression can be explored. On the other hand, a more representative feature can be learned by treating RNN model as a hierarchical feature learning stage.

Image Recapturing Detection Scheme

In this section, we first introduce the CNN architecture which is employed to capture the intrinsic properties of image block. Subsequently, we propose the RNN based technique for inter-block dependency exploitation.

Intra-block Feature Extraction

Recently, CNN has been widely adopted for multimedia forensics analysis [15, 9], and most of the schemes share a similar module with a pre-processing step (e.g. high pass filter for steganalysis [16], median filter for median filtering detection and Laplacian filter for recapturing detection [9]) before feeding the input into CNN model. However, such filters are usually designed as predefined filters with fixed coefficients such that the specific task and corresponding inherent data properties have been largely ignored. Here, instead of directly adopting a filter, we seek to learn an appropriate pre-processing filter for our specific task.

Our adopted CNN model is based on LeNet [13], and modifications have been made based on the recapturing detection problem. In particular, although Laplacian filter [9] works well as a pre-processing step for image recapturing detection, it still cannot fully extract useful information since recapturing process induces not only high-frequency information loss but also other distortions (e.g. color). However, directly employing the Laplacian filter may remove such valuable information which is very helpful in image recapturing detection. Therefore, instead of directly applying a existed filter for pre-processing, we aim to learn a desired filter. In our work, a 5×5 convolution layer with zero padding is incroporated as the pre-processing stage, and the filter coefficients are adaptively learned based on the training data.

Moreover, Batch-Normalization (BN) is adopted in the convolutional layer. The motivation behind employing BN module is that CNN model is likely to be overfitted to image contents. In particular, when the training and testing images are captured from different environments, the so-called covariate shift problem [14] can cause overfitting problem especially when the contents of images are not diverse enough. Therefore, BN module is employed to reduce the covariate shift [14] induced by image content variations. Typically, such module has also been considered in [9].

To summarize, our proposed image recapturing detection network comprises three convolutional layers (the last two convolutional layers are coupled with pooling and ReLU layers) and two fully connected layers (the last layer has 2 nodes). The structure of our adopted CNN model is illustrated in Fig. 1.

Inter-block Feature Extraction

Besides detecting image blocks with the 32×32 block, we are also interested in larger image block size such as 64×64. One possible solution for recapturing detection regarding larger image block size is to construct a neural network with the size of 64×64. Another solution is to propose a voting strategy based on our proposed network by dividing the input 64×64 image block into sub-blocks. Here, we adopt the second strategy and divide the 64×64 image block into nine different sub-blocks with an overlapping size of 16.

In this paper, we propose an advanced voting strategy by leveraging the neighbor blocks dependencies. In particular, given an image, there are two issues need to be addressed. First, only focusing on extracted information based on local image block may not fully capture the image representative information, and the non-local image information should be taken into account as well. Second, inter-block dependency should also be considered in the recapturing detection process. Since recurrent neural network (RNN) is a powerful method to learn robust feature representation of a sequential input, we hereby propose to combine CNN and RNN to learn both intra-block and inter-block information.

To extract the block dependency information, we follow the framework of directed acyclic graphs recurrent neural network

