Intelligent content based Logo Retrieval System for Industrial Application

Kunal Dilip Chandiramani, Rishabh Verma
SCSE, Vellore Institute of Technology

ABSTRACT

- In industries categorization of invoices in a major problem which happens manually and consumes a lot of time.
- This problem can be eradicated by automating the process by categorizing these invoices using content based image retrieval techniques.
- In this paper our main objective is to develop an efficient system to extract the logo out of the invoices and identify the organization it belongs to.
- To develop this system, we have used OpenCV library in python. Furthermore, the system will use efficient searching algorithm to reduce the time complexity.
- This system will not only significantly reduce the time to categorize invoices manually but also will help reduce the human errors.

CONTACT

Rishabh Verma
Vellore Institute of Technology, Chennai
Email: rishabh.verma.