JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IMAGE FORGERY DETECTION USING OPEN-CV AND MD5

A. Vadivelu, Asst Professor, Department of CSE, SRM IST Ramapuram campus, Chennai, India.

M.Lohithdakshan, Maithili Jha, S. Maitri, Students, Department of CSE, SRM IST Ramapuram campus, Chennai, India.

Abstract*— Digital images play a major role in the digital sector, but image counterfeiting is a developing worry that threatens the legitimacy of data. Digital photographs were intended to increase accessibility and efficiency, however some users have misused the system. Although there are conventional techniques for identifying phony photographs, more sophisticated methods are needed due to the rising incidence of image counterfeiting. Although OpenCV and MD5-based techniques have improved the detection of picture fraud, they are not as effective in identifying all forms of fraud. Efficiency must therefore be developed and enhanced.

Keywords*— Digital images, Image forgery, Authenticity, Traditional methods, OpenCV, Fraud detection, MD5.*

I. INTRODUCTION

Thanks to technological advancements and globalization, digital cameras are now widely available and reasonably priced. People so use a variety of camera sensors to take and gather a large number of photos, which are then frequently saved in soft copy for use in online documents and social media sharing. Because they are readily available and can be understood by people who lack literacy, images are crucial for the storage and distribution of data in the digital age. Several tools for image editing are accessible; these tools were first created to improve the quality of the pictures. Nevertheless, some people abuse these resources to disseminate misleading information and produce fraudulent photos. Because of the serious and frequently irreversible effects of altered images, this is a serious concern.

The purpose of this notion is to lessen the amount of forgeries that occur and assist in distinguishing between real, legitimate images and fakes. This will make identification easier and more accurate, improving efficiency.

Image splicing and copy-move are the two categories under which image forgeries fall.

Image Splicing: A sort of digital picture forgeries known as "image splicing" involves combining one or more sections from separate photos to create a new one. This is frequently done to produce a composite image that combines elements from various sources to look authentic. Sophisticated methods are needed to detect image splicing, such as examining variations in the texture, color, and lighting of various image regions [1][5].

Copy-Move: In order to conceal or duplicate content, a portion of an image is copied and pasted into a different area inside the same image. This technique is known as copy-move image forging. This method is frequently applied to remove undesired aspects from an image or to make many instances of a person or object. Copy-move forgeries and duplicate elements in an image can be detected using advanced digital image processing techniques such content-based picture retrieval, feature extraction, and pattern recognition [6].

The following is how the rest of the essay is structured: Section 2 discusses significant studies on image forgery detection using image-splicing convolutional neural networks. Section 3 describes the study's approach, as well as the research, and explains the mathematical principles used in this work. Finally, we discussed the analysis of our project's results in Section 4.

II. LITERATURE SURVEY

Xiao et al [1]

They offered a two-pronged method for identifying image splicing forgeries. It extracts variations in picture attributes between tampered and untampered regions using an adaptive clustering architecture and a C2RNet. In comparison to state-of-the-art techniques, the suggested method yields promising results and detects splicing frauds successfully.

Furthermore, the suggested methodology exhibits computational efficiency and yields encouraging outcomes across various assault scenarios. C2RNet and adaptive clustering work together to understand the differences in image attributes between tampered and untampered sections, enabling accurate detection of splicing forgeries. All things considered, the technique offers a reliable way to identify image splicing fake.

Kwon et al [2]

A completely convolutional neural network designed for photo splicing localization is called CAT-Net. In order to learn the

forensic characteristics of compression artifacts in both domains, the network integrates RGB and DCT streams. While the DCT stream is pre-trained on double JPEG detection to take use of JPEG artifacts, the RGB stream uses resolutions to deal with the shapes and sizes of the spliced item. The proposed method is a valuable tool against malicious image forgeries, as it performs better than the state-of-the-art neural networks in both JPEG and non-JPEG image localization.

Zheng et al[3].
A survey on photo tampering and how to spot it in real-world photos found that while it's now fairly simple to edit images using software, it appears suspicious if an object has been hidden or a person's face has been altered. It is vital to ascertain whether aspect of the image has been altered prior to scrutinizing their intentions. This calls for the creation of automated tools that can discern between real and altered images. This article examines common image modification techniques, previously accessible datasets of modified images, and novel techniques for detecting tampering. Additionally, it offers a fresh perspective on reevaluating the presumptions of tampering clues supporting various detection systems, encouraging the research community to develop general techniques for tampering localization instead of relying just on single-type tampering detection.