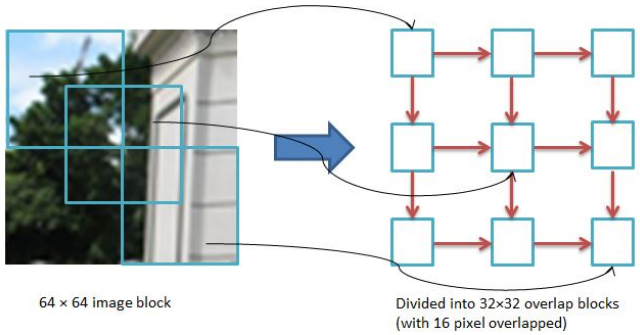


Figure 2. *An example of DAG in southeastern direction.*

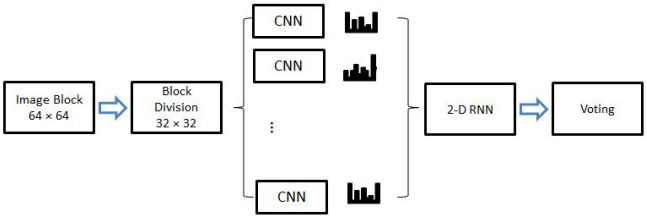


Figure 3. *The proposed feature extraction framework for recapturing detection.*

(DAG-RNN) proposed in [6]. Specifically, it is defined as

$$\begin{aligned} \hat{h}^{(v_i)} &= \sum_{v_j \in \mathscr{P}_g(v_i)} h^{(v_j)} \\ h_{(v_i)} &= f(Ux^{(v_i)} + W\hat{h}^{(v_i)} + b) \\ o^{(v_i)} &= g(Vh^{(v_i)} + c), \end{aligned} \tag{1}$$

where $v_i, x^{(v_i)}, h^{(v_i)}, o^{(v_i)}$ are the ith vertex of DAG, input, hidden representation and output respectively, $\mathscr{P}_g(v_i)$ is the direct predecessor set of v_i defined in [6]. In Fig. 2, we show an example which extracts blocks dependency in southeastern direction.

The employed 2D RNN can be treated as a hierachical step to learn a more representative feature based on CNN feature, where the feature obtained by RNN $h_{(v_i)}$ comes from both $x^{(v_i)}$ and its neighbor $h_{(v_j)}$. The framework of our proposed algorithm is shown in Fig. 3. In particular, to fully extract the block dependency property, four DAG directions, southeastern, northeastern, southwestern and northwestern, are considered, as suggested in [6]. Once the output $o^{(v_i)}$ is determined, we then propose the voting strategy based on $o^{(v_i)}$.

Experimental Results

In this section, we evaluate the proposed scheme based on three image recapturing databases. In particular, we first introduce the image recapturing databases and the evaluation baseline methods. Subsequently, the detection accuracy for the proposed and the baseline schemes are provided. Finally, we perform the filter visualization to further investigate the properties of the learned filter.

Image Recapturing Database

In this section, we evaluate our proposed image recapturing detection algorithm on three public databases, including Astar database [8], NTU-ROSE [1] database and ICL database [10]. The recaptured images from these three databases were obtained under different environmental settings. In particular, for the recaptured images in Astar database, both images taken from printed paper and digital screen were considered. For NTU-ROSE database, the recaptured images were acquired in a lighting controlled room, and only LCD screen was employed as the recapturing medium. For ICL database, images were taken in strictly controlled conditions with a certain distance between the LCD screen and the camera. Due to the diverse recapturing environments, the distortion types regarding these three databases are also different. For NTU-ROSE database, the dominant artifact we observe is the unexpected texture which is caused by aliasing. For Astar database, loss-of-the-detail artifacts, texture artifacts and even the artifacts caused by illumination and lighting reflection appear due to the uncontrolled capturing condition. However, as claimed in [10], only loss-of-the-detail artifact occurs in the ICL database since a novel capturing strategy was proposed to avoid anti-aliasing pattern.

Data Preparing

For each database, we randomly select multiple 64×64 images patches from each image. To reduce content similarity, we consider 64 pixels as the stride when extracting image patches. Half of the image patches are used for training and the remaining ones are for testing. It should be noted that the patches for training and testing should not be collected from the same image to ensure the content divergence. We repeat the process for 10 times and the average results are reported in the experiment.

Baseline Methods

To illustrate the performance of the proposed scheme, we compare our proposed framework with the state-of-the-art handcrafted features as well as deep learning features with and without pre-processing.

