International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

TEXT AND IMAGE PLAGIARISM DETECTION USING LONGEST COMMON SUBSEQUENCEAND FAST MULTIPOLE METHOD

K VIJAY KRISHNA [1], MD.RAHEEM [2], MATTA RAHUL [3], SIRIPURAM ACHYUTH [4], J.NEERAJ KUMAR [5]
[1,2, 3, 4,5]UG Student,DepartmentofCSE,*NOVA COLLEGE OF ENGINEERING AND TECHNOLOGY,* Jafferguda, R.R Dist., Ranga Reddy, Telangana India. 501512

ABSTRACT:

In an educational environment, plagiarism is a crucial task that needs to be identified, in recent years all known journals and conferences, as well as universities, request a plagiarism report from students and researchers to prove the originality of published text or scientific paper. Plagiarism detection usually checks the text content via many of the platforms which areavailable for productive use reliably identifying copied text or near-copies of text and these systems usually fail to detect the images, and Files plagiarism since it is originally built for text mainly. In this paper, we suggest an adaptive, scalable, and extensible, robust method for image plagiarism which is tested in designs collect from department of architecture University of Technology, this method mainly compare the data (designs images) entered to the system with data sets saved in the database mainly these designs are saved as feature which is one of the artificial intelligence algorithms and match by using k-mean clustering and the similarity check is done with threshold used 40% which can be changed to an accepted levels when needed. Using the k-mean algorithm in clustering, whichis a robust artificial intelligence clustering algorithm giving us a strong system that is not discarding any feature extracted from the image. In this paper, data sets consist of 45 samples as training images saved and used in the system as the system database and using 48 samples as testing images which consist of original and forgery designs. These testing images were evaluated with 100% matching rate and 81% matching accuracy rating.

Keywords:Text, Image, plagiarism, ML.

1. INTRODUCTION:

Plagiarism is any identical or lightly-altered use of one's own or someone else’s work (ideas, texts, structures, images, plans, etc.) without adequate reference to the source. There are two main types of plagiarism as Text Based Plagiarism and Image Based Plagiarism. Text Based Plagiarism includes ‘copying textual information available from internet or other resources without proper permission and presenting it as their own” Image Based plagiarism includes "copying an image or portions of an image from the Internet or from classroom resources without permission or proper acknowledgment.” Hashing techniques are used in the process of plagiarism detection. Hashing is like calculating a fingerprint value of the image considering different parameters. There are different algorithms for calculating hash value like Average Hash, Difference hash, Perceptual Hash. A. Average Hash (Ahash): In this approach image is reduced to 8*8 and 64 bits are set based on whether colour is greater than average colour of the image or not. B. Difference Hash (DHash): In this type of hashing function image is reduced to 8*8 and then difference between pixels are calculated and their gradient is used to find whether image is plagiarised. C. Perceptual Hash (PHash): In this there are many functions which calculates footprint of the image based on the features present in the image [3][4]. Based on literature survey it is proven that perceptual hash is more efficient than other hashing functions. Hence in the proposed methodology PHash is used.

2. LITERATURE SURVEY

Many systems use features to compare images. In 2010 [5], the authors proposed a fast detection method for copy-move forgery which is mainly based on Speeded Up Robust Feature (SURF) descriptors, SURF provides ability to system to discard rotation, scaling, etc. Since its scalability to work effectively with it. Theauthors claim that their system was valid to detect regions duplication, and it works strongly with noise and blurring. Although the SURF provides high-speed feature extraction, the SIFT algorithm extracts more features than SURF and provides more accuracy.

In 2017 [6], the authors suggested a novel passive image forgery detection method which is mainly built based on Local Binary Pattern (LBP) and Discrete Cosine Transform (DCT) which was used to detect forgeries, features obtained from a tested image by applying 2D DCT in LBP space. Then the system is trained and matched by using a support vector machine; the system was tested with 3 image forgery datasets and got an accepted level of detection accuracy. The accuracy of George Bebis system is high, and the results were obtained after testing the system for three datasets, but the training time for the system is high, and it required special system requirements to handle the training process which is necessary with every update to the system.

Wei Wang (2017) in [7] proposes an optimized 3D lighting estimation method by working on a reflection model. Within this model, the authors took the occlusion geometry and surface texture information into consideration. They show that the system is effective and accurate when comparing the proposed method with existing 3D lighting-based forensic methods. A new technique that is used to detect the forgery parts but it is not effective properly with the images that have irregular shapes or models.

