Jabeen, S.; Khan, UG.; Iqbal, R.; Mukherjee, M.; Lloret, J. (2021). A deep multimodal system for provenance filtering with universal forgery detection and localization. Multimedia Tools and Applications. 80(11):17025-17044. https://doi.org/10.1007/s11042-020-09623-w

Noname manuscript No.
(will be inserted by the editor)

A Deep Multimodal System for Provenance Filtering with Universal Forgery Detection and Localization

Saira Jabeen · Usman Ghani Khan · Razi Iqbal · Mithun Mukherjee · Jaime Lloret

Received: date / Accepted: date

Abstract Traditional multimedia forensics techniques inspect images to identify, localize forged regions and estimate forgery methods that have been applied. Provenance filtering is the research area that has been evolved recently to retrieve all the images that are involved in constructing a morphed image in order to analyze an image, completely forensically. This task can be performed in two stages: one is to detect and localize forgery in the query image, and the second integral part is to search potentially similar images from a large pool of images. We propose a multimodal system which covers both steps, forgery detection through deep neural networks(CNN) followed by part based image retrieval. Classification and localization of manipulated region are performed using a deep neural network. InceptionV3 is employed to extract key features of the entire image as well as for the manipulated region. Potential donors and nearly duplicates are retrieved by using the Nearest Neighbour

S. Jabeen
Department of Computer Science and Engineering,
University of Engineering & Technology, Lahore, Pakistan
E-mail: saira.jabeen@kics.edu.pk

U. Ghani Khan
Department of Computer Science and Engineering,
University of Engineering & Technology, Lahore, Pakistan
E-mail: usman.ghani@kics.edu.pk

R. Iqbal
Al-Khawarizi Institute of Computer Science,
University of Engineering & Technology, Lahore, Pakistan
E-mail: razi.iqbal@ieee.org

M. Mukherjee
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis,
Guangdong University of Petrochemical Technology, Maoming 525000, China
E-mail: m.mukherjee@ieee.org

J. Lloret
Universitat Politecnica de Valencia, 46022 Valencia, Spain
E-mail: jlloret@dcom.upv.es

Algorithm. We take the CASIA-v2, CoMoFoD and NIST 2018 datasets to evaluate the proposed system. Experimental results show that deep features outperform low-level features previously used to perform provenance filtering with achieved Recall@50 of 92.8%.

Keywords Provenance Filtering · Convolutional Neural Networks · Forgery Detection and Localization · Manipulation Detection

1 Introduction

Social networks are expanding day by day through the electronic web. These networks have become a great source of communication and connection for people from all over the world. People around the globe upload their private and general information on these social platforms, such as Facebook, Twitter, YouTube, and Instagram. These digital information is not only uploaded but is also consumed with a considerable ratio on a daily basis. When this extensive amount of data is shared and viewed around the world, the authenticity of this multimedia becomes a major concern. A large number of software tools such as Photoshop and GNU Image Manipulation Program (GIMP) happens to be the cause of image manipulation. Both image manipulation tools and social networks play a significant role in creating and spreading fake news that can reach to the target audience without any significant effort [39]. The Internet users can easily download, crop, and copy/paste material from other images and can repost the altered version on Internet websites. Such an attitude can adversely affect the reputation of any public celebrity as well as any individual. Hence, media forensics becomes an integral part of investigating cybercrimes.

Image authentication techniques are broadly divided into: 1) *active* image authentication technique that involves digital watermarking and signatures and 2) *passive* image authentication techniques. However, the later is considered more useful despite being challenging because of its independence upon foregoing image templates. Passive image authentication techniques are based on the assumption that any manipulation demonstrated upon an image, leaves its articrafts. This assumption has led to many effective solutions for forgery detection. Several attempts in digital forensics were made from the last few decades to retain trust in digital media by recognizing forged and original images [12, 24, 25, 34].

Initially, low-level features such as Discrete Cosine Transform (DCT) [6], Scale Invariant Features Transforms (SIFT) algorithm [35] and Speed up Robust Features (SURF) based techniques [9] were employed along with the supervised classifiers to classify forged versus original image [7, 25, 44]. These techniques worked well with small datasets and some specific forgery type. ~~Neural networks have shown remarkable progress in eclectic applications [11, 30, 31].~~ Recently, Bhatti et al. [11] proposed an image classification based on the optimal feature selection leveraging neural network. In addition, Convolutional Neural Networks (CNN)-based deep unified model was suggested in [30]. With the advantage of computational resources of the edge nodes,

the data latency has been significantly reduced while providing acceptable image detection accuracy. Besides, the capabilities of Deep Neural Networks (DNN) [43, 45] to learn more complex underlying statistical representation of images while keeping it general has engaged the interest of researchers in discipline of forgery detection and classification. A few techniques were proposed using CNN targeting copy-move forgery detection [42], splicing detection [10], and near-duplicate forgery detection [17]. Besides classification, CNN were also utilized to localize forged regions [52]. With the rise in fake information, a trend towards multimedia forensics has gained prominent interest in provenance analysis. Provenance filtering, is the process of suspecting the origin of fake images which was initially addressed in [40] as multimedia forensics problem. However, the efforts for manipulation detection, localization, and provenance filtering has been made separately.

In this paper, we aim to propose a multimodal neural network-based system, which is capable of solving both problems while giving comparable results. The main contributions are summarized as follows:

- Firstly, we propose a CNN-based architecture that classifies and localizes universal forgery in the images.
- Secondly, we transfer the learned parameters of the pre-trained neural network to produce robust key features of images and their manipulated regions in order to perform provenance filtering.

The rest of the paper is organized as follows. Section 2 presents the related work. The proposed multimodal system and methodology are discussed in Section 3. Evaluation benchmark datasets and experimental results with comparative analysis are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 Related Work

Researchers have been making attempts to address the issue of manipulated media for several years. Many frameworks have been developed to prove the authenticity of an image. Initially, this research was called as *copy-move detection* [8], where forgery identification is performed on an image that is changed by either copying and pasting some part into image or removing sub-part of the original image. The efforts to identify forgery can be broadly divided into two categories: a) low-level key features-based and b) deep learning-based techniques.