R. Shao et al [4]
The system for identifying altered areas in scanned photos through deep learning methods is presented in the study. The system is trained using well-known convolutional neural network designs, such as InceptionV3, Resnet34, and Xception Net, on a dataset of more than 3,800 scanned photos from 169 distinct scanner models. Any areas of the image that might have been altered are highlighted on the dependability map that the system creates. It distinguishes between features of various scanner models and detects any areas that might have been altered by utilizing sophisticated deep-learning techniques and a sizable dataset of scanned images.

Several authors [7-11] contributed to the development of deep learning and machine learning models to anticipate forgeries using techniques such as Convolutional Neural Networks and Support Vector Machines. These algorithms have yielded encouraging results in detecting forgeries, highlighting the possibility of automated image forgery detection systems. However, further study is needed to investigate and compare the performance of various algorithms and their combinations to produce more accurate and dependable image forgery detection models.

Experiments show that the suggested method detects picture forgeries faster than current methods, both in terms of image detection time and efficiency. This is a viable solution to the spreading problem of image fraud and has potential uses in forensics, security, and media authentication, among many other fields.

To reduce noise and boost image quality, convert the picture to grayscale, resize it to a standard size, and apply any necessary filtering or enhancement methods. OpenCV may be used to extract texture, color, and shape from images. Methods like histograms of oriented gradients (HOG), scale-invariant feature transform (SIFT), and local binary patterns (LBP) might be used to achieve this. .

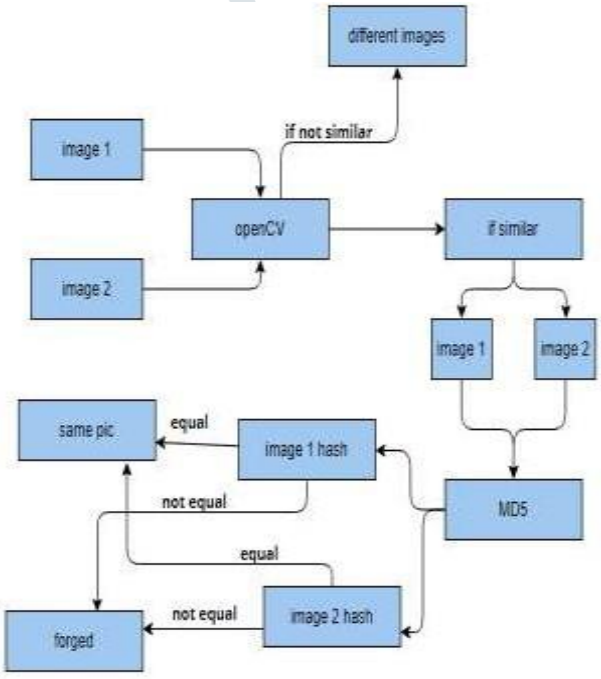


Figure 1. Architecture

III. METHODOLOGY

In digital forensics, image forgery detection is a method used to identify if an image or report is fake. During this process, the scanned document's properties are analyzed and distinctive features that can be utilized for identification are extracted using computer vision libraries like OpenCV. In addition, the scanned document's distinct digital fingerprint is produced by the MD5 hashing technique and can be compared.

The suggested technique uses an open-source computer vision and machine learning software library to extract image features, which are then run through a hash function to produce a digital signature. To find any discrepancies that would point to picture manipulation, the signature is compared to the original signature. A dataset of real-world images that had undergone different manipulations, such as copy-move, splicing, and removal, was used to test this.

In this study, the histogram approach is being used to calculate the images' grayscale. generating the image's MD5 hash value. This will help us decide which image is superior and act as a distinctive digital signature for it. comparing the hash value of the possibly manipulated image to the hash value of the original image. A substantial discrepancy between the two hash values suggests that the image has been altered. Lastly, present the analysis's findings, mentioning the position and size of any tampered regions in addition to the degree of confidence in the detection. For legal or investigative purposes, this information may be used.

A. Dataset Explanation

We are use the CASIA public dataset, which includes both authentic and altered photos, for detection. We can compare any two images and determine whether or not they are forged by using the openCV image processing library. A popular benchmark in image forgery detection research is the CASIA dataset. It contains a large number of original photos in addition to modified versions created through a variety of picture altering methods. Both the individual photos and the characteristics or attributes that define each image are referenced in the rows and columns of the dataset. A single image is represented by each row, and a different feature or characteristic of that image is represented by each column. Image resolution, color distribution, texture, and edge information are some of these attributes.

