International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

An Overview of Image Duplicacy Detection using Machine Learning

Shreya Moghe[1], Vaishnodevi Pardeshi[2], Dixita Patel[3], Arya Zinje[4], Prof. Nilam Honmane[5]

[1] Student, Information Technology, ZCOER, Pune 411041, India (email: shreyamoghe20@gmail.com)
[2] Student, Information Technology, ZCOER, Pune 411041, India (email: vaishnopardeshi29@gmail.com)
[3] Student, Information Technology, ZCOER, Pune 411041, India (email: dixitapatel100@gmail.com)
[4] Student, Information Technology, ZCOER, Pune 411041, India (email: aryazinje2020@gmail.com)
[5] Guide, Information Technology, ZCOER, Pune 411041, India (email: nilam.honmane@zealeducation.com)

ABSTRACT

In the era of digital content proliferation, the detection of duplicate images plays a pivotal role in various domains, including copyright protection, content-based image retrieval, and security. This survey paper provides a comprehensive overview of existing techniques for duplicate image detection, ranging from content-based methods to perceptual hashing and deep learning approaches. In addition, we propose a novel methodology that leverages K-Nearest Neighbors (KNN), perceptual hashing, and bit-packed encoding to enhance the accuracy and efficiency of duplicate image detection. Our survey highlights the strengths and limitations of each approach and offers insights into the evolving landscape of this field. This work contributes to the body of knowledge on duplicate image detection and suggests directions for future research.

Keywords: *Duplicate Image Detection, Machine Learning, Hashing, Perceptual Hashing, Deep Learning, K-Nearest Neighbors.*

I. INTRODUCTION

Duplicate image detection is the process of identifying images that are either exact copies or exhibit substantial visual similarity. These duplications can arise from a myriad of sources, including image theft, inadvertent redundancy, or format conversions. Whether it's protecting copyrights, improving the efficiency of image search engines, or maintaining data integrity, the significance of this task cannot be overstated. However, the path to accurate and efficient duplicate image detection is laden with complexities. The diverse nature of digital images, coupled with variations in image resolution, format, and the presence of transformations, presents formidable challenges. Images may undergo resizing, cropping, filtering, or other subtle modifications, further complicating the task of recognizing

their similarity. Noise, compression artifacts, and alterations due to different image encoding and storage technologies add another layer of intricacy.

In the ever-expanding digital realm, the magnitude of the challenge is evident. An effective duplicate image detection system must contend with immense datasets, sometimes containing thousands, if not millions, of images. Scaling up existing methodologies to meet these demands without compromising accuracy is a formidable undertaking. Yet, the advantages of a proficient duplicate image detection system are manifold. In a world where image-based plagiarism is on the rise, protecting intellectual property becomes paramount. Detecting duplicate images can assist in identifying cases of copyright infringement, ensuring fair compensation for creators, and preventing content theft. Moreover, in image search engines and content management systems, eliminating redundant or similar images can significantly enhance search results and database efficiency, streamlining the user experience and reducing computational overhead.

This survey paper embarks on a comprehensive exploration of duplicate image detection. Our objective is to offer readers a comprehensive understanding of existing methodologies and techniques, while also introducing an innovative approach. Leveraging the power of K-Nearest Neighbors (KNN), perceptual hashing, and bit-packed encoding, while promising, represent only a fraction of the vast landscape of machine learning algorithms available. While our current methodology has shown promise in addressing certain aspects of the challenge, we recognize the need for continued exploration and experimentation. Our proposed method aims to enhance the precision and computational efficiency of duplicate image detection. By synthesizing the wealth of existing knowledge, this survey intends to serve as a roadmap for researchers, developers, and practitioners in the field.

II. LITERATURE SURVEY

Aditi Joshi, Aman V Shet, Adarsh S Thambi, Sunitha R in “Quality Improvement of Image Datasets using Hashing Techniques” [1], Journal - International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), states that the dataset becomes more efficient by detecting duplicates and near duplicates. Implementation on Difference Hash (dHash), Average Hash (aHash), and Perceptual Hash (pHash). Hash functions are supervised to detect near-identical images because it predicts minor changes and minimizes the number of false positive collisions. The model presented in this paper is used for various real-time image processing applications.

R. Kaur, Jhilik Bhattacharya, Inderveer Chana in “A Deep CNN based online image deduplication technique for cloud storage system” [2], Journal – Multimedia tools and Applications. This paper detects exact and near-exact images using cross-functional, even in the presence of disruptions in the form of blur, noise, compression. The experimental results in this paper title deep CNN for online image deduplication technique outperforms in terms of image matching accuracy and performance too.

Dr. Manjunath K, Yogeen S. Honnavar, Rakesh Pritmani, Rakesh Kumar, Sethuraman K in "Detection of duplicate and non- face images in the e Recruitment applications using machine learning techniques" [3], Journal - International Journal of Robotics and Automation IJRA. To recognize image uploaded to recruitment portal contains a human face or not & to check whether images uploaded by 2 or more applications are same or not. This is achieved by using ML algorithm to generate similar score bet 2 images & then identify the duplicate images.

K. Thyagharajan, G. Kalaiarasi in "A Review on near duplicate detection of images using Computer Vision Techniques" [5], Journal - Archives of Computational Methods in Engineering. Feature extraction methods are used for detecting near duplicate images. Image understanding is the main application of computer vision. Automatic extraction, analysis and understanding of useful information from digital images is what computer vision is related to. The near duplicates also affect the search engine performance.

S.Manjunatha, M. Patil in " Deep Learning- based technique for "Image Tamper Detection" [4], Journal – Third International Conference on Intelligent Communication Technologies and virtual mobile networks (ICICV) 2021. In this paper it detects that Pixels plays an important role in deep learning-based feature extraction techniques and in detecting the tampering of the image. Not all methods have good accuracy for attacks such as splicing copy-move, etc.

III. METHODOLOGY

1) Data Collection

(Figure-1)

The CASIA dataset on Kaggle is a subset of the original CASIA duplicate image detection evaluation database, consisting of 749 images in 'real' and 'fake' folders. The 'Real' folder contains 249 duplicate images, while the 'Fake' folder contains 500 images subjected to various duplication methods. This dataset is highly diverse, covering a wide range of scenes, objects, people, and events, making it valuable for evaluating duplicate image detection in real-world scenarios. The inclusion of both real and manipulated images allows for extensive testing, which increases the robustness of the method. The

'fake' folder provides a practical assessment of the method's ability to detect changes and similarities in the face of diverse transformations.

2) Model Selection

K-Nearest Neighbors (KNN) is a widely known machine learning method known for its simplicity and effectiveness in data classification. Its basic principle rests on the observation that similar data points often cluster in similar regions of feature space. The choice of KNN for this task is driven by its image comparison and straightforward implementation. Ongoing Exploration for Additional models, including CNN. Our exploration extends beyond KNN as we actively explore the potential of Convolutional Neural Networks (CNN). CNNs have garnered recognition for their remarkable performance in a wide range of image-related tasks. By including CNN techniques in our study, we exemplify our dedication to harnessing cutting-edge approaches to enhance the accuracy and robustness of our duplicate image detection system.

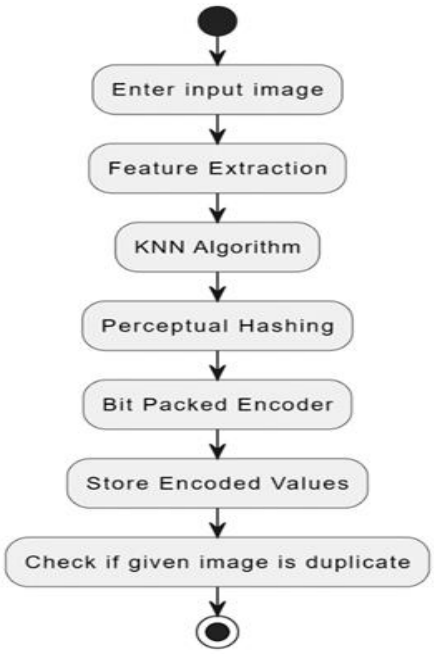
3) Perceptual Hashing

Perceptual hashing is a method that generates a unique fingerprint, or hash, for multimedia files like images or audio, based on their distinctive features and content. This technique is instrumental in combating online piracy, detecting plagiarism in multimedia content, and identifying duplicate or closely similar files within extensive databases or collections. Perceptual hashing offers advantages in terms of speed and efficiency compared to other hashing techniques, as it only analyses the features and content of files, not the entire file itself.

4) Bit-Packed Encoding

Bit-packed encoding is a data compression technique that condenses multiple values into a single byte or word, employing a consistent number of bits for each value. This approach organizes the data into a bitstream, where each element occupies a pre-defined bit count. Bit-packed encoding enhances the storage, processing, and comparison of perceptual hash values, significantly improving the efficiency and practicality of duplicate image detection.

5) Flow Chart



(Figure-2)

The flowchart describes the sequence of image detection, which requires input of the image by the user. The system starts with subtraction, follows the KNN algorithm, uses hash sense and encodes the image using a bit encoder. The encoded data is then saved, and the system checks whether the input images are duplicates.

IV. CONCLUSION

In conclusion, our extensive research and analysis underscore the complexity of image discovery, showcasing the diversity of techniques and methods at our disposal. The proposed approach, employing the K-Nearest Neighbours (KNN) algorithm, sensitivity washing, and bitwise coding, holds promise for efficiently identifying duplicate images within the CASIA Kaggle database. However, this marks only the initial phase of our exploration. Further research is imperative to elevate the accuracy and scalability of the proposed approach. This involves the exploration of feature extraction techniques, the integration of advanced machine learning models such as Convolutional Neural Networks (CNN), and the investigation of ensemble techniques. The quest for perfection in duplicate image detection extends beyond the academic realm, with practical implications for content management, copyright protection, and image search engines. As the field continues to evolve, our contribution lies in expanding the knowledge base for researchers and practitioners, providing insights that pave the way for more sophisticated, efficient, and adaptive solutions, ushering in a promising future in the realm of duplicate image detection.

REFERENCES

[1] Joshi, A., Shet, A. V., Thambi, A. S., & Sunitha, R. (2023). "Quality Improvement of Image Datasets using Hashing Techniques."
[2] Kaur, R., Bhattacharya, J., & Chana, I. (2022). "A Deep CNN based online image deduplication technique for cloud storage system.", DOI:10.1007/s11042-022-13182-7
[3] Manjunath, Dr., Honnavar, Y. S., Pritmani, R., Kumar, R., & Sethuraman, K. (2021). "Detection of duplicate and non-face images in the e-Recruitment applications using machine learning techniques.", DOI:10.11591/ijra.v10i2.pp114-122
[4] Manjunatha, S., & Patil, M. (2021). "Deep Learning-based technique for Image Tamper Detection."
[5] Thyagharajan, K., & Kalaiarasi, G. "A Review on near duplicate detection of images using Computer Vision Techniques.", DOI: 10.1007/s11831-020-09400-w
[6] Zhang, Q., & Cheng, W. (2020). "A novel near-duplicate image detection method based on convolutional neural network features and visual attention model." Signal Processing: Image Communication, 85, 115852.
[7] Taskesen, E. "Detection of Duplicate Images Using Image Hash Functions." Towards Data Science.
[8] Hu, Y., Wang, Y., & Ai, X. (2021). "Image-Based Copy-Paste Tamper Detection Technology Based on Improved SURF."
[9] Gusev, A., & Xu, J. (2022). "Evolution of a Web-Scale Near Duplicate Image Detection System."
[10] PyImageSearch. "Detect and Remove duplicate images from a dataset for deep learning."
[11] Lamberti, F. "Benchmarking unsupervised near-duplicate image detection."
[12] DragonOfMath. "dupe-images." GitHub. https://github.com/DragonOfMath/dupe-images.
[13] Thyagharajan, K. K., & Kalaiarasi, G. "Duplicate Image Detection in Large Scale Databases." Research Gate.
[14] Stack Overflow. "Near-Duplicate Image Detection."
[15] Yafeng, L. (2021). "A Fast Algorithm for Near-Duplicate Image Detection." Communication University of China, Beijing, China. IEEE, 23 June 2021.
[16] Thakur, R., & Rohilla, R. "Recent advances in digital image manipulation detection techniques: A brief review." ScienceDirect.