2.1 Low-level key features-based techniques

Fridrich et al. [25] specifically targeted the copy-move type of manipulation. They proposed two methods for such forgery detection: one method uses exact match for the detection and other method considers an approximation.

For the exact match, the user-specified size of the considered segment is slid throughout the whole image. When two consecutive rows are found identical, that region of an image is considered as *copied* image. Moreover, in an approximate match, the comparison is not made between pixel values but between the DCT representation of the pixels. Through this approach, they were successfully able to detect copy-pasted regions in a forged image.

Similar type of problem was solved by Amerini et al. [7], where they even made attempts to estimate geometric transformations. The SIFT algorithm [35] was used to extract the robust features from an image. The matching between SIFT features was adopted to determine possible tampering in *copy-pasted* forged image. The technique followed the idea that tampered and original region have quite similar key features. Once the key points are matched, agglomerative hierarchical clustering [46] is performed at the spatial locations of matched key points to detect the tampered region. After classification of image, they applied Random Sample Consensus algorithm to estimate homography. This homography is applied to all the matched key points and transformed key points are compared with original ones. So the homography method that yields higher number of inliers is considered as estimation method. They achieved a true positive rate of 100% for MICC-F220 dataset. SURF descriptors were used in methodology proposed by researchers in [13] and [44] for copy-move forgery. SURF descriptors proved to be rotation invariant so if the duplicate region is rotated and rescaled, key points will be same. In [13], SURF descriptors were computed and compared between images to localize tampering and in similar way after matching descriptors, nearest neighbors are identified by KD-trees searching algorithm in [44]. Visual Attention Model was used to propose a fixation point in methodology introduced by Qu et al. [41]. This visual cue was used to determine features from the spliced region of image. Primarily, all the efforts made for splicing detection and copy-move detection similar DCT, SURF, and SIFT techniques were used to detect forgery in images [16, 32, 33, 36]. The fact that image manipulation may alter the micro-texture in an image inspired researchers in [38] to employ Local Binary Pattern (LBP) and Steerable Pyramid Transform (SPT) and achieved higher detection performance.

2.2 Deep learning-based techniques

By the capability of DNN [43, 45, 53] to extract complex statistical dependencies from high-dimensional inputs and efficiently learning their hierarchical representations while being generalized for variety of computer vision tasks, passive image forensics techniques are being implemented through deep learning approaches. As in [42], 10-layer supervised CNN was trained to learn the hierarchical features of the tampered images with labelled regions. During training, the contents of images other than tampered regions were surpassed at first layer where the weights of first layer were initialized similar to the residual maps in Spatial Rich Model (SRM) [49]. The final discriminative fea-

tures were obtained by passing through pre-trained CNN and were fed into the SVM classifier for the binary classification of the image. Noise removal and image enhancement tool, the median filter is adopted by anti-forensics authorities to destroy correlation between pixels of manipulated images and to block the artifacts of an image that are left in JPEG compression.

Several work were presented to detect median filtering in images as part of forgery detection [15, 29, 54]. Chen et al. [17] have considered CNN for the median filter forensics with five convolutional layers and three fully connected layers to classify tampering. Moreover, deep learning techniques were employed in [14] to detect and localize tampered regions in the image. The authors in [14], proposed two methods: first method was an end-to-end system for the detection and localization through Radon transform [27] and CNN while second method was used to classify tampered region resampling features based on probability-maps (pmaps) and Long Short-Term Memory (LSTM). Universal image manipulations rather than specific copy-move and median filter, were targeted by work in [10]. They proposed a convolutional layer that learns manipulation detection features and outstrip whole image content. It was trained for both binary and multi-class classification like median filtering, Gaussian with additive white Gaussian blurring and re-sampling versus original image. Besides, copy-move and spliced regions detection, people have been working for detecting near-duplicates or semantically similar images where an image undergoes some geometrical transformations over time but does not change its semantic representations [19,23,31]. The methodology proposed by [4], utilized image segmentation by simple linear iterative clustering (SLIC), extracted features through VGGNet, and matched the keypoints by using adaptive patch matching. The resultant image marked the region of splicing by a block of red color. Moreover, authors in [28] used resnet-50 features and classfied images of CASIA-2.0 in their paper. In addition, a Mask-RCNN has also been proposed for the problem of splicing region detection by [5]. They have also utilized weights of pre-trained resnet-t0 as backbone feature extractor.

2.3 Image phylogeny techniques

Most of the traditional forgery detection techniques rely on analyzing isolated objects that were alerted. Interestingly, recent multimedia phylogeny approaches [21] analyze the whole evolutionary process that has influenced the probed image. In 2016 and 2017, NIST Nimble Challenge [2] was conducted to support the research in image and video forensics technologies. Using the dataset [2] published in 2016, Pinto et al. [40] performed a provenance analysis of altered images. The proposed methodology was composed of two stages as follows. Firstly, using image retrieval algorithm, the images with similar content were extracted from the database. The query images were compared for fast retrieval using SURF [9]. Once the similar images were fetched at first attempt of retrieval, a second step was the geometrical transformations for the best matched images with a query image leveraging homography. The

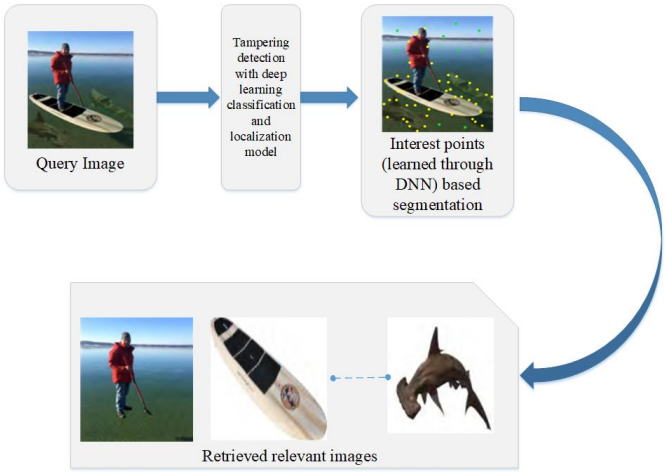


Fig. 1 Flow diagram of our proposed provenance filtering framework: A manipulated image is given as input to the system, then the system analyzes the input image and extracts the key features of an image and segmented regions. These key features are compared with a large image database, and approximate contributors are retrieved.

