ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

PLAGIARISM DETECTION IN ACADEMIC DOCUMENTS USING COMBINED DCT+WAVELET+INVARIANT TRANSFORM BASED COPY PASTE DETECTION

Manpreet Kaur, ER. Manit Kapoor(Hod) , DR. Naveen Dhillon(Principal)
R.I.E.T Phagwara, India

ABSTRACT

The paper describes in brief about the plagiarism detection and various works done by different authors and the final section describes about the proposed method's efficiency as it eradicates and solves all the above mentioned occurring problems in the base technique i.e. Linear Programming technique. The importance of the technique is that it makes use of a method which detects and extracts more features as compared to method used in base technique. In comparison to the base technique the proposed will detect plagiarism in documents for text copied from other documents, for edited photo and images from various documents. It will identify accuracy of plagiarism detection in document images and will further compare the proposed work with the previous techniques under defined parameters. In the proposed method, the images document will be selected to process to cut and paste. Further the improvement filters will be applied if needed. Next step involves application of hybrid similarity and arte-fact detection using multi-temporal filtering using DCT analysis and classification using SVM and feature extraction that results in segmenting the image into parts. The proposed method uses the extracted feature for matching with documents suspected to be copied from and after matching test the accuracy, efficiency, time of the processing algorithm we further use image documents with / without distortions for testing fluency of the algorithm.

.

KEYWORDS: SVM, DCT, Wavelet, Linear Programming, DWT

INTRODUCTION

Plagiarism identification is surely understood sensation in the scholastic stadium. Duplicating other individuals is considered as genuine offense that should be checked. There are numerous written falsification recognition frameworks, for example, turn-it-in that has been created to give these checks [1]. Most, if not all, dispose of the figures and graphs before checking for counterfeiting. Disposing of the figures and diagrams brings about look openings that individuals can take advantage. That implies individuals can counterfeit figures and diagrams effectively without the present copyright infringement frameworks identifying it. The availability of powerful digital image processing programs, such as Photoshop, makes it relatively easy to create digital forgeries from one or multiple images and also the word files are copied. The plagiarism detection can be used in image manipulative detection in the evidence of the law enforcement agencies and in general publication document forgery, the detection is also a measure of how much plagiarism does the document consist of and from where the material was taken. Plagiarism detection is used where it is to be found whether the data available in the document is being copied from another document which is illegal. Content cannot be copied from any other document which has preserved its rights. Copying any data from such a document is illegal [2]. There are many definitions of what constitutes plagiarism, and we will look at some of them in more detail below. However, according to research resources at plagiarism.org, the things that immediately come to mind as description of plagiarism are:

- turning in someone else's work as your own
- copying words or ideas from someone else without giving credit
- failing to put a quotation in quotation marks
- giving incorrect information about the source of a quotation
- changing words but copying the sentence structure of a source without giving credit
- copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not

Plagiarism is derived from the Latin word "plagiarius" which means kidnapper. It is defined as "the passing off of another person's work as if it were one's own, by claiming credit for something that was actually done by someone else". Plagiarism is not always intentional or stealing some things from someone else; it can be unintentional or accidental and may comprise of self stealing [3]. The broader categories of plagiarism include:

- Accidental: due to lack of plagiarism knowledge, and understanding of citation or referencing style being practiced at an institute
- Unintentional: the vastness of available information influences thoughts and the same ideas may come out via spoken or written expressions as one's own
- Intentional: a deliberate act of copying complete or part of someone else's work without giving proper credit to original creator
- Self plagiarism: using self published work in some other form without referring to original one

There is a long list of plagiarism methods commonly in practise [4]. Some of these methodologies include

- Copy-paste: copying word to word textual contents.
- Idea plagiarism: using similar concept or opinion which is not common knowledge.
- Paraphrasing: changing grammar, similar meaning words, re-ordering sentences in original work. Or restating same contents in different words.
- Artistic plagiarism: presenting someone else's work using different media, such as text, images, voice or video.
- Code plagiarism: using program code, algorithms, classes, or functions without permission or reference.
- Forgotten or expired links to resources: addition of quotations or reference marks but failing to provide information or up-to-date links to sources.
- No proper use of quotation marks: failing to identify exact parts of borrowed contents.
- Misinformation of references: adding references to incorrect or non existing original sources.
- Translated plagiarism: cross language content translation and use without reference to original work.