- **Multi-scale Local Binary Pattern**: As stated in [1], aliasing causes the unexpected texture pattern during recapturing process. We adopt here the multi-scale LBP feature which is proved to be effective in texture classification.
- **Multi-scale wavelet statistics**: JPEG coding is widely adopted in RGB format during the capturing process, which introduces distortion artifacts that cannot be recovered. Such distortions can be more severe when a scene goes through the compression process twice, such that stronger loss-of-the-detail artifacts appear in the recaptured image compared with the single captured one. Therefore, we employ the multi-scale wavelet statistics in multi-scale frequency domain to compute the energy as the feature for detecting the loss-of-the-detail artifacts.
- **CNN model with and without Laplacian filter**: Preprocessing step is widely adopted in multimedia forensics analysis which aims to extract useful information and filter out the noisy signal. We follow the work in [9] and employ Laplacian filter as pre-processing module before feeding the image into the CNN model. Meanwhile, we also evaluate the performance by using CNN without the pre-processing step.
- **CNN model with learned filter**: We also consider our pro-

posed framework without RNN as one of the baseline methods. In this manner, we expect to demonstrate the effectiveness of learned filter compared with predefined filter. The filter can be a delta function if no pre-processing is favored for image recapturing detection.

Results

Table 1. Experimental Results on Image Recapturing Detection. The results are evaluated by accuracy (%).

Method \ Database	ASTAR	NTU-ROSE	ICL
Multiscale LBP [1] (64 × 64)	72.31%	88.80%	71.44%
Multiscale Wavelet [1] (64 × 64)	69.95%	78.40%	76.36%
Proposed CNN without filtering (32 × 32)	86.05%	94.56%	95.27%
Proposed CNN with Laplacian filter (32 × 32)	85.59%	95.80%	97.01%
Proposed CNN with learned filter (32 × 32)	86.78%	96.93%	97.79%
Proposed CNN with voting (64 × 64)	89.89%	97.72%	98.95%
Proposed CNN with 2D RNN (64 × 64)	93.29%	98.67%	99.54%

The experimental results are listed in the Table 1. Based on the results, we can observe that CNN achieves significant improvement compared with the traditional hand-crafted features, which is not surprising since the hand-crafted features cannot fully extract useful information regarding the image recapturing artifacts. Moreover, such hand-crafted features are also bias towards the image content. We can also observe that the pre-processing module can consistently boost the performance of image recapturing detection. However, the results show that the predefined filter does not always improve the performance. This observation implies that useful information may be filtered out by the pre-defined filter (e,g. Laplacian filter). Finally, we show that by employing RNN model as a hierarchical feature learning step, the performance can be further improved. Such results demonstrate that the block dependency information is important for image recapturing detection.

Filter Visualization

We are particularly interested in the filter learned as the preprocessing module. Here, we visualize the frequency spectrum of the filter learned by back-propagation and the frequency spectrum is shown in Figure 4.

As we can see, for Astar and NTU-Rose database, the spectrum of our learned filter can be treated as a band-pass filter. Such observation indicates that both low-frequency and high-frequency components are important for image recapturing detection. We conjecture that the low-frequency component comes from the color distortion while the high-frequency comes from the loss-of-the-detail and aliasing texture artifacts. For ICL database, the spectrum is closer to a high-pass filter. This is reasonable since the recapturing condition strictly controlled, only loss-of-the-detail artifact appears in ICL database.

Conclusion

In this paper, we tackle the problem of image recapturing detection by deep learning. We propose a novel framework that considers both intra-block information and inter-block dependencies. In particular, we first propose to learn a preprocessing filter for this specific problem by back-propagation to extract the meaningful information in recapturing detection. We subsequently establish the 2D-RNN structure to further exploit the block dependencies and improve the detection performance. The experimental results on three different databases show the superior performance of our proposed scheme compared to both state-of-the-art handcrafted and deep learned features.

Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

References

[1] H. Cao and A. C. Kot. "Identification of recaptured photographs on LCD screens." 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010.

[2] D. Wen, H. Han and A. K. Jain. "Face spoof detection with image distortion analysis." IEEE Transactions on Information Forensics and Security 10.4 (2015): 746-761.

[3] J. Doe, Recent Progress in Digital Halftoning II, IS&T, Springfield, VA, 1999, pg. 173.

[4] J. Yin and Y. Fang. "Markov-based image forensics for photographic copying from printed picture." Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012.

[5] H. Farid and S. Lyu. "Higher-order wavelet statistics and their application to digital forensics." IEEE workshop on statistical analysis in computer vision. Vol. 8. Madison, Wiscomsin, 2003.