above resulted in obtaining the region that was spliced through any of the matched images. A mask was generated with the unmatched regions left in query image. Both of the steps were iterated up till 2 tiers. They achieved recall of 100% for host image and 67.71% for donor image using NC2016 [2] dataset with World1M [1] dataset. Similar, low-level interest points and local features as SURF were used by Moreira et al. [37]. To filter the query for the next-tier image retrieval task, a Reciprocal Condition Matching Measure (RCMM) was used. A large RCMM value indicated that the compared two images were nearly duplicate. Therefore, these images were surpassed from the next retrieval query.

The above has focused separately on solving forgery detection and handled provenance filtering tasks as another problem. Though both of these tasks add upto forensic analysis of multimedia. Moreover, provenance filtering task has been performed by using low level image features up till now as per our best knowledge. Our main contribution in this paper is to combine these tasks using the state-of-the-art deep learning techniques. Furthermore, for forgery detection and localization, our proposed model captures the strategies that integrate *copy-move* and *splicing* by post-processing operations such as blurring or retouching termed as *universal forgery detection*. Fig. 1 illustrates the process of multimedia forensic analysis demonstrated by this paper.

3 Proposed Methodology

We propose a multimodal system combining a DNN for forgery detection and localization with InceptionV3 [47] as a key feature extractor. Fig. 2 demon-

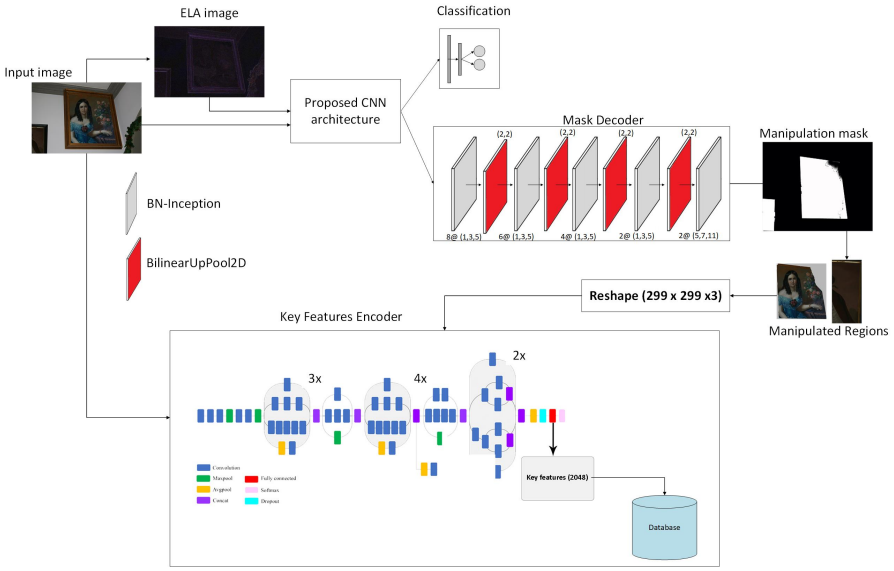


Fig. 2 Proposed Architecture for end-to-end provenance filtering task. Architecture explains the classification, detection and feature extraction steps.

strates the main integrants of our purposed methodology. If an image is classified as original then it is discarded at this initial step and no provenance is performed for this image. However, if a tampered image is encountered then the provenance filtering step is executed. In the following, we discuss the steps involved in our purposed methodology.

3.1 Forgery Detection and Localization

Our proposed end-to-end forgery detection and localization model is solution for universal manipulation techniques. In real world scenarios, when images are manipulated through copy-move or splicing, they are subjected to post processing methods for example blurring the sharp edges by employing median or Gaussian filter. Proposed network has been trained on diverse dataset which has enabled it to learn articrafts of manipulation even if these articrafts have been made obscrue by several operations. We propose a CNN-based network combined with Error Level Analysis (ELA) of an input image.

3.1.1 Error Level Analysis

Image format proves to be as informative as image itself. Hence, analysing the image format can help in forecasting about image manipulation. Error level analysis works on the intention that whenever an image is saved in lossy compression amount of error it introduces is not linear. When manipulation

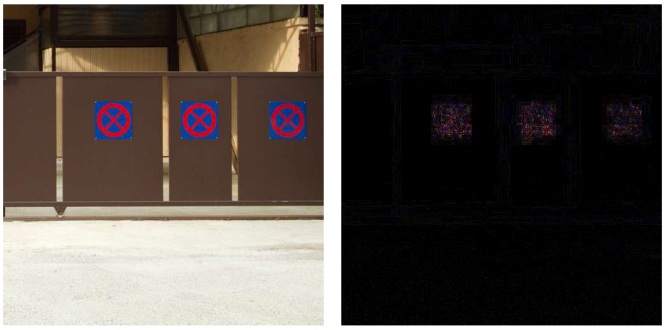


Fig. 3 Left is manipulated image and right depicts the ELA image.

is applied on image, manipulated region (8×8 blocks) reaches at different error rate than unmodified regions. This way error level analysis enhances the regions with manipulation which enable convolution neural networks to learn articrafts of forgery.

The ELA is performed by re-saving image into disc at some known error rate which is 90% in our case. Difference is computed between input image and resaved image. Once the difference matrix is extracted, local extrema are computed. Local maxima gives the value of maximum difference in two images. Finally, image brightness is enhanced by a scale of maximum difference. The ELA operation can be expressed as

$$I_{\text{ELA}} = \alpha \times (I_{\text{IN}} - I_{\text{RESAVED}}) \,, \tag{1}$$

where $\alpha = \dfrac{255}{\max(I_{\text{IN}} - I_{\text{RESAVED}})}$. Where I_{IN} is input image and I_{RESAVED} is resaved I_{IN} at 90%. Fig. 3 illustrates an example for an image after resaving at 90% and its ELA.