TEXT BASED PLAGIARISM

This sort of plagiarism spotlights on distinguishing the likenesses between records by utilizing the vector space model. It additionally can figure and check the repetition of the word in the report, and after that they utilize the fingerprints for every record for coordinating it with fingerprints in different reports and discover the closeness. This technique is suitable for non incomplete copyright infringement as specified before utilize the entire report and utilization vector space to match between the records, yet in the event that the record has been part of the way copied it can't accomplish great results. It may incorporate duplicate and glue, adjustment or changing a few expressions of the first data from the web book magazine, daily paper, research, diary, individual data or thoughts [5, 6]. Text based plagiarism detection stages are:

- **Stage One Collection:** This is the first phase of Plagiarism Detection Process, and it involves the understudy or specialist to transfer their assignments or attempts to the web motor, the web motor goes about as an interface between the understudies and the framework.
- **Stage Two Analysis:** In this stage all the submitted corpus or assignments are gone through a comparability motor to figure out which records are like different reports. There are two sorts of likeness motors, first intra-corpal motor and second additional corpal motor. The intra-corpal motors work by returning requested rundown between each comparable sets. By complexity, the additional corpal motors return suitable web joins.
- **Stage Three Confirmations:** The capacity of this stage is to figure out whether the significant content has been counterfeited from different writings or to figure out whether there is a high level of comparability between a source record and whatever other archive.
- **Stage Four Investigation:** This is the last phase of a Plagiarism Detection Process and it depends on human intercession. In this stride a human master is in charge of figure out whether the framework ran accurately and also figuring out whether an outcome has been genuinely counterfeited or just cited.

IDENTIFYING CITATION PATTERNS

Finding similar patterns in the citations used within two scientific texts is a strong indicator for semantic text similarity and the core idea of CbPD. Patterns are subsequences in the citation tuples CA and CB of two texts A and B that (partially) consist of shared references and are therefore similar to each other.

- The degree of similarity between patterns depends on the number of citations included in the pattern, and the extent to which their order and/or the range they cover is alike. Thus, literally matching subsequences of citations in two documents are a strong indicator for semantic similarity.
- Unlike e.g. in string pattern matching the subsequence's of citations to be extracted from a suspicious text and searched for within an original are initially unknown [7, 8]. Citations that are shared by the two documents are easily identified. However, it is unlikely that all of those shared citations represent plagiarized text passages. For instance, two documents might share 8 citations, of which 3 are contained within a plagiarized text section and 4 are distributed over the length of the text and used along with other non-shared citations without representing any form of plagiarism. The citation sequences of the two documents might therefore look like the following:

1. $Original$: 1 2 3 x x x 4 x x 5 x 6 x 7 8
2. $Plagiarism$: x x 5 x x x 4 x 3 x 1 x 2 x x 7 x 8
3. Numbers 1-8 represent shared citations, the letter x non-shared citations. The shared citations 1-3 are supposed to represent a plagiarized passage.

CBPDS SYSTEM ARCHITECTURE

For the Citation-based Plagiarism Detection an Open Source programming framework in Java instituted CitePlag was created. These strides are performed in our counterfeiting discovery framework:

1. The record is parsed and a progression of heuristics connected to process the references, including their position inside of the archive.
2. Citations are coordinated with their entrances in the list of sources.
3. The reference based likeness of the records is figured.