In 2018 [4], the author proposed a new detection approach that using perceptual hashing and tested with newly developed similarity assessments that the author did for images, the system is using a ratio hashing and positioning aware Object Character Recognition (OCR) system tested with 15 images. According to authors system output results were accurate and also very low error rate. Using OCR requires using a proper.

3. METHODOLOGY

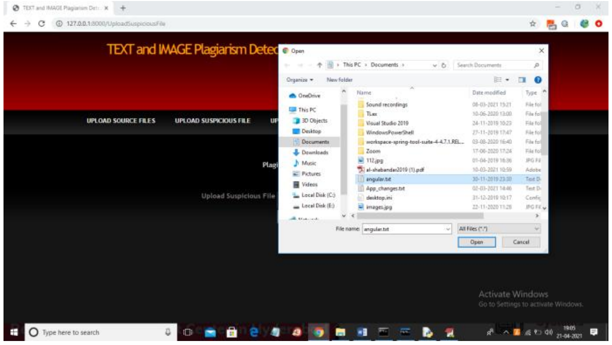
There are two main types of plagiarism as Text Based Plagiarism and Image Based Plagiarism. Text Based Plagiarism includes ‘copying textual information available from internet or other resources without proper permission and presenting it as their own” Image Based plagiarism includes "copying an image or portions of an image from the Internet or from classroom resources without permission or proper acknowledgment.” Hashing techniques are used in the process of plagiarism detection.There are different algorithms for plagiarism. here we are using corpus for image and Text.

PROBLEM DEFINITION

The corpus and the measures form the first controlled evaluation environment dedicated to plagiarism detection. Unlike other tasks in natural language processing and information retrieval, it is not possible to publish a collection of real plagiarism cases for evaluation purposes since they cannot be properly anonymized. Therefore, current evaluations found in the literature are incomparable and often not even reproducible. Our contribution in this respect is a newly developed large-scale corpus of artificial plagiarism and new detection performance measures tailored to the evaluation of plagiarism detection algorithms

OBJECTIVE OF PROJECT

We aimed to create a corpus that could be used for the development and evaluation of plagiarism detection systems that reflects the types of plagiarism practiced by students in an academic setting as far as realistically possible.



TEXT and IMAGE Plagiarism Detec

Open

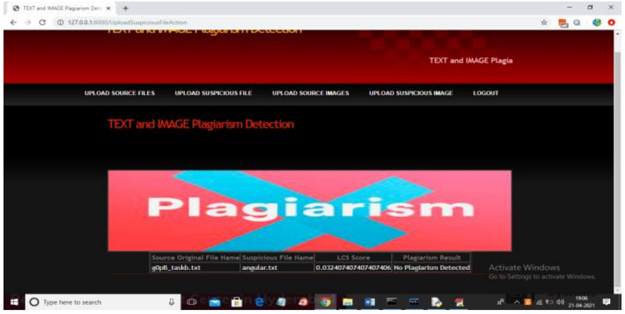
This PC › Documents ›

Name	Date modified	Type
Sound recordings	08-03-2021 13:21	File fol
TLax	10-06-2020 13:00	File fol
Visual Studio 2019	24-11-2019 10:23	File fol
WindowsPowerShell	27-11-2019 17:47	File fol
workspace-spring-tool-suite-4-4.7.1.REL...	03-08-2020 16:40	File fol
Zoom	17-06-2020 17:24	File fol
112.jpg	01-04-2019 16:36	JPG Fil
al-shabandar2019 (1).pdf	10-03-2021 10:59	Adobe
angular.txt	30-11-2019 23:30	Text D
App_changes.txt	02-03-2021 14:46	Text D
desktop.ini	31-12-2019 10:17	Confic
images.jpg	22-11-2020 11:28	JPG Fil

File name: angular.txt

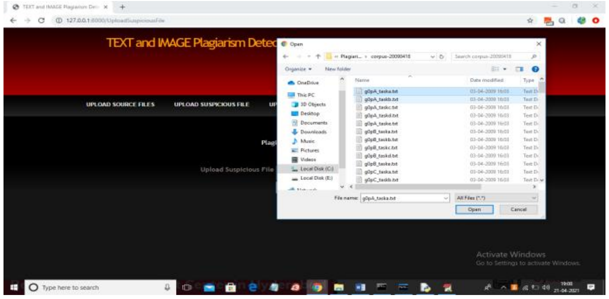
All Files (*.*)

In above screen I am selecting and uploading ‘angular.txt’ file and then click on ‘Open’ button to get below result and then click on ‘Check Plagiarism’ button to get result



Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result
g0pB_taskb.txt	angular.txt	0.032407407407407406	No Plagiarism Detected