3.1.2 Proposed Deep Architecture

The architecture of our proposed DNN model is demonstrated in Fig. 4. Our proposed network architecture consists of 2 convolutional layers and 1 pooling layer. Network is then splited into two streams, one is forgery detection and other stream is forgery localization. Two dense layers with softmax activation at last layer are combined with the max pooling layer to output probabilities of positive and negative class. In parallel to dense layers, CNN features are up-sampled through alternative set of BN-Inception blocks and `BilinearUpPool2D` [50] layers called Mask Decoder. As a result of these, mask of manipulated region is obtained so-called, forgery localization.

Following are the characteristics of proposed architecture:

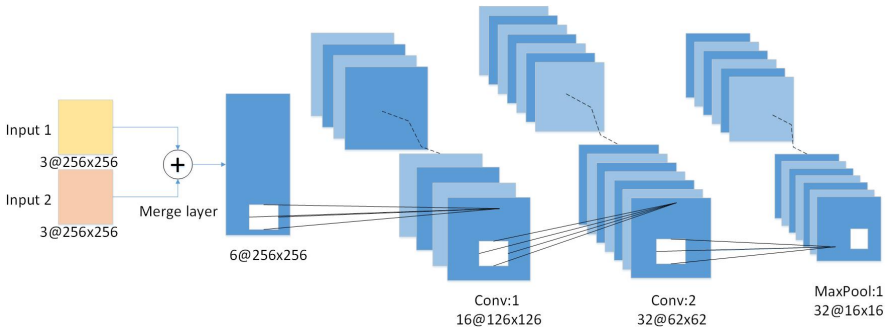


Fig. 4 Proposed CNN network for feature extraction.

1. Two input tensors are passed to the CNN of size $256 \times 256 \times 3$ both, where 256×256 is the height & width of ELA image and RGB image, 3 denotes the number of channels (red, green, and black).
2. Very first layer of our proposed CNN is merge layer. It takes two tensors of $256 \times 256 \times 3$ as input and outputs a tensor of $256 \times 256 \times 6$.
3. First convolution layer has 32 filters with receptive field of 5×5.
4. Second convolution later has 64 filters with 5×5 receptive field as well.
5. Both convolutional layers are activated by using Rectified Linear Units (ReLU) which respond selectively to the useful input signals.
6. Second convolutional layer is followed by non-overlapping, max-pooling layer of 2×2 filter size. Max-pooling layer discards 75% of input activations and reduces the size of feature map.
7. A dropout of 0.25 is applied to generalize the convergence. Dropout ignores the neurons with probability less than 0.25.
8. A dense layer with ReLU changes the 4D feature map to 256 features.
9. 256 features obtained from dense layer are fed to another dense layer which computes the probability scores for 2 classes using softmax activation.
10. Output from step 5 is sent to mask decoder block. Which converts feature map of size $16 \times 16 \times 32$ to manipulated region mask of size 256×256.

Mask decoder block has been shown in Fig. 2.

- BN-Inception is basic module of Google Inception model with batch normalization and ReLU activation. Intuition of inception model to be wider while learning local as well as global image features has been utilized to analyse articrafts of forgery. This small BN-Inception network consists of three Conv2D layers in parallel described as $n@[s_1, s_2, s_3]$. Here, n denotes the number of filters for each layer and s_1, s_2 and s_3 denote the filter size of three Conv2D layers. Batch normalization is applied on concatenated output of these Conv2D layers and final output is activated using ReLu.
- Mask decoder block enhances the feature map obtained from CNN to 16 times by using `BilinearUpPool2D` layer of size 2×2. Each one of the `BilinearUpPool2D` layers brings 2 times rise in the width and height of input feature map. To obtain the manipulated mask of input image size,

`BilinearUpPool2D` layer has been injected four times with alternative BN-Inception block in mask decoder. For a filter size of $a \times b$, output size after `BilinearUpPool2D` layer can be defined as:

$$X_{new}, Y_{new} = a \times X_{old}, b \times Y_{old} \tag{2}$$

- Output at last BN-Inception block is $256 \times 256 \times 6$, having 6 channels due to the number of filters which are 2 for each Conv2D layer.
- This tensor of $256 \times 256 \times 6$ is processed through last Conv2D layer having filter of size $3 \times 3 \times 1$ with *sigmoid* as an activation function. This layer outputs the mask of 256×256.

3.2 Provenance Filtering

Once the query image is classified as forged and manipulated region is localized, relevant images are extracted from database based on image semantics called Provenance Filtering. Relevant images are those which contributed towards creation of manipulated image i.e. donors or which are semantically similar to query image i.e. nearly duplicate. Given a query image I which has been classified as forged or tampered image from a repository of images C , a set of donors and nearly duplicate r_{nn} are retrieved. To trace the remnants of I with time, retrieval of nearly duplicate images of I is also essential. Our proposed provenance filtering method demonstrate the robustness of convolutional features in two tiers. in first tier we retrieve the images that are more likely to entire image while in second tier we obtain images from databases which are likely to contain the manipulated regions.

Convolutional features for image representation are extracted by exploiting InceptionV3 model trained on ImageNet. This dataset is composed of wide variety of 1000 classes, such as animals, objects, and events. The NIST dataset used for provenance filtering is not drastically different from ImageNet. Therefore, the knowledge from network trained on this dataset is useful. Note that Google's InceptionV3 network has been utilized for many classification problems. Inception module applies different sized convolutions on same input and stack them up at output. This way allowing the network to choose right path. As Fig. 5 depicts the basic structure of one inception module, our problem best fits with this type of convolutional network. Manipulation somehow deteriorates the underlying structure of images at pixel level as well as at patch level. Convolution filters of size 1×1 learns pixel level facts while 3×3 and 5×5 gives patch level feature representation to input image. In InceptionV3, 5×5 and 3×3 convolutions have been factorized into two 3×3 and 1×3 after 3×1, respectively. This factorization performs the similar feature extraction to InceptionV1 while reducing number of parameters. Architecture of InceptionV3 is summarized in Table 1. The classification layer of InceptionV3 is declined and 2048 transfer values are considered as rich image representation.