[6] B. Shuai, Z. Zuo, B. Wang and G. Wang. "DAG-Recurrent Neural Networks For Scene Labeling." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] X. Gao, T.-T. Ng, Q. Bo, S.-F. Chang, "Single-View Recaptured Image Detection Based On Physics-Based Features," IEEE International Conference on Multimedia & Expo (ICME), 2010.

[8] X. Gao, Q. Bo, J. Shen, T.-T. Ng, Y.-Q. Shi, "A Smart Phone Image Database for Single Image Recapture Detection," International Workshop on Digital Watermarking (IWDW), 2010

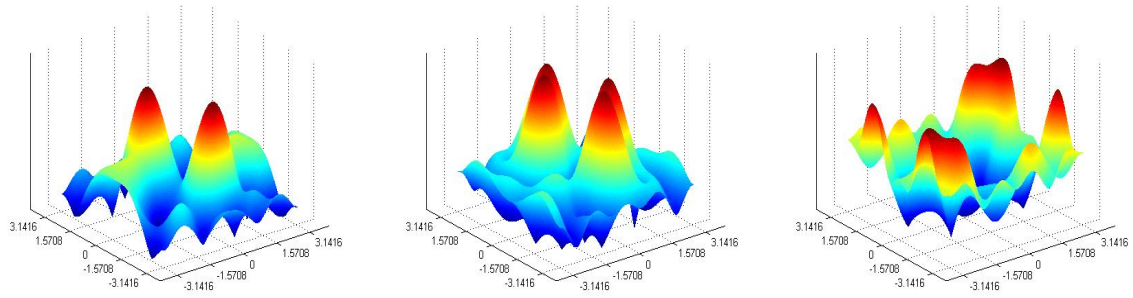
[9] P. Yang, R. Ni and Y. Zhao, "Recapture Image Forensics Based On Convolutional Neural Networks With Laplacian Filter Layer," International Workshop on Digital Watermarking (IWDW), 2016

[10] T. Thongkamwitoon, H. Muammar, and P. Dragotti, "An image recapture detection algorithm based on learning dictionaries of edge profiles." IEEE Transactions on Information Forensics and Security (2015): 953-968.

[11] T.-T. Ng, S.-F. Chang, J. Hsu, L. Xie, and M.-P. Tsui, "Physics-motivated Features for Distinguishing Photographic Images and Computer Graphics," in ACM Multimedia, 2005.

[12] Yun Q. Shi, C. Chen, and W. Chen. "A Markov process based approach to effective attacking JPEG steganography." International Workshop on Information Hiding (IWDW) 2006.

[13] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner "Gradient-based learning



(a) Learned filter on Astar database (b) Learned filter on NTU-Rose database (c) Learned filter on ICL database

Figure 4. *Frequency spectrum visualization of learned filter.*

applied to document recognition." Proceedings of the IEEE (1998): 2278-2324.

[14] S. Ioffe, and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

[15] J. Chen, X. Kang, Y. Liu, and Z. Jane Wang. "Median filtering forensics based on convolutional neural networks." IEEE Signal Processing Letters 22, no. 11 (2015): 1849-1853.

[16] G. Xu, H. Wu, and Yun Q. Shi. "Ensemble of CNNs for Steganalysis: An Empirical Study." Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. ACM, 2016.

Author Biography

Haoliang Li received his B.S. degree from University of Electronic Science and Technology of China in 2013. He is currently pursuing the Ph.D. degree in Nanyang Technological University, Singapore. His research interest is multimedia forensics.

Shiqi Wang received the B.S. degree in computer science from the Harbin Institute of Technology in 2008, and the Ph.D. degree in computer application technology from the Peking University, in 2014. He is currently with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore, as a Research Fellow. His research interests are image and image/video coding, processing, quality assessment and analysis.

Alex C. Kot is with the Nanyang Technological University, Singapore. He is currently a Professor and Associate Dean for the College of Engineering, the Director of Rapid-Rich Object Search (ROSE) Laboratory, and the Director of the NTU-PKU Joint Research Institute. He has authored or coauthored works in the areas of signal processing for communication, biometrics, data-hiding, image forensics, information security, and image object retrieval and recognition.