The created model CbPDS comprises of three principle parts. The principal is a Relational Database System (RDBS) termed CbPD database putting away information to be procured from reports and also location results. The second is the location programming called CbPD Detector that recovers information from the CbPD Database. The third part, the CbPD Report Generator, makes condensed reports of location results for individual archive sets in light of flexible channel criteria. The three-level structural planning is outlined in the accompanying figure 1.

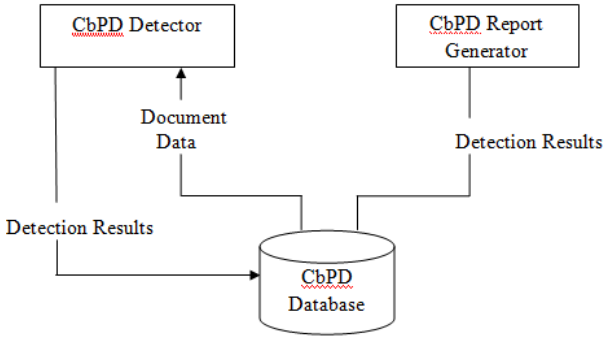


Figure 1 ***Citation based Plagiarism Detection System Architecture***

DETECTION BASED ON DOCUMENTS COMPARISON

The real objective of any unoriginality location framework is to highlight copyright violations. A infringement can happen when a piece of content of whatever size and circulation is copied between two or more records fitting in with diverse creators, for this situation the framework grammatically scans for any such covers. Notwithstanding, because of the multifaceted nature of normal dialects, it is conceivable that the same substance are exhibited in diverse semantics (e.g., rewording), or the same words or expressions could have distinctive implications in distinctive settings, for this situation a profound examination must be utilized by the framework, and some Natural Language Processing (NLP) strategies could be utilized. In both cases it is obliged that a referential gathering of records (corpus) exist. This segment quickly examines strategies for both semantic and syntactic literary theft recognition.

SEMANTIC-BASED DETECTION

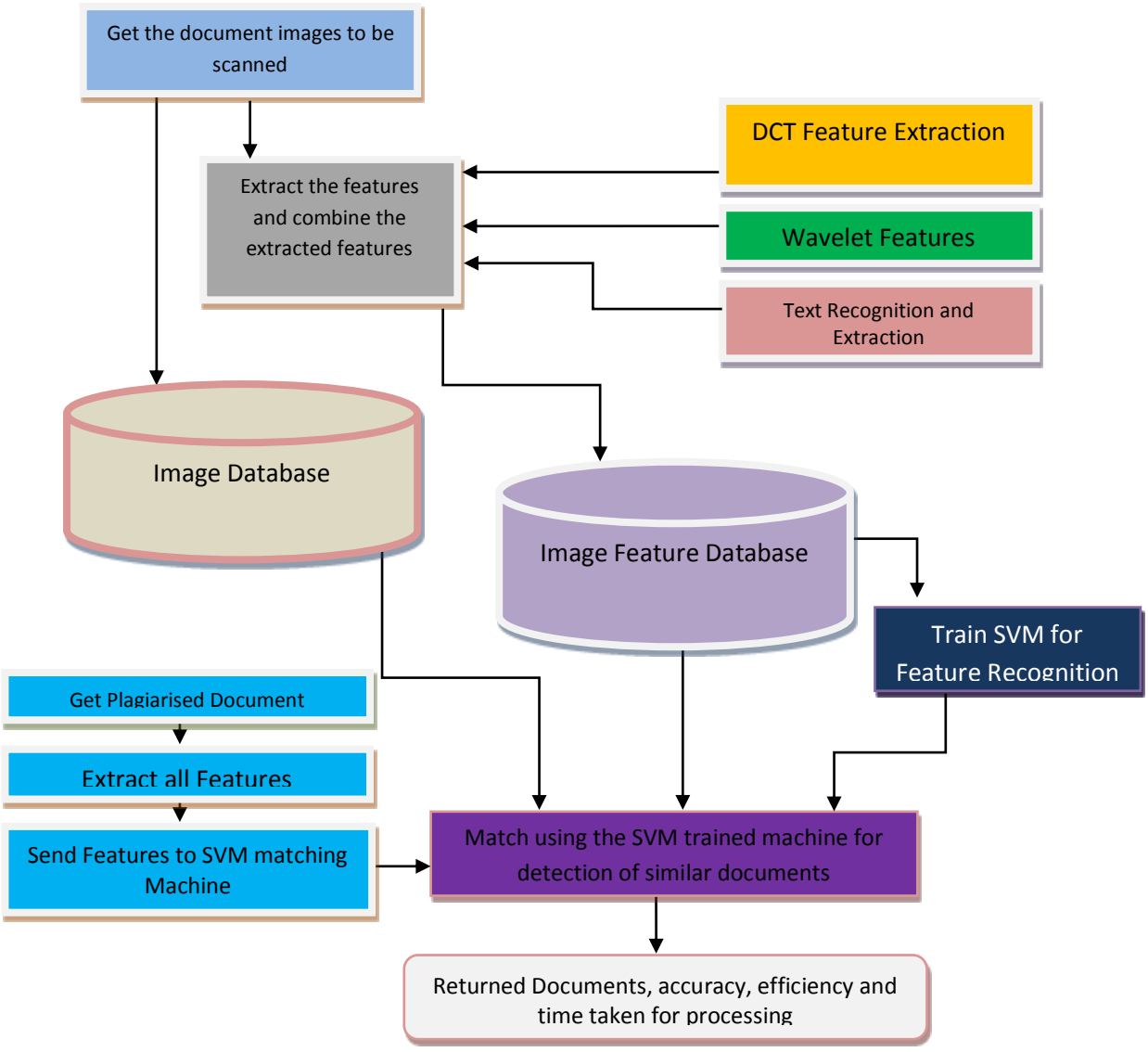
Most duplicate recognition framework can just stand up in comparison linguistically