Assuming that every image has a unique hash value, we are using the CASIA dataset to select text and edge features for our project. We will then compare the features by computing the grayscale from the text feature and the histogram from the edge features. We grab a single picture from the system itself instead of importing the dataset and checking it for forgeries as we are using the hashing function md5.

The techniques used in the CASIA dataset for image forgery detection with OpenCV and MD5 are designed to identify several types of image manipulation, including as retouching, splicing, and copy-pasting.

These methods entail altering an image's color, texture, or edges, among other elements, to create an authentic-looking fake. These kinds of alterations can be found and photo forgeries can be identified by looking at an image's visual components and comparing its hash to that of an original.

B. Techniques Used

1. OpenCV (Open-Source Computer Vision Library):

It is an open-source computer vision and machine learning software library that is widely used in academic and industrial research and development for real-time image and video processing, object detection and tracking, face recognition, and many other applications. It was created by Intel in 1999 and is now maintained by the OpenCV.org community.

OpenCV is written in C++ and includes a C++ API as well as APIs for Python, Java, and other programming languages. It includes over 2,500 optimized algorithms for image and video capture, filtering, segmentation, feature detection and extraction, object detection and recognition, face detection and recognition, tracking, and 3D reconstruction [12].

OpenCV's main features include:

Cross-platform compatibility: OpenCV can run on Windows, Linux, Mac OS, iOS, and Android.

High performance: OpenCV is designed to take advantage of multi-core CPUs and GPUs to achieve real-time performance.

Simple APIs: OpenCV offers both high-level APIs for basic computer vision tasks and low-level APIs for more complex customization. OpenCV has a large and active community that contributes to its development and maintenance, as well as many third-party libraries and tools that integrate with OpenCV.

Many real-world applications rely on OpenCV, including self-driving cars, security and surveillance systems, medical imaging, robotics, augmented reality, and video game development.

2. MD5:

MD5 (Message Digest 5) is a widely used cryptographic hash function that generates a fixed-size, unique hash value from an arbitrary-length message. Ronald Rivest created it in 1991 as a replacement for the earlier MD4 algorithm.

The MD5 algorithm processes an input message through a series of mathematical operations to produce a fixed-size, 128-bit hash value. This hash value is unique to the input message, which means that any changes to the input message will result in a different hash value.

MD5 is widely used for integrity checking, authentication, and data validation in a variety of applications. For example, it is often used to hash passwords in online applications to ensure that the password is not stored in plaintext and cannot be easily recovered if the password database is stolen. It is also employed in digital signatures to ensure that a message has not been tampered with or altered during transmission [13].

3. Django Web Framework:

Django is a high-level Python web framework that adheres to the model-view-controller (MVC) architectural pattern. It has a sleek and practical architecture that allows you to construct web applications quickly and with less code. Django's key strengths are its stability, scalability, and adaptability. It includes several built-in capabilities such as an ORM (Object-Relational Mapping) for interfacing with databases, a sophisticated template engine, built-in authentication, and a complete admin interface.

Django is also highly extendable, with various third-party packages and plugins available to extend its capability. Many famous websites use it, including Instagram, Pinterest, and Mozilla.

Overall, Django is a robust and adaptable web framework that enables developers to create sophisticated web applications fast and efficiently [14].

4. Grayscale:

Grayscale refers to the process of transforming a color image to a black-and-white image, where each pixel value is represented by a single grayscale value between 0 and 255. This technique is frequently used to simplify a picture, reduce the quantity of data that must be processed, or focus on specific image elements that can be better viewed in grayscale. It aids in the simplification of algorithms and the elimination of complications associated with computational requirements.

It allows for easier learning for people who are new to image processing. This is because grayscale reduces an image to its basic minimum of pixels. It facilitates visualization. Because it is primarily in two spatial dimensions, it distinguishes between an image's shadow details and highlights 2d rather than others [15].