In above screen angular.txt file matched very little with g)pB_taskb.txt corpus file and we got similarity score as 0.03 so no plagiarism detected and now upload any file from corpus and see result



TEXT and IMAGE Plagiarism Detec

Open

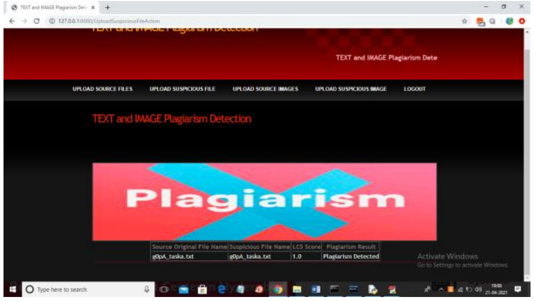
Plagiari... › corpus-20090418

Name	Date modified	Type
g0pA_taska.txt	03-04-2009 16:03	Text D
g0pA_taskb.txt	03-04-2009 16:03	Text D
g0pA_taskc.txt	03-04-2009 16:03	Text D
g0pA_taskd.txt	03-04-2009 16:03	Text D
g0pA_taske.txt	03-04-2009 16:03	Text D
g0pB_taska.txt	03-04-2009 16:03	Text D
g0pB_taskb.txt	03-04-2009 16:03	Text D
g0pB_taskc.txt	03-04-2009 16:03	Text D
g0pB_taskd.txt	03-04-2009 16:03	Text D
g0pB_taske.txt	03-04-2009 16:03	Text D
g0pC_taska.txt	03-04-2009 16:03	Text D
g0pC_taskb.txt	03-04-2009 16:03	Text D

File name: g0pA_taska.txt

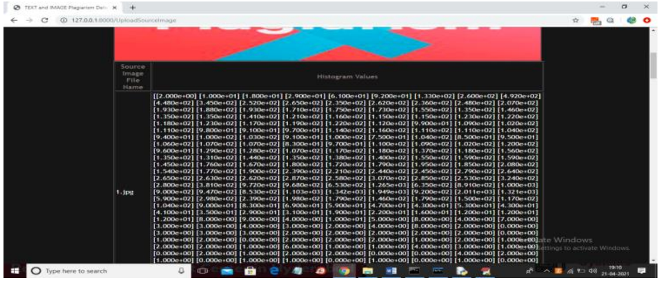
All Files (*.*)

In above screen I am selecting and uploading first file and then click on button to get below result



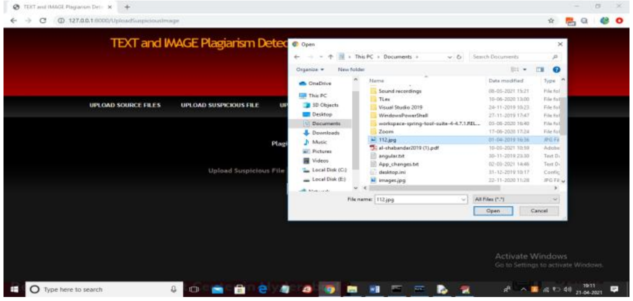
Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result
g0pA_taska.txt	g0pA_taska.txt	1.0	Plagiarism Detected

In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on ‘Upload Source Images’ link to upload all images from ‘images’ folder

