ISSN:0975-9646

Cut Paste Detection in Document Images Using Neural Network

Sarabjot Singh , Nishu Bansal

Indo Global Colleges,
Abhipur

Abstract:- **To manipulate and modify digital images are very easy due to rapid advances of image processing software. So, to judge the authenticity of a given image is very difficult for a viewer. Many documents are created by Cut-And-Paste (CAP) of existing documents. In this thesis, we proposed a novel technique to detect CAP in document images using Neural Network. This can help in detecting unethical CAP in document of image collections. Recognition free and scalable is the solution to large collection of documents. The formulation is also independent of the imaging process (camera based or scanner based) and does not use any language specific information for matching across documents. After that, we model the solution as finding a mixture of homo-graphics, and design a linear programming (LP) based solution to compute the same. The proposed method is presently limited by the fact that we do not support detection of CAP in documents formed by editing of the textual content. The proposed results demonstrate that without loss of generality (i.e. without assuming the number of source documents), it can be correctly detect and match the CAP content in a questioned document image by simultaneously comparing with large number of images in the database. For the implementation of this proposed work we use the Image Processing Toolbox under MATLAB software.**

Keywords: - Document retrieval, CAP detection, Plagiarism detection, Neural Networks.

I.INTRODUCTION

Document retrieval is referred to as a branch of Text Retrieval. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. With the emergence of large document repositories, many new documents get created by cut and paste (CAP) of documents. We believe there can be two directions to detect such cases of document forgeries – recognition based and recognition free. In this paper, we focus on a recognition free solution to the problem of detecting CAP in document images. To the best of our knowledge, this is the first attempt on detection of recognition free CAP and plagiarism. Since our method is recognition free, our closest work is that of retrieving similar documents given a query document, which is popularly known as recognition free document retrieval. Our problem, in a way, is a retrieval problem where the query is a part of a document (say a paragraph or few lines) and therefore, falls in between the two categories of work described above, which retrieves results based on query as entire document or query as words. In this work, we detect the forgery and plagiarism in documents by detecting CAP content in it. Because of the increasing availability of low-priced digital cameras, many applications have been built over the recent years to work on images captured using camera. The images captured from cameras are of often low quality and suffers from perspective distortion. We also use homography to remove perspective distortion while detecting CAP. There are some parameters given which is useful in our implementation.

Document retrieval: Document retrieval is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual.

Security: Everyone in your organization does not need to see every document you have. If you have sensitive files or are under strict government privacy regulations, you want to make sure that your documents can only be viewed by the people with the need to know.

Access: At the same time, you want to ensure that the people who should have access can view the files. Not only this, but they need access to them instantaneously. This is a standard part of any document management system.

II. LINEAR PROGRAMMING AND OPTIMIZATION

Linear programming (LP or linear optimization) is a method to achieve the best outcome (such as maximum profit or lowest cost) in a mathematical model whose requirements are represented by linear relationships. Linear programming is a special case of mathematical programming (mathematical optimization). More formally, linear programming is a technique for the optimization of a linear objective function, subject to linear equality and linear inequality constraints. Its feasible region is a convex polyhedron, which is a set defined as the intersection of finitely many half spaces, each of which is defined by a linear inequality. Its objective function is a real-valued affine function defined on this polyhedron. A linear programming algorithm finds a point in the polyhedron where this function has the smallest (or largest) value if such a point exists.

III. PLAGIARISM DETECTION

Plagiarism detection is the process of locating instances of plagiarism within a work or document. The widespread use of computers and the advent of the Internet have made it easier to plagiarize the work of others. Detection of plagiarism can be either manual or software-assisted.

Manual detection requires substantial effort and excellent memory, and is impractical in cases where too many documents must be compared, or original documents are not available for comparison. Systems for text-plagiarism detection implement one of two generic detection approaches, one being external, the other being intrinsic. External detection systems compare a suspicious document with a reference collection, which is a set of documents assumed to be genuine. Based on a chosen document model and predefined similarity criteria, the detection task is to retrieve all documents that contain text that is similar to a degree above a chosen threshold to text in the suspicious document. Intrinsic PDS solely analyze the text to be evaluated without performing comparisons to external documents. This approach aims to recognize changes in the unique writing style of an author as an indicator for potential plagiarism. PDS are not capable of reliably identifying plagiarism without human judgment. Similarities are computed with the help of predefined document models and might represent false positives.

IV. NEURAL NETWORKS

Neural network is set of interconnected neurons. And used for universal approximation. Artificial neural networks are composed of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons). The artificial neural network either is used to gain an understanding of biological neural networks; or for solving artificial intelligence problems without necessarily creating a model of a real biological system. Then real; biological nervous system is mostly complex: artificial neural network algorithms attempt to abstract this complexity and focus on what may hypothetically matter most from an information processing point of view. Good performance (e.g. as measured by good predictive ability; low generalization error) or performance mimicking animal or human error patterns.

Delta Rule: The delta rule is a gradient descent learning rule for updating the weights of the artificial neurons in a single-layer perceptron. It is a special case of the more general back propagation algorithm. For a neuron j with activation function g(x), the delta rule for j's, ith weight is given by

$$\Delta W_{ij} = (t_j - y_j)\, g'(h_j) x_i$$

The delta rule is commonly stated in simplified form for a perceptron with a linear activation function as $\Delta W_{ij} = \alpha (t_j - y_j) x_i$, where α is known as the learning rate parameter.

V. METHODOLOGY

Phase 1: Firstly code is developed for a particular opening GUI for this implementation. After that we develop a code for loading the document image in the MATLAB database.

Phase 2: Develop a code for the document retrieval from the loaded document image.

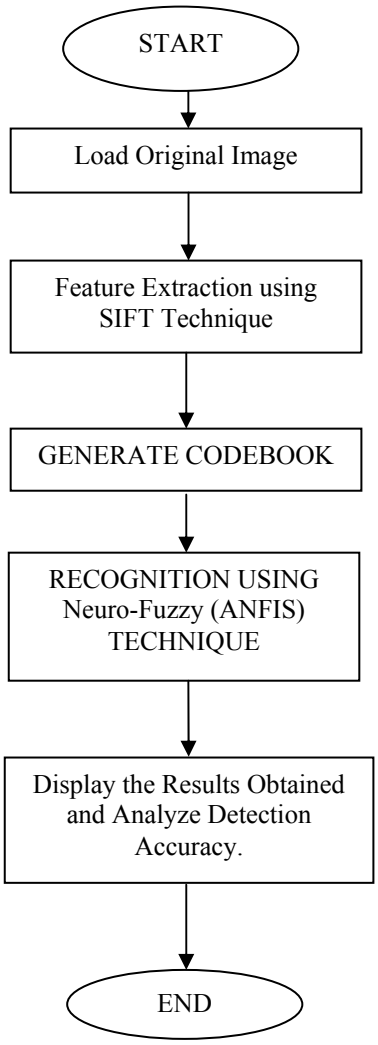

Figure 1: Flowchart of Proposed work

Phase 3: Develop a code for the Neural Network and apply this on the loaded document image.

Phase 4: Code is developed to analyze final CAP documents obtained using various parameters like Scalability and detection accuracy.

VI. RESULTS AND DISCUSSION

In the following figures, result of proposed algorithm is highlighted.

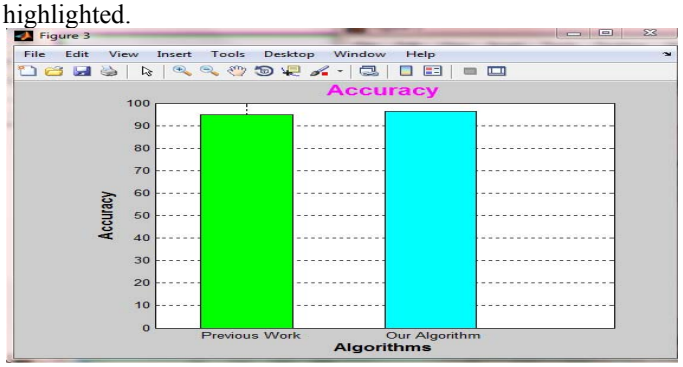

Figure 2:

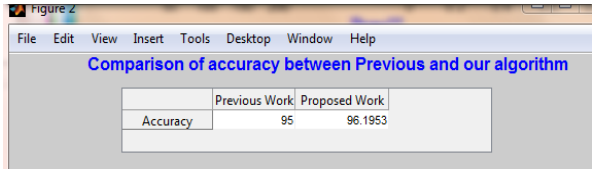

Figure 3:

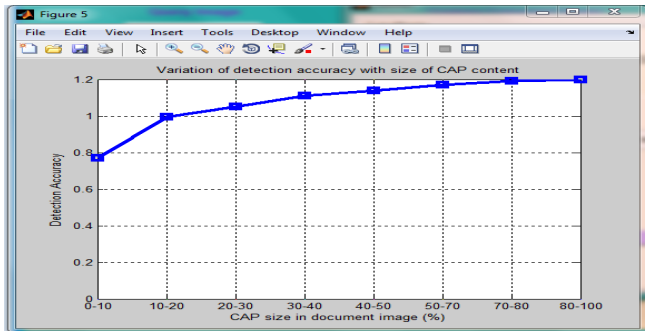

Figure 4:

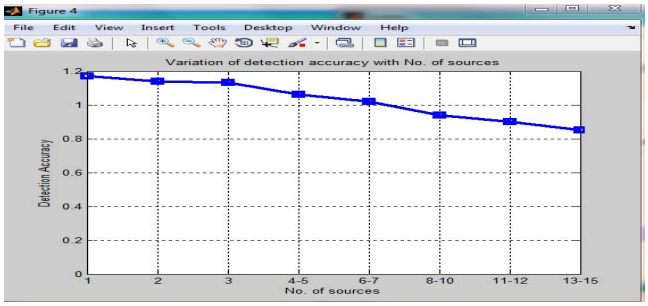

Figure 5:

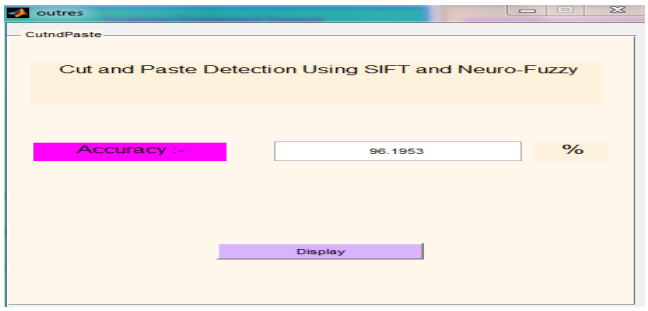

Figure 6:

CONCLUSION

In this research, Gaussian mixture model is used for foreground object estimation in which an additional step of filtering by median filter is incorporated to remove noises. Moving purpose classification algorithm is used separate human being (i.e., pedestrian) from other foreground objects (viz., vehicles). Shape and boundary information is used for this moving target classification. The width vector of outer outline of binary silhouette and Gait PAL AND PAL ENTROPY coefficients are used for extracting the feature vector. Extracted feature vectors are used to recognizing individual identification. The Surf Feature is also used for recognizing persons which are based on gait. There are various parameters like distance between hand and distance between legs are calculated .Finally SVM and K-means results are calculated which is far better in comparison to previous research paper.

Future work will be the combination of some other biometric technique with present gait recognition technique to make identification more accurate and for more secure authentication.

ACKNOWLEDGMENT

Thanks to my Guide and family member who always support, help and guide me during my dissertation. Special thanks to my father who always support my innovative ideas.

REFERENCES

[1] Y. Cao, T. Gao, L. Fan and Q. Yang, “A Robust Detection Algorithm for Copy-Move Forgery in Digital Images”, Forensic Science International, vol.214, Jan. 2012, pp. 33-43.
[2] M. Hussain,K. Khawaji,G. Bebis and G. Muhammad, "Passive Copy Move Image Forgery Detection Using Undecimated Dyadic Wavelet Transform", Digital Investigation, vol. 9, 2012, pp. 49-57.
[3] R. Shekhar and C. V. Jawahar, “Word image retrieval using bag of visual words,” in DAS, 2012.
[4] T. M. Rath and R. Manmatha, “Word spotting for historical documents,” IJDAR, 2007.
[5] A. Kumar, C. V. Jawahar, and R. Manmatha, “Efficient search in document image collections,” in ACCV, 2007.
[6] K. Takeda, K. Kise, and M. Iwamura, “Real-time document image retrieval on a smart phone,” in DAS, 2012.
[7] T. Nakai, K. Kise, and M. Iwamura, “Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval,” in DAS, 2006.
[8] [Online]. Available: http://www.google.co.in/mobile/goggles/
[9] R. I. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.
[10] V. Blankers, C. Heuvel, K. Franke, and L. Vuurpijl, “ICDAR 2009 signature verification competition,” in ICDAR, 2009.
[11] M. Malik, S. Ahmed, A. Dengel, and M. Liwicki, “A signature verification framework for digital pen applications,” in DAS, 2012.
[12] S. Srihari, “Evaluating the rarity of handwriting formations,” in ICDAR, 2011.
[13] C. Su and S. N. Srihari, “Evaluation of rarity of fingerprints in Forensics,” in NIPS, 2010.
[14] J. van Beusekom and F. Shafait, “Distortion measurement for automatic document verification,” in ICDAR, 2011.
[15] H. Li, “Two-view motion segmentation from linear programming relaxation,” in CVPR, 2007.
[16] M. Iwamura, T. Kobayashi, and K. Kise, “Recognition of multiple characters in a scene image using arrangement of local features,” in ICDAR, 2011.
[17] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in ICCV, 2003.
[18] M. Muja and D. G. Lowe, “Fast approximate nearest neighbours with automatic algorithm configuration,” in VISSAPP, 2009.