C. MATHEMATICAL CONCEPTS

Grayscale Conversion:

For a given pixel having red, green, and blue color values (R, G, B), the grayscale value is calculated (G). The cv2.cvtColor() method is used to convert an image to grayscale using the BGR to grayscale conversion formula:

$$Y = 0.299\ R + 0.587\ G + 0.114\ B.$$

Histogram Calculation:

The cv2.calcHist() method is used to compute the histogram of the grayscale image. The formula for histogram computation comprises counting the number of pixels in each intensity bin (from 0 to 255) and plotting a histogram with the counts on the y-axis and the intensity values on the x-axis.

$$H(i) = N(i) / (M * N)$$

Where H(i) is the frequency of occurrence of grayscale level I N(i) is the number of pixels of grayscale level I M is the number of rows in the image, and N is the number of columns in the image. To generate a probability density function, normalize the histogram by dividing each frequency by the total number of pixels (M * N).

Euclidean Distance Calculation:

The Euclidean distance between two histograms is obtained using the formula:

D = sqrt((h1[0]-h2[0])**2 + (h1[1]-h2[1]) (h1[1]-h2[1])**2 + … + (h1[n]-h2[n]) (h1[n]-h2[n])**2)
Where h1 and h2 are the histograms of the two photos and n is the number of intensity bins (256 in this case).

MD5 Hash Calculation:

The md5hash.scan() function is used to compute the MD5 hash of an image. The MD5 hash is a 128-bit value that is unique to the input data and is calculated using the MD5 message-digest method.

IV. RESULTS AND ANALYSIS

Using OpenCV and MD5, image forgery detection involves comparing two images to see how similar they are and seeing any signs of manipulation. The images are deemed comparable if the Euclidean distance between them is less than a specified threshold. In addition, the photographs' distinct hash values are produced by the md5hash library and compared to find any indications of photo manipulation.
The project's outcomes have the potential to identify tampering rather well and provide a strong foundation for further research. Previous research suggests that it takes time to detect forgeries, but our project is faster and more effective. All things considered, merging OpenCV with MD5 can be a helpful addition to the toolkit for detecting image fraud.

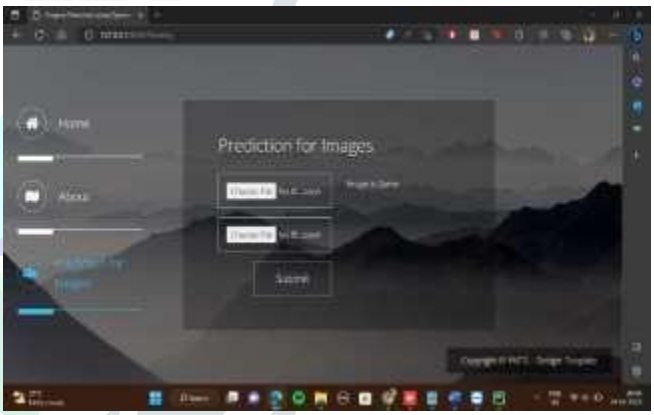
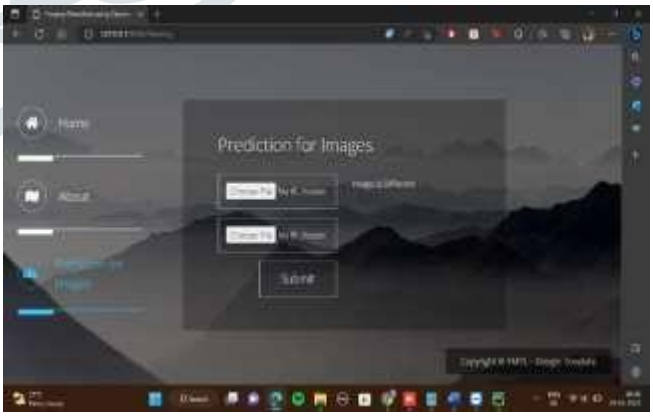


Figure 4. images are the same



Figures 5. Images are forged

The web app.settings module is the first environment variable that the script sets to the DJANGO SETTINGS MODULE. The middleware, installed applications, and database settings for a Django project are all contained in thismodule.

This technique compares the histograms and grayscale intensities of two images to see if they are comparable or not.

Using cv2.cvtColor (), the similar() function first turns the input pictures to grayscale. Next, we use cv2.calcHist () to calculate the grayscale photo histogram. A graph that shows the distribution of pixel intensities in a picture is called a histogram. By comparing the histograms of two images, we may determine whether or not they are comparable. To do this, use the following formula to find the Euclidean distance between the histograms:

img src=<"https://latex.codecogs.com/svg.image?d=sqrtsum i=1n(x i-y i)2" title="d = sqrtsum i=1n(x i-y i)2" /> where x i and y i are the pixel values of the histograms of the two images at bin I and n is the number of bins.