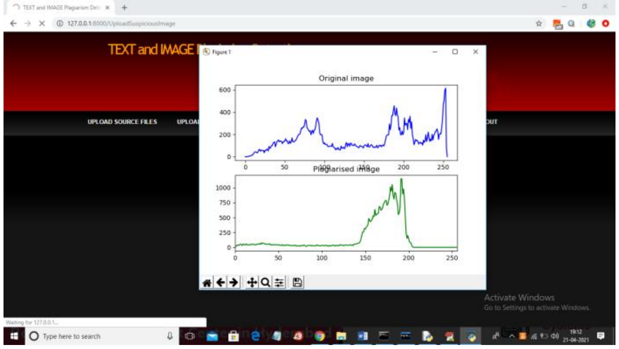


Source Image File Name	Histogram Values
1.jpg	[[2.000e+00] [1.000e+01] [1.800e+01] [2.900e+01] [6.100e+01] [9.200e+01] [1.330e+02] [2.600e+02] [4.920e+02] [4.480e+02] [3.450e+02] [2.520e+02] [2.650e+02] [2.350e+02] [2.620e+02] [2.360e+02] [2.480e+02] [2.070e+02] [1.930e+02] [1.880e+02] [1.930e+02] [1.710e+02] [1.750e+02] [1.730e+02] [1.550e+02] [1.350e+02] [1.460e+02] [1.350e+02] [1.350e+02] [1.410e+02] [1.210e+02] [1.160e+02] [1.150e+02] [1.150e+02] [1.230e+02] [1.220e+02] [1.180e+02] [1.230e+02] [1.170e+02] [1.190e+02] [1.220e+02] [1.120e+02] [9.900e+01] [1.090e+02] [1.020e+02] [1.110e+02] [9.800e+01] [9.100e+01] [9.700e+01] [1.140e+02] [1.160e+02] [1.110e+02] [1.110e+02] [1.040e+02] [9.400e+01] [1.000e+02] [1.030e+02] [9.100e+01] [1.000e+02] [7.500e+01] [1.040e+02] [8.500e+01] [9.500e+01] [1.060e+02] [1.070e+02] [1.070e+02] [8.300e+01] [9.700e+01] [1.100e+02] [1.090e+02] [1.020e+02] [1.200e+02] [9.600e+01] [1.290e+02] [1.280e+02] [1.070e+02] [1.170e+02] [1.180e+02] [1.370e+02] [1.180e+02] [1.560e+02] [1.350e+02] [1.310e+02] [1.440e+02] [1.350e+02] [1.380e+02] [1.400e+02] [1.550e+02] [1.590e+02] [1.590e+02] [1.450e+02] [1.760e+02] [1.670e+02] [1.800e+02] [1.720e+02] [1.790e+02] [1.950e+02] [1.850e+02] [2.080e+02] [1.540e+02] [1.770e+02] [1.900e+02] [2.390e+02] [2.210e+02] [2.440e+02] [2.450e+02] [2.790e+02] [2.640e+02] [2.650e+02] [2.630e+02] [2.620e+02] [2.870e+02] [2.580e+02] [3.070e+02] [2.850e+02] [2.530e+02] [3.240e+02] [2.800e+02] [3.810e+02] [9.720e+02] [9.680e+02] [6.530e+02] [1.265e+03] [6.350e+02] [8.910e+02] [1.000e+03] [9.000e+02] [9.470e+02] [8.530e+02] [1.103e+03] [1.342e+03] [1.949e+03] [9.200e+02] [2.011e+03] [1.321e+03] [5.900e+02] [2.980e+02] [2.390e+02] [1.980e+02] [1.790e+02] [1.460e+02] [1.790e+02] [1.500e+02] [1.170e+02] [1.040e+02] [9.000e+01] [8.300e+01] [6.900e+01] [5.900e+01] [4.700e+01] [4.300e+01] [5.300e+01] [4.300e+01] [4.100e+01] [3.500e+01] [2.900e+01] [3.100e+01] [1.900e+01] [2.200e+01] [1.600e+01] [1.200e+01] [1.200e+01] [1.200e+01] [8.000e+00] [9.000e+00] [4.000e+00] [1.000e+01] [5.000e+00] [8.000e+00] [4.000e+00] [7.000e+00] [3.000e+00] [3.000e+00] [4.000e+00] [3.000e+00] [2.000e+00] [4.000e+00] [8.000e+00] [2.000e+00] [0.000e+00] [3.000e+00] [3.000e+00] [3.000e+00] [2.000e+00] [2.000e+00] [2.000e+00] [2.000e+00] [2.000e+00] [0.000e+00] [1.000e+00] [2.000e+00] [0.000e+00] [2.000e+00] [2.000e+00] [2.000e+00] [1.000e+00] [2.000e+00] [1.000e+00] [2.000e+00] [2.000e+00] [1.000e+00] [6.000e+00] [1.000e+00] [1.000e+00] [4.000e+00] [3.000e+00] [1.000e+00] [0.000e+00] [2.000e+00] [1.000e+00] [2.000e+00] [2.000e+00] [0.000e+00] [0.000e+00] [4.000e+00] [2.000e+00] [1.000e+00] [0.000e+00] [1.000e+00] [1.000e+00] [0.000e+00] [1.000e+00] [0.000e+00] [1.000e+00] [0.000e+00]

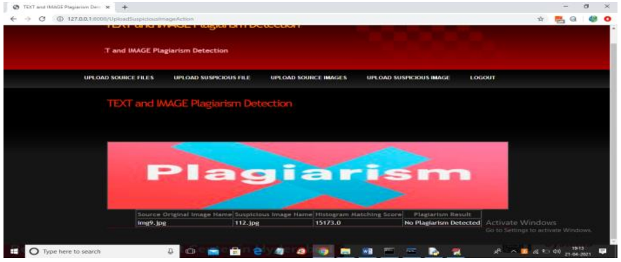
In above screen from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched and now click on 'Upload Suspicious Image' link to upload some image



