



Cut-And-Paste Detection on Document Images and Document Image Retrieval -A Review

Harmandeep Kaur*

Department of CE, Y.C.O.E,
Punjabi university, India

Ashok Kumar Bathla

Department of CE, Y.C.O.E,
Punjabi university, India

Abstract- Many document image collections are now being scanned and made available over the internet or in digital libraries. Effective access to such information sources is limited by the lack of efficient retrieval schemes. The use of text search methods requires efficient and robust optical character recognizers (OCR), which are presently unavailable for Indian languages. Another possibility is to search in the image domain using word spotting. First, words are automatically segmented. Then features are computed at word level and indexing is performed. Word retrieval is done efficiently by using an appropriate technique. Document image retrieval is a task of searching document images relevant to a user's query. With the emergence of large document repositories, many new documents get created by cut and paste (CAP) of documents. Often these could also lead to unethical CAP. There can be two types of document forgeries- recognition based and recognition free. There exist many software tools for text based document comparison. However, text is not always available. So, recognition free CAP detection on document images is used where text is not available. Many techniques for extracting the features of document images are used. Indexing is done to retrieve the document images. In this work, forgery and plagiarism are detected in document images by detecting CAP content in it. Feature extraction is done using SIFT. Key point - based feature method SIFT is robust to rotation and scaling.

Keywords— Document Image retrieval, OCR, CAP detection, plagiarism, SIFT.

I. INTRODUCTION

The advent of digital cameras and cheap commodity hardware, has paved the way for an exponential growth in the amount of image and video content generated. Retrieval of electronic documents is necessary for efficient management of a large scale document database. Electronic documents can be retrieved by keywords when their textual contents are available. On the other hand, in order to retrieve documents whose textual contents are not available, another technique such as document image retrieval is needed. Document image retrieval is the task of retrieving documents represented as images by the query of scanned or captured document images. Cut and paste detection in document images is a task of detecting the parts of document images which are copied from other document images. Cut and paste detection is somewhat similar that of retrieving similar documents giving a query document.[5] This document retrieval is of two types- recognition based and recognition free. Recognition based document retrieval is based on OCRs and it is fairly advanced. There exist many software tools for recognition based document image retrieval. However, text is not always available. So, recognition free document image retrieval is used. Here query can be used as a complete document images or query can be used by specific word examples i.e. a smaller portion of the document image.

II. DOCUMENT IMAGE RETRIEVAL

Document retrieval from digital libraries depends mainly on three components: document storage or indexing, query formulation, and similarity computation with relevant ranking of the indexed documents images. Digital libraries are used to browse collections of documents by searching for individual pages, stored in the image databases, with the help of user interfaces. Optical Character Recognition (OCR) techniques have used that follow the recognition-based framework. The recognition-based approach has some limitations when dealing with documents having a high level of noise, layout variance, or when documents contain multi-lingual text. Several recognition-free approaches have been proposed recently to tackle these problems. Two important factors lead to this research are the increased functionality of modern computers, and interest for the evaluation of historical documents.

In case of recognition-free approaches, the comparability between the indexed documents and the query is calculated at the raw data or at the feature level, avoiding the explicit recognition during the indexing. This can be done by using following techniques-

Word indexing and keyword spotting

Word spotting is a recognition free method of searching and locating words in document images by treating a collection of documents as a collection of word images. These words are then clustered and the clusters are annotated for enabling indexing and searching over the documents. Segmentation of each document into its corresponding lines and then into words is done very carefully. Each document is indexed by the visual image features of the words present in it. The work of Rath and Manmatha [15]

discussed the problem of matching handwritten words in historical documents. They explored different word image matching techniques: translation invariant Euclidean Distance Mapping (EDM), invariant word transformation based on Scott and Longuet-Higgins algorithm (SLH), dynamic time warping (DTW), etc.

Graphical items

The retrieval of graphical items allows the user to recognize interesting documents from a new perspective. Graphical Retrieval techniques are appropriate since graphical symbols can have different sizes and are prone to segmentation problems being frequently connected with other parts of the documents. Logo retrieval and the retrieval of architectural symbols are the popular examples.

Handwriting

The retrieval of handwritten documents at the image or feature level is still at the beginning stage. Large variability in writing style and the large size of the vocabulary is the main problem. Applications are the processing of online handwritten documents [9] and the signature-based document retrieval.

Layout retrieval

Document image retrieval based on layout similarity offers to users a new retrieval strategy that was possible before only by manually browsing documents. The retrieval by layout similarity is similar to the concept of Content Based Image Retrieval (CBIR). Mainly layout retrieval is achieved using a fixed-size feature vector by computing some features in the regions defined by a grid superimposed to the page.

III. CLASSIFICATION METHODS

K-Nearest Neighbour (KNN)

One of the most popular classifiers is a nearest neighbour classifier. Its extension to K-nearest neighbor (KNN) is a supervised learning algorithm in which the new instance query is classified based on majority of the category of K-nearest neighbors [3].

This algorithm is used to classify a new object based on attributes and training samples. Given a query point, we find K number of objects or (training points) closest to the query point. The classification is performed using voting among the classification of the K objects. K Nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance. Distance measures used to obtain the nearest neighbour are Euclidean distance, Bray Curtis distance, Manhattan distance, Minkowski distance, Chebyshev distance, Canberra distance etc [3].

Approximate Nearest Neighbour (ANN)

Computing exact nearest neighbors in high dimensional space is a very difficult job. It has been shown that by computing nearest neighbors approximately, it is possible to achieve remarkably faster running times and a relatively very small errors. Many of the ANN algorithms allow the user to specify a maximum approximation error bound, thus allowing the user to control the trade-off between accuracy and running time [1].

Decision Tree Classifiers (DTC)

This classifier is used to classify samples by a series of successive decisions. Decision Tree Classifiers (DTC's) are used for character recognition, radar signal classification, remote sensing applications, medical diagnosis, expert systems, data mining approaches and speech recognition [3, 11].

Perhaps, the most important feature of DTC's is their capability to break down a complex decision- making process into a collection of simple decisions, thus providing an easier to interpret solution. The basic idea involved in any multistage approach is to break up a complex decision into a union of several simple decisions, wishing the final solution obtained this way would resemble the intended desired solution.

Multi-layer Perceptron (MLP)

We also explore the performance of neural network classifiers. Multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and it is much powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable by a hyper-plane. The supervised learning problem of the MLP can be solved with the back-propagation algorithm.

Convolutional Neural Networks (CNN)

Convolutional Neural Networks are a special kind of multi-layer neural networks [13]. Like MLPs, they are also trained with a version of the back-propagation algorithm. But they differ in the architecture of the network. Convolutional Neural Networks are designed to recognize visual patterns directly from pixel images with minimal preprocessing. They can identify patterns with extreme variability such as hand-written characters, and these are robust to distortions and simple geometric transformation.

Support Vector Machines (SVM)

SVMs have received considerable attention in recent years. SVMs are a set of related supervised learning methods used for classification and regression. Given a set of points belonging to two classes, a Support Vector Machine (SVM) finds the hyper-plane that separates the largest possible fraction of points of the same class on, while it maximize the distance from one class to the hyper-plane. SVMs are used to minimize the structural. SVMs use the positive and negative samples for a class to find the support vectors, the samples that have high probability of getting misclassified. It has high generalization capability.

SVMs are of two possible types. First one computes the majority of all the classifiers. We mention to this as SVM-1. The second SVM classifier (SVM-2) integrates the decisions using a decision oriented acyclic architecture (DDAG) [4, 6].

IV. INDEXING AND RETRIEVAL TECHNIQUES

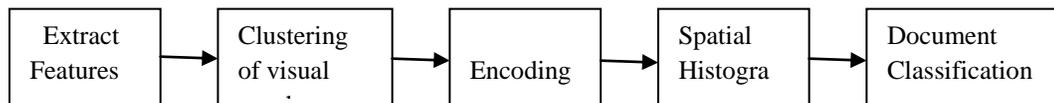
LSH: Locality-sensitive hashing (LSH), proposed by Indyk & Motwani [7], is an approximate similarity search technique that works efficiently even for high-dimensional data. Here indexing is done using locality sensitive hashing (LSH) - a technique which computes multiple hashes - using word image features calculated at word level. Efficiency and scalability is attained by content-sensitive hashing implemented through approximate nearest neighbor computation.

The index is built by hashing word level features of document images. These features are then hashed using content sensitive hash functions, such that the likelihood of finding words with similar content in the same bucket is of great extent. These content sensitive hash functions are used to query similar words during the search. Content-sensitive hash functions are used to hash the features such that similar word images are grouped in the same index of the hash table.

LLAH: In LLAH, a document image is transformed into a set of feature points. After that, features are calculated from arrangements of the feature points. Every feature point in the image is stored into the document image database using its feature. During the retrieval process, the document image database is accessed with features to retrieve images by voting. When a database is scaled up, a large amount of memory is required and retrieval accuracy drops due to insufficient discrimination power of features [14]. To solve these problems, three improvements are made: memory reduction by sampling feature points, making improvement of discrimination power by increasing the number of feature dimensions and stabilizing features by reducing redundancy [18].

Bag of visual Words Method for CAP detection:

The bag of visual words model in computer vision is inspired from the bag of words model in the text domain where a document is represented as an unordered collection of words [16]. The typical Bag of Words (BoW) pipeline is composed of the following steps:



Document is formally represented with the help of frequency of occurrences i.e. histograms of the words in the vocabulary. These histograms are then used to perform document classification and retrieval. Parallely, an image is represented by an unordered set of non-distinctive discrete visual features. Features are computed using scale invariant feature transform (SIFT).

The bag of visual words model has been first introduced by Sivic and Zisserman [17] and Csurka et al. [2]. It describes images as sets of elementary local features called visual words. The whole set of visual words is called the visual vocabulary. The representation of an image database using bags of visual words relies on two steps:

1. Construction of a visual vocabulary.
2. Description of images using this vocabulary.

FLANN-

The accuracy of the approximation is measured in terms of precision, which is defined as the percentage of query points for which the correct nearest neighbor is found. In our experiments, one of two algorithms obtained the best performance, depending on the dataset and required precision. These algorithms used either the hierarchical k-means tree or multiple randomized kd-trees. We can see that for the cases when the build time or the memory overhead had the highest weight, the

algorithm chosen was the kd-tree with a single tree because it is both the most memory efficient and the fastest to build. When no importance was given to the tree build time and the memory overhead the algorithm chosen was k-means [12].

V. CONCLUSION

Today information technology has proved that there is a need to store, query, search and retrieve large amount of electronic information efficiently and accurately. Many documents are created by cut and paste of existing documents. So there is a need to detect cut and paste in document images efficiently. 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. This paper surveys various methods to retrieve the document images and to detect the forgeries in the document images. Thus there is a need to combine different classifiers with different feature vectors in future work to enhance performance.

REFERENCES

- [1] Arya, S., Fu, H.Y.A., "Expected-case complexity of approximate nearest neighbor searching," in proceedings of Symposium on Discrete Algorithms, 2000, pp. 379–388.
- [2] Csurka, G., Christopher, R., "Visual categorization with bags of keypoints.", In ECCV: Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, May 2004.
- [3] Duda, R.O., Hart, P.E., Stork, D.G., "Pattern Classification", Wiley-Interscience Publication, 2000.
- [4] Duan, K.B., Keerthi, S.S., "Which is the best multiclass svm method? an empirical study," in proceedings of Multiple Classifier Systems(MCS), June 2005, pp. 278–285.
- [5] Gandhi, A., Jawahar, C.V., "Detection of Cut-And-Paste in Document Images," in International Conference on Document Analysis and Recognition, 2013.
- [6] Hsu, C.W., Lin, C.J., "A comparison of methods for multiclass support vector machines," in IEEE Transactions on Neural Networks, vol. 13, 2002, pp. 415–425.
- [7] Indyk, P., Motwani, R., "Approximate nearest neighbor towards removing the curse of dimensionality", In *Proceedings of Symposium on Theory of Computing*, 1998.
- [8] Iwamura, M., Kobayashi, T., Kise, K., "Recognition of multiple characters in a scene image using arrangement of local features," in ICDAR, 2011.
- [9] Jain, A. K., Nambodiri, A. M., "Indexing and Retrieval of On-line handwritten documents", Proc. 7th ICDAR, 2003, pp. 655–659.
- [10] Kumar, A., Jawahar, C.V., Manmatha, R., "Efficient search in document image collections," in Asian Conference of Computer Vision, 2007.
- [11] Lim, T.S., Loh, W.Y., Shih, Y.S., "An empirical comparison of decision trees and other classification methods," Tech. Rep. 979, Madison, WI, 30 1997. 44
- [12] Muja, M., Lowe, D.G., "Fast approximate nearest neighbors with automatic algorithm configuration," in VISSAPP, 2009.
- [13] Nakei, T., Kise, K., Iwamura, M., "Real-Time Retrieval for Images of Documents in Various Languages using a Web Camera", Proceedings of the 10th international Conference on Document Analysis and Recognition (ICDAR2009), pp. 146-150 (2009).
- [14] Pansare, A., Bhatia, S., "Handwritten Signature Verification using Neural Network", International Journal of Applied Information Systems, Vol. 1, pp: 44-49, 2012.
- [15] Rath, T., Manmatha, R., "Word image matching using dynamic time warping" Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 521-527, 2003.
- [16] Shekhar, R., Jawahar, C.V., "Word image retrieval using bag of visual words," in Document Analysis System, 2012.
- [17] Sivic, J., Zisserman, A., "Video Google: A text retrieval approach to object matching in videos.", In Proceedings of ICCV, volume 2, pages 1470–1477, Nice, France, 2003.
- [18] Takeda, K., Kise, K., Iwamura, M., "Real-Time Document Image Retrieval for a 10 Million Pages Database with a Memory Efficient and Stability Improved LLAH", International Conference on Document Analysis and Recognition, 2009, pp. 1520-5363.