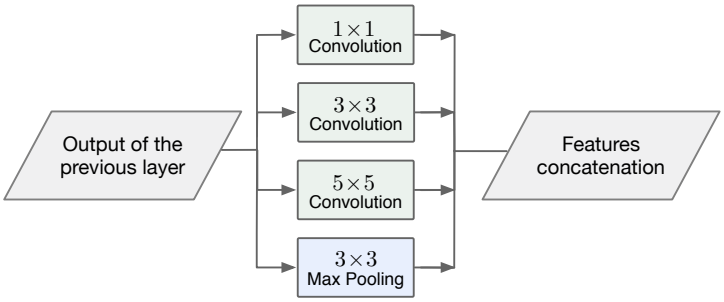


Fig. 5 Basic inception module.

Table 1 InceptionV3 Architecture

Layer	**Filter size**	**Stride**	**Input Size**
Conv2D	3×3	2	299×299×3
Conv2D	3×3	1	149×149×32
Conv2D + Padding	3×3	1	147×147×32
MaxPool2D	3×3	2	147×147×64
Conv2D	3×3	1	73×73×64
Conv2D	3×3	2	71×71×80
Conv2D	3×3	1	35×35×192
3×Inception	As shown in key feature encoder block fig. 2	–	35×35×288
4×Inception	As shown in key feature encoder block fig. 2	–	17×17×768
2×Inception	As shown in key feature encoder block fig. 2	–	8×8×1280
MaxPool2D	8×8	1	8×8×2048
Flatten	logits	-	1×1×2048
Output Size = 2048			

3.3 Database Indexing and Retrieval

Image encodings generated through InceptionV3 feature extractor are indexed in databases. These feature encodings are calculated for entire image as well as for manipulated region in case of manipulation. For a collection of C images, features set collection can be described as,

$$F_i \mid i \in Img_{\text{ind}} = \begin{cases} 0, 1, \ldots |C| \times p \times n & \text{if manipulation} \\ 0, 1, \ldots |C| \times p & \text{otherwise,} \end{cases} \tag{3}$$

where $p = 2048$ and n denotes the number of disjoint manipulated regions. Img_{ind} is the set of image indexes. Once image features in database are indexed, search procedure is performed via feature-wise query.

Algorithm 1: Image Forgery Detection with Provenance Filtering

```
Input Image i, Image Features I_i & Manipulated Region Features M_i ;
Output Class cls and a set of contributors S_i
cls,masks = ForgeryDetection(i)
S_i = []
if cls=1 then
    for m in masks do
    end
end
```

To retrieve relevant images, the key features points learned through convolutional neural networks are obtained on query image. We combine the input image with multiple segments that improve the retrieval task. In fact, the performance and effectiveness of retrieval are increased by indexing with respect to a small region of the image. For a query image Q, if an image is classified as *forged* image in the initial classification module. The manipulated region of the query image is obtained form its manipulated mask. The key features representation is created from both images and submitted to the framework. Afterward, the proposed framework returns the indices of top-k similar images to the query image using nearest neighbour algorithm. At the second level of retrieval, the indices of top-k images which are similar to the manipulated region of the images are returned. A pseudo code is depicted in Algorithm 1.

4 Experimental Results

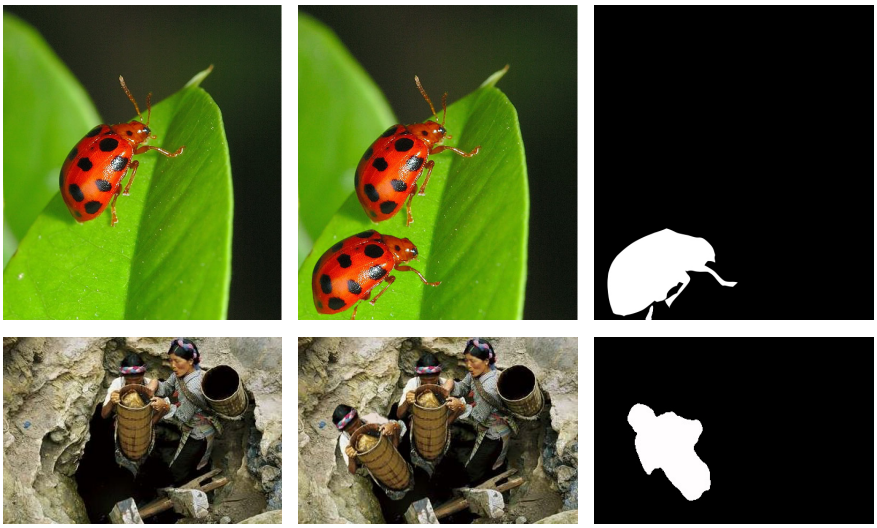
In this section, we present the experimental results to validate the performance of the proposed methodology. We show the effectiveness of the results in terms of a) percentage accuracy for forgery detection, b) precision, and c) recall measures using *pixel-level* evaluation for the manipulated mask decoder and *recall@k* for provenance filtering that measures the ratio of accurate images among the top-k retrieved results.

4.1 Dataset

We have used three benchmark datasets for the forgery detection and localization problem and one for the provenance filtering. Details of these datasets are discussed as follows.

4.1.1 CASIA V2.0 Dataset

CASIA V2.0 [22] is the dataset containing color images with realistic tampering operations including *copy-move* and *splicing*. This dataset has been widely used in image forensics problems. Basically, this dataset contains labelled images for binary classes, i.e., authentic and tampered. This dataset



(a) Original images (b) Tampered images (c) Manipulation mask

Fig. 6 Samples from CASIA V2.0 dataset [22] with simulated masks.

contains total 7,200 authentic images and 5,123 tampered images of a resolution ranging from 320×240 to 800×600. However, this dataset does not provide manipulation masks ground truth. We have followed the procedure described in [52] to generate the ground truth masks which is to subtract a tampered image with its corresponding original image and further verified the mask by an human intervention. A few examples are illustrated in Fig. 6.