Comparative words and sentences, therefore if the duplicated materials are changed extensively it is hard to recognize written falsification in such frameworks. The change can run from supplanting words by their equivalent words, to presenting the same idea under diverse semantics [9, 10].

PROBLEM STATEMENT

This is recognition free approach which is similar to recognition free document image retrieval system. Document image retrieval is very challenging field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. The main objective of thesis is to develop a plagiarism detection technique which provides better results as compared to the base technique (Linear Programming technique). The importance of the technique is that it makes use of a method which detects and extracts more features as compared to method used in base technique.

1. To detect plagiarism in documents for text copied from other documents.
2. To detect plagiarism of edited photo and images from various documents.
3. To identify accuracy of plagiarism detection in document images.
4. To compare the proposed work with the previous techniques under defined parameters.



Proposed Methodology

- Select the images document to process for cut and paste
- Use the improvement filters if needed / pre process the image raw
- Applying hybrid similarity and artefact detection using multi-temporal filtering using DCT analysis and classification using SVM.
- Perform the feature extraction or segment the image into parts
- Use the extracted feature for matching with documents suspected to be copied from
- After matching test the accuracy, efficiency, time of the processing algorithm

Use image documents with / without distortions for testing fluency of the algorithm

RESULTS AND DISCUSSIONS

The following are simulation Results for a number of documents which were forged with text and image copying in order to test and compare the output of the proposed system with that of the previous system.

Figure 3 shows the output of the proposed system with most plag matched documents

The above visual result shows the output of the proposed system which is same for the base system, but the difference is in the match and detection accuracy of the plag detected documents, the proposed system has a definitive edge on the base system as the number of features extracted from the image documents is high as compared to the base system output, which is shown in below graphical outputs for base and proposed systems.

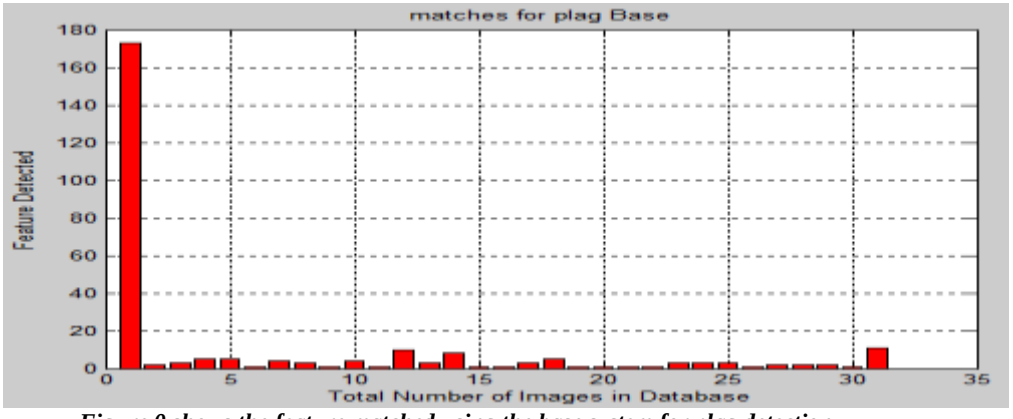


Figure 0 shows the feature matched using the base system for plag detection

The above figure shows the output of the base system for matched features, the maximum match number has reached a value of 170 for the best matched and the rest showed match below 20 features, which indicated loss occurring in feature extraction during database construction.

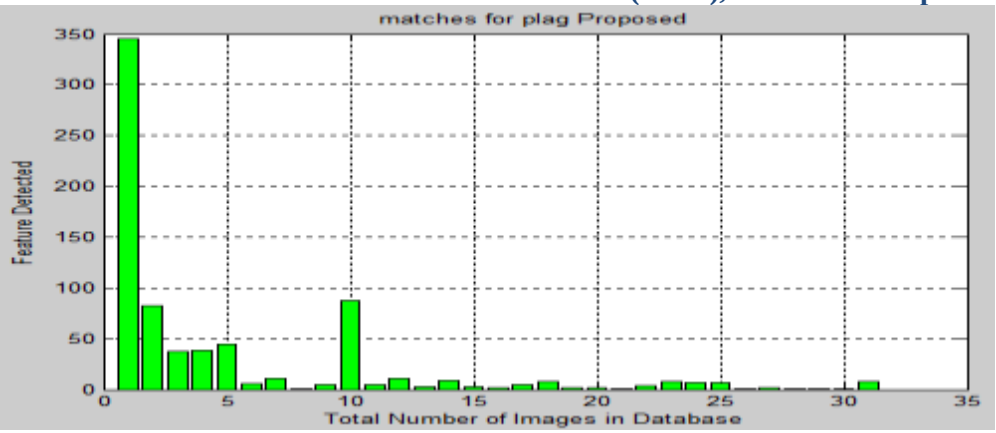


Figure 0.1 shows the feature matched using the proposed system for plag detection

The above figure shows the output of the proposed system for matched features, the maximum match number has reached a value of 340+ features for the best matched and the rest showed match below 100 but above 20 features, which indicated reduction in the loss occurring in feature extraction during compared to the base system.

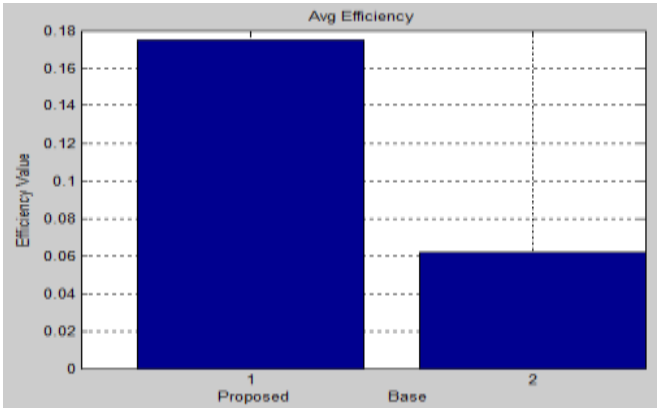


Figure 0.2 shows average efficiency in matching the detected features by proposed and base filters

The above figure shows the average efficiency for all the returned documents with respect to search document image and the overall difference in the efficiency of detect and match of proposed and base system.

Table 0.1 shows the extracted feature match for both base and proposed systems

Match	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Proposed Feature match	345	82	37	38	45	6	11	1	5	88
Base Feature match	173	2	3	5	5	1	4	3	1	4
Match	11	12	13	14	15	16	17	18	19	20
Proposed Feature match	5	11	3	9	3	2	5	8	2	2
Base Feature match	1	10	3	8	1	1	3	5	1	1
Match	21	22	23	24	25	26	27	28	29	30
Proposed Feature match	1	4	8	7	7	1	2	1	1	1
Base Feature match	1	1	3	3	3	1	2	2	2	1

The above table shows the matched features for all the 300 images with the query/ search document image from the base and proposed extracted feature databases

CONCLUSION

The proposed system of work deals with the extraction and match of image documents with the application in detection of duplication work detection for both image and text forgery, the system has different approach from the previous as it extracts the most stable feature points which are not affected by either crop or color changing attacks done by authors for hiding the copy paste activity, the proof of the system accuracy improvisation is done with comparison of the base technique which only incorporated the text detection and match based plagiarism detection, the overall average increase in efficiency and accuracy of detection in similar documents has proved the robustness of the proposed system over the base. The proposed system uses the advantage of singling out the best and worst match on the basis of the features of text and image combined with the total propagation of the image descriptors profiles, the efficiency of the system depends on the basis of maximum match values in the image document, the accuracy of the system is also not affected by the noise generated with scanners.

ACKNOWLEDGEMENTS

I would like to thank my guide (HOD) Er. Manit Kapoor and (Principal) Dr.Naveen Dhillon for their guidance and support and Department of Electronics and Communication Engineering, R.I.E.T College of Engineering, Phagwara.

REFERENCES

[1] U. Ozlem, K. Boris and N. Thade, "Using Syntactic Information to Identify Plagiarism," in proceedings of Massachusetts Institute of Technology Computer Science and Artificial Intelligence (MITCSAI) , Laboratory Cambridge, USA, pp. 37-44, 2005.

[2] B.G OvGU, "Citation-based Plagiarism Detection – Idea, Implementation and Evaluation," Germany / UC Berkeley, California, USA, 2010.

[3] A.Juan, C. Nicholas and C. Rafael, "Applying Plagiarism Detection to Engineering Education," in proceedings of School of Electrical and Information Engineering University of Sydney, NSW, pp. 722-731, 2006.

[4] M. Zimba and S. Xingming, "DWT-PCA (EVD) Based Copy-move Image Plagiarism Detection," in International Journal of Digital Content Technology and its Applications (IJDCTA), vol. 5, no. 1, 2011.

[5] M. Asim et al., "Overview and Comparison of Plagiarism Detection Tools," Department of Computer Science, Germany & UC Berkeley JCDL, 2011.

[6] B. Stein et al., "Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07," in SIGIR Forum, vol. 41, pp. 68-71, 2007.

[7] R. Karp and M. Rabin, "Efficient Randomized Pattern-Matching Algorithms," in IBM Journal of Research and Development, 2007.

[8] S. Tachaphetpiboon, N. Facundes and T. Amornraksa, "Plagiarism Indication by Syntactic-Semantic Analysis," in proceedings of Asia-Pacific Conference on Communications, 2007.

[9] J. Y. Bao, X. D. Shen and H. Y. Liu, "Finding plagiarism based on common semantic sequence model," in proceedings of the 5th International Conference on Advances in Web-Age Information Management, vol.31, no. 29, pp.640–645, 2004.

[10] P. Clough, "Old and new challenges in automatic plagiarism detection," Plagiarism Advisory Service, vol. 10, Department of Computer Science, University of Sheffield, 2003.