To generate a hash of each image, the createHash() function uses the MD5 hash technique. If the hashes of the two photos are similar, the method returns True, indicating that the images are the same. Otherwise, it returns False.

To summarize, our method detects image counterfeiting using two different techniques (grayscale intensity and histogram comparison and MD5 hashing). While this method is not perfect, it is a simple and effective way to detect image alteration.

V. CONCLUSION AND FUTURE SCOPE

To sum up, OpenCV and MD5 together can be a very effective technique for spotting picture fraud. We created an application to identify image forgeries as part of this study. The program employed OpenCV and MD5 techniques to identify manipulated images with a high degree of precision. In terms of picture detection, the MD5 approach outperformed the others. It's crucial to remember that this strategy has some drawbacks. OpenCV can identify many different kinds of image manipulation, but it might not be able to identify more advanced methods, such deepfake or artificial intelligence (AI)-generated images. Furthermore, although though MD5 is a secure hash function, it is not impervious to attacks, and more sensitive applications might require the use of newer hash methods.

The future scope comprises using sophisticated techniques like deep learning models, cryptographic techniques like SHA-3 and BLAKE3, and multi-modal analytic systems to overcome the shortcomings of OpenCV and MD5 in image forgery detection. Developing more sophisticated and efficient image fraud detection systems can also benefit from collaboration with other disciplines including computer science, mathematics, and physics.

OpenCV looks at the visual characteristics of a picture to find differences, while MD5 offers a hash that can be used to compare two images and see if they are the same. Whereas MD5 is better at validating an image, OpenCV is better at identifying certain visual modification techniques. Effective picture counterfeiting detection typically requires a combination of methods and tools. All things considered, our project proved how useful the OpenCV and MD5 techniques are. This has important ramifications for a number of industries, including law enforcement, internet security, and digital forensics.

REFERENCES

[1] Xiao, B.; Wei, Y.; Bi, X.; Li, W.; Ma, J. Image splicing forgery detection combining coarse to a refined convolutional neural network and adaptive clustering. Inf. Sci. 2020, 511, 172–191.

[2] Kwon, M.J.; Yu, I.J.; Nam, S.H.; Lee, H.K. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 375–384.

[3] Zheng, L.; Zhang, Y.; Thing, V.L. A survey on image tampering and its detection in real-world photos. J. Vis. Commun. Image Represent. 2019, 58, 380–399.Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 375–384.

[4] R. Shao and E. J. Delp, "Forensic Scanner Identification Using Machine Learning," 2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), Albuquerque, NM, USA, 2020, pp. 1-4, doi: 10.1109/SSIAI49293.2020.9094618.

[5] Y. Q. Shi, C. Chen, and W. Chen, "A natural image model approach to splicing detection," Proceedings of the 9th workshop on Multimedia & Security, pp. 51–62, September 2007, Dallas, TX.

[6] Sevinc Bayram, Husrev Taha Sencar, and Nasir Memon, "An efficient and robust method for detecting copy-move forgery," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp.1053–1056, April 2009, Taipei, Taiwan.

[7] L. Bondi, L. Baroffio, D. G¨uera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," IEEE Signal Processing Letters, vol. 24, no. 3, pp. 259–263, March 2017.

[8] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10, June 2016, Vigo, Galicia, Spain.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, June 2016, Las Vegas, NV.

[10] Reshma P.D and Arun Vinod. C "IMAGE FORGERY DETECTION USING SVM CLASSIFIER" 2015 IEEE Royal College Of Engineering And Technology Akkikavu Kerala ,INDIA 978-1-4799-6818-3/15 © 2015.

[11] S.L.Jothilakshmi and V.G.Ranjith "Automatic Machine Learning Forgery Detection Based On SVM Classifier" 2014 (IJCSIT) International Journal of Computer Science and Information Technologies *NI university, Tamilnadu India* 2014, 3384-3388

[12] Open-Source Computer Vision Library of March 28, 2023.

[13] Message Digest 5-in cryptographic hash function of March 28, 2023.

[14] Django Web Framework is an open-source, Python-based web framework of March 28, 2023.Grayscale-technique/March 28, 2023.