4.1.2 CoMoFoD

CoMoFoD dataset [48] differs from other multimedia forensics datasets due to its suitability for the evaluation of post-processing methods, such as smoothing, color reduction, and noise addition. Furthermore, distortion and combination have also been added as broad categories of manipulation. For the images of resolution 512×512, each of manipulation category has 40 images. After applying post processing methods to both original and fake images, a total of 10,400 images are prepared. These ground truth masks demonstrate the localization of source region as well as manipulated region. Therefore, for our problem, we change the masks to only manipulated localized region using similar method used for CASIA V2.0 dataset. Fig. 7 shows a few samples from CoMoFoD dataset [48].

4.1.3 MFC 2018 NIST Challange

We take the Media Forensics Challenge 2018 (MFC18) dataset which has been provided by the National Institute of Standards and Technology (NIST) [3].

Fig. 7 Samples from CoMoFoD dataset [48] with the simulated masks. (a) Original images, (b) Fake images, (c) Provided manipulation mask, and (d) Generated manipulation mask.

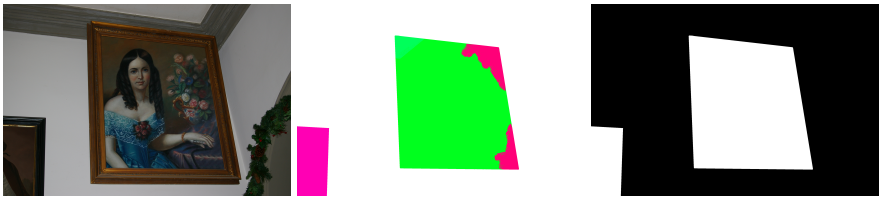


Fig. 8 Example from MFC 2018 dataset [3] for the manipulation detection. (a) Query Image, (b) Manipulation mask, and (c) Binary Manipulation mask.

This dataset focuses on the media forensics tasks, provenance filtering, and phylogeny problems. For our media forensics problem, we use this dataset for manipulation detection as well as for provenance filtering. For manipulation detection task it contains $6,700$ images for manipulation detection task and $3,000$ images for splicing detection task. It consists of images with varying resolutions. Images from this dataset is presented in Fig. 8.

Provenance filtering domain of this dataset comprise a query set which contains different types of manipulated images such as splicing, compositions and copy-move etc, and a world image set which contains the source images employed to produce probe/query images, nearly duplicates and distractors. The probe image set of MFC2018 dataset contain $6,707$ images and world set contains $13,672$ images. Fig. 9. shows an example of query image with its corresponding source images and nearly duplicate images.

(a) Query/probe image

(b) Nearly duplicates (c) Components (d) Components

Fig. 9 Example from MFC 2018 dataset [3] for provenance filtering.

4.2 Forgery Detection and Localization

We discuss the experimental setup and performance of our proposed network. We have trained our multi-inputs and multi-outputs network by performing end-to-end optimization. Network converges after 500 epochs with learning rate of $10e^{-4}$. Moreover, the RMSprop algorithm is employed to optimize the network weights. RMSprop algorithm can be defined as follows.

$$w_{new} = w_{old} - \frac{\eta}{\sqrt{E\left[g^2\right]_{new}}} . \frac{\partial f}{\delta w} \ , \tag{4}$$

where f is the cost function, $\frac{\partial f}{\delta w}$ is the gradient of f with respect to weight w. w_{new} is updated value of previous weight w_{old}. RMSprop embraces the idea of adaptive learning rate with moving average squared gradients $\mathbb{E}\left[g^2\right]$. Particularly, the adaptive learning rate makes faster convergence of the network.

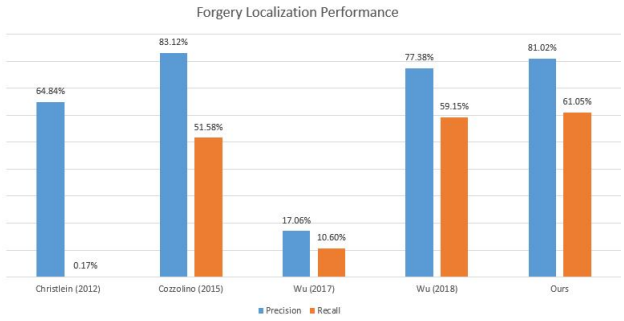


Fig. 10 Comparison of manipulation detection performance with available methods.

Table 2 Performance of forgery detection and localization module for different combinations of datasets

Dataset	**Classification Performance (Accuracy)**	**Localization Performance (Precision)**	**Forgery Technique**
CASIA-V2	93.04%	85.47%	Copy-Move, Splicing
CoMoFoD + CASIA-V2	88.90%	80.63%	Copy-Move, Splicing, Resampling, Median Filters, Contrast enhancement, Blurring
CoMoFoD + CASIA-V2 + NIST 2018	89.01%	81.02%	Copy-Move, Splicing, Resampling, Median Filters, Contrast enhancement, Blurring

We have combined three datasets for training of the network to tackle the universal forgery techniques. The dataset has been divided into training, validation, and test sets with a ratio 80%, 10%, and 10%, respectively. Table 2 summarizes the results of the proposed methodology for forgery classification and manipulation detection. Fig. 10 shows that our proposed methodology outperforms other localization techniques [18], [20], [51] in terms of recall. *Pixel-level* recall measures the correctness of retrieved pixels according to the pixels of manipulated regions. The precision of the methodology presented by [20] is slightly higher than the precision achieved by our model. Because, to achieve higher recall value that tends to localize all pixels that are the part of manipulated region, a trade-off exists between between precision and recall.

4.3 Provenance Filtering

We now evaluate the proposed approach of provenance filtering task, we have used MFC18 NIST dataset [3]. Probe repository in this dataset contain composite images which are used as queries to perform experiments. Some of these composite images overlap with the training data for manipulation detection,

Table 3 End-to-end System Evaluation

Method	Dataset	Classification Accuracy	Localization Recall	Provenance Filtering (End-to-end)
Pinto et. al [40]	NIST 2016 World1M	–	–	R@10 83.5%
Moreira et. al [37]	NIST 2017	–	–	R@50 90.7%
Proposed	MFC NIST 2018	89.01%	61.05%	R@50 92.8%

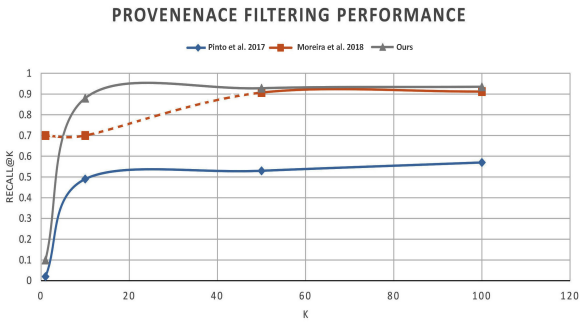


Fig. 11 Performance comparison of provenance filtering with recent approaches.

hence gives negligible error in manipulation localization. However, unseen composite images also achieve good manipulation results as discussed in previous section. Experimental results for retrieval from large image databases shows that the encoded key features extracted through InceptionV3 prove to be robust for near duplicate(with disparate compositions) retrieval as well as for retrieving origin of manipulated regions.

An end-to-end system performance is presented in Table 3. Recall@k with $k = \{1, 10, 50, 100\}$ is chosen to measure the performance of provenance filtering algorithm. Recall@k means that ratio of relevant images have been measured for top k retrieval results. Retrieval results are verified using references provided in MFC18 references. Fig. 11. shows the performance of our proposed methodology with recent techniques. Results shows that recall is significant for smaller values of k, this is because of considering nearly duplicate images that have global features correlating to query image. At second stage, incorporating the manipulated region based filtering helps in achieving higher recall values. Results depict that our proposed methodology outmatches the recent techniques.

5 Conclusion

In this paper, we proposed an entire framework for multimedia forensics task. We have developed an end-to-end system that combines three media forensics tasks, i.e, classification of image as authentic or fake image for a manipulated image, localization of manipulated region, and provenance filtering which is retrieval of manipulation donors from a large set of images. We have presented a robust deep learning-based universal network to classify the type of a forged image. This network is further enhanced to localize the manipulated region in case of *copy-move* and *splicing* forgery techniques. We have proposed an end-to-end trainable for forgery detection and localization using RGB image as well as ELA image. For provenance filtering, unlike previous efforts, to capture the retrieval of small manipulated regions, we have incorporated forgery localization. Rich key features are extracted and host, nearly duplicate and donor filtration is executed in two steps. We have employed Inception-V3 to extract key deep features due to out performance of deep learning mechanisms over traditional features. Key features of query image and its manipulated regions are compared with indexed features of world image set in databases using K-nearest neighbour algorithm. This is upto our knowledge first ever system which incorporates forgery detection techniques with provenance filtering as well as performs provenance filtering task with deep features. We have achieved performance in terms of accuracy, precision and recall comparable to the state of the art methods using NIST MFC18 dataset. Although, a separate learning is employed to localize the forged region, in future we will focus on multistream single network to extract key features of forged part from the localization model. This work can also be extended to find the similarity between images through key points retrieval using deep learning techniques. Moreover, it is worthwhile to investigate the forensics task in several multimedia applications, such as IPTV [26].

References

1. Medibrowser. URL http://medifor.rankone.io/
2. Nist media forensics challenge (2016). URL https://www.nist.gov/itl/iad/mig/media-forensics-challenge. Accessed: 2018-07-26
3. Nist media forensics challenge (2018). URL https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018. Accessed: 2019-06-11
4. Agarwal, R., Verma, O.P.: An efficient copy move forgery detection using deep learning feature extraction and matching algorithm. Multimedia Tools and Applications pp. 1–22 (2019)
5. Ahmed, B., Gulliver, T.A., et al.: Image splicing detection using mask-rcnn. Signal, Image and Video Processing pp. 1–8 (2020)
6. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE transactions on Computers **100**(1), 90–93 (1974)
7. Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., Serra, G.: A sift-based forensic method for copy–move attack detection and transformation recovery. IEEE Trans. Informat. Forensics and Security **6**(3), 1099–1110 (2011)
8. Asghar, K., Habib, Z., Hussain, M.: Copy-move and splicing image forgery detection and localization techniques: a review. Taylor & Francis Australian Journal of Forensic Sciences **49**(3), 281–307 (2017)

9. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European conference on computer vision, pp. 404–417. Springer (2006)
10. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10. ACM (2016)
11. Bhatti, M.H., Khan, J., Khan, M.U.G., Iqbal, R., Aloqaily, M., Jararweh, Y., Gupta, B.: Soft computing-based eeg classification by optimal feature selection and neural networks. IEEE Transactions on Industrial Informatics **15**(10), 5747–5754 (2019)
12. Birajdar, G.K., Mankar, V.H.: Digital image forgery detection using passive techniques: A survey. Elsevier Digital investigation **10**(3), 226–245 (2013)
13. Bo, X., Junwen, W., Guangjie, L., Yuewei, D.: Image copy-move forgery detection based on surf. In: Proc. IEEE Int. Conf. on Multimedia Information Netw. and Security, pp. 889–892 (2010)
14. Bunk, J., Bappy, J.H., Mohammed, T.M., Nataraj, L., Flenner, A., Manjunath, B., Chandrasekaran, S., Roy-Chowdhury, A.K., Peterson, L.: Detection and localization of image forgeries using resampling features and deep learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1881–1889. IEEE (2017)
15. Cao, G., Zhao, Y., Ni, R., Yu, L., Tian, H.: Forensic detection of median filtering in digital images. In: 2010 IEEE International Conference on Multimedia and Expo, pp. 89–94. IEEE (2010)
16. Cao, Y., Gao, T., Fan, L., Yang, Q.: A robust detection algorithm for copy-move forgery in digital images. Forensic science international **214**(1-3), 33–43 (2012)
17. Chen, J., Kang, X., Liu, Y., Wang, Z.J.: Median filtering forensics based on convolutional neural networks. IEEE Signal Processing Letters **22**(11), 1849–1853 (2015)
18. Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E.: An evaluation of popular copy-move forgery detection approaches. IEEE Transactions on information forensics and security **7**(6), 1841–1854 (2012)
19. Chum, O., Philbin, J., Zisserman, A., et al.: Near duplicate image detection: min-hash and tf-idf weighting. In: BMVC, vol. 810, pp. 812–815 (2008)
20. Cozzolino, D., Poggi, G., Verdoliva, L.: Efficient dense-field copy–move forgery detection. IEEE Transactions on Information Forensics and Security **10**(11), 2284–2297 (2015)
21. Dias, Z., Goldenstein, S., Rocha, A.: Toward image phylogeny forests: Automatically recovering semantically similar image relationships. Forensic science international **231**(1-3), 178–189 (2013)
22. Dong, J., Wang, W., Tan, T.: Casia image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing, pp. 422–426. IEEE (2013)
23. Dong, W., Wang, Z., Charikar, M., Li, K.: High-confidence near-duplicate image detection. In: Proceedings of the 2nd acm international conference on multimedia retrieval, p. 1. ACM (2012)
24. Farid, H.: Image forgery detection. IEEE Signal Process. Mag. **26**(2), 16–25 (2009)
25. Fridrich, A.J., Soukal, B.D., Lukáš, A.J.: Detection of copy-move forgery in digital images. In: in Proceedings of Digital Forensic Research Workshop. Citeseer (2003)
26. Garcia, M., Canovas, A., Edo, M., Lloret, J.: A QoE management system for ubiquitous IPTV devices. In: The Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM) (2008)
27. Helgason, S., Helgason, S.: The radon transform, vol. 2. Springer (1999)
28. Jaiswal, A.K., Srivastava, R.: Image splicing detection using deep residual network. Available at SSRN 3351072 (2019)
29. Kang, X., Stamm, M.C., Peng, A., Liu, K.R.: Robust median filtering forensics using an autoregressive model. IEEE Transactions on Information Forensics and Security **8**(9), 1456–1468 (2013)
30. Khan, M.Z., Harous, S., Hassan, S.U., Khan, M.U.G., Iqbal, R., Mumtaz, S.: Deep unified model for face recognition based on convolution neural network and edge computing. IEEE Access **7**, 72622–72633 (2019)
31. Li, M., Wang, B., Ma, W.Y., Li, Z.: Detecting duplicate images using hash code grouping (2010). US Patent 7,647,331

32. Lin, S.D., Wu, T.: An integrated technique for splicing and copy-move forgery image detection. In: Proc. IEEE 4th International Congress on Image and Signal Processing, vol. 2, pp. 1086–1090 (2011)
33. Lin, Z., He, J., Tang, X., Tang, C.K.: Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis. Pattern Recognition **42**(11), 2492–2501 (2009)
34. Lloret, J., Bosch, I., Sendra, S., Serrano, A.: A wireless sensor network for vineyard monitoring that uses image processing. Sensors **11**(6), 6165–6196 (2011)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
36. Mahmood, T., Mehmood, Z., Shah, M., Saba, T.: A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform. Journal of Visual Communication and Image Representation **53**, 202–214 (2018)
37. Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K.W., Flynn, P.J., Rocha, A., Scheirer, W.J.: Image provenance analysis at scale. IEEE Transactions on Image Processing **27**(12), 6109–6123 (2018)
38. Muhammad, G., Al-Hammadi, M.H., Hussain, M., Bebis, G.: Image forgery detection using steerable pyramid transform and local binary pattern. Machine Vision and Applications **25**(4), 985–995 (2014)
39. Murabayashi, A.: The problem of fake photos in fake news (2018). URL https://petapixel.com/2017/01/19/problem-fake-photos-fake-news/. Accessed: 2019-01-11
40. Pinto, A., Moreira, D., Bharati, A., Brogan, J., Bowyer, K., Flynn, P., Scheirer, W., Rocha, A.: Provenance filtering for multimedia phylogeny. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1502–1506. IEEE (2017)
41. Qu, Z., Qiu, G., Huang, J.: Detect digital image splicing with visual cues. In: Springer Int. workshop on Information hiding, pp. 247–261 (2009)
42. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE (2016)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). DOI 10.1007/s11263-015-0816-y
44. Shivakumar, B., Baboo, S.S.: Detection of region duplication forgery in digital images using surf. International Journal of Computer Science Issues (IJCSI) **8**(4), 199 (2011)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
46. Steinbach, M., Karypis, G., Kumar, V., et al.: A comparison of document clustering techniques. In: KDD workshop on text mining, vol. 400, pp. 525–526. Boston (2000)
47. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)
48. Tralic, D., Zupancic, I., Grgic, S., Grgic, M.: Comofod—new database for copy-move forgery detection. In: Proceedings ELMAR-2013, pp. 49–54. IEEE (2013)
49. Wang, P., Wei, Z., Xiao, L.: Pure spatial rich model features for digital image steganalysis. Multimedia Tools and Applications **75**(5), 2897–2912 (2016)
50. Wojna, Z., Ferrari, V., Guadarrama, S., Silberman, N., Chen, L.C., Fathi, A., Uijlings, J.: The devil is in the decoder. arXiv preprint arXiv:1707.05847 (2017)
51. Wu, Y., Abd-Almageed, W., Natarajan, P.: Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 1480–1502. ACM (2017)
52. Wu, Y., Abd-Almageed, W., Natarajan, P.: Busternet: Detecting copy-move image forgery with source/target localization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 168–184 (2018)

53. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Prof. Advances in neural information processing systems, pp. 3320–3328 (2014)
54. Yuan, H.D.: Blind forensics of median filtering in digital images. IEEE Transactions on Information Forensics and Security **6**(4), 1335–1345 (2011)