In above screen I am selecting and uploading '112.jpg' file and then click on 'Open' button to get below result

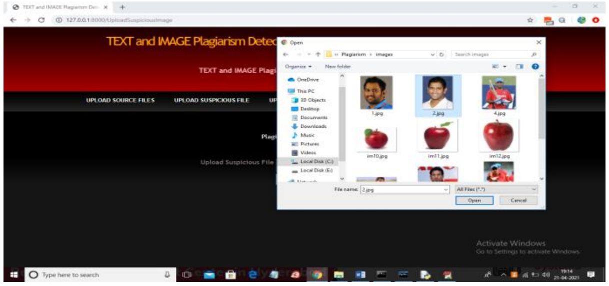


In above screen we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected and now close above graph to get below result

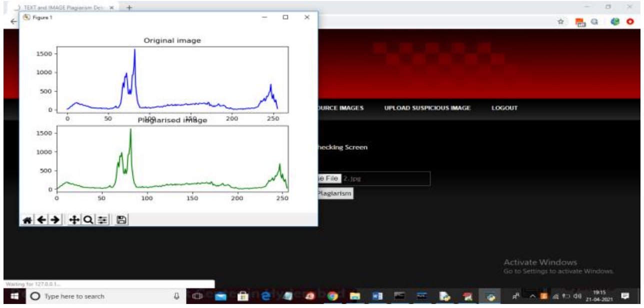


Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result
img9.jpg	112.jpg	15173.0	No Plagiarism Detected

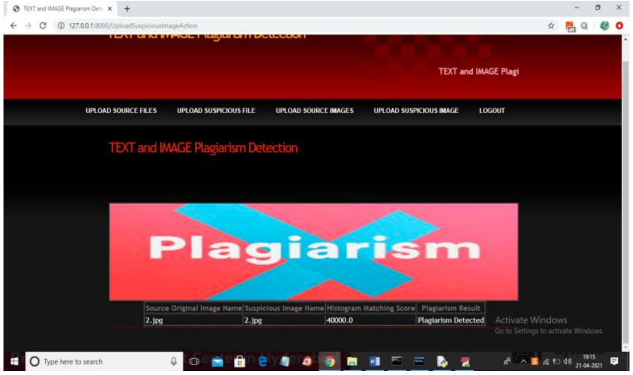
In above screen histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from "images" folder and see result



In above screen I am selecting and uploading '2.jpg' file from "images" database folder and below is the result



In above screen we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now close above graph to get below result



Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result
2.jpg	2.jpg	40000.0	Plagiarism Detected

In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

CONCLUSION

Corpus is the first standardized corpus dedicated to the evaluation of automatic plagiarism detection and was successfully employed in the First International Competition on Plagiarism Detection. We believe that our corpus and the performance measures will become an effective means to evaluate future plagiarism detection research. Currently, an improved version of the corpus is being constructed.

REFERANCES

[1] Tengyu Yu, Blockchain operation principle analysis: 5 key technologies, iThome, https://www.ithome.com.tw/news/105374

[2] JingyuanGao, The rise of virtual currencies! Bitcoin takes the lead, and the other 4 kinds can't be missed. Digital Age, https://www.bnext.com.tw/article/47456/bitcoinether-litecoin-ripple-differences-betweencryptocurrencies

[3] Smart contractswhitepaper, https://github.com/OSELab/learning-blockchain/blob/ master/ethereum/smart-contracts.md

[4] Gong Chen, Development and Application of Smart Contracts, https://www.fisc.com.tw/Upload/b0499306-1905-4531-888a-2bc4c1ddb391/TC/9005.pdf

[5] Weiwei He, Exempted from cumbersome auditing and issuance procedures, several national junior diplomas will debut next year.iThome, https://www.ithome.com.tw/news/119252

[6] Xiuping Lin, "Semi-centralized Blockchain Smart Contracts: Centralized Verification and Smart Computing under Chains in the EthereumBlockchain",Department of Information Engineering, National Taiwan University, Taiwan, R.O.C., 2017.

[7] Yong Shi, "Secure storage service of electronic ballot system based on block chain algorithm", Department of Computer Science, Tsing Hua University, Taiwan, R.O.C., 2017.

[8] ZhenzhiQiu, “Digital certificate for a painting based on blockchain technology”, Department of Information and Finance Management, National Taipei University of Technology, Taiwan, R.O.C., 2017.

[9] Weiwen Yang, Global blockchain development status and trends,

[10] Benyuan He, “An Empirical Study of Online Shopping Using Blockchain Technology”, Department of Distribution Management, Takming University of Science and Technology, Taiwan, R.O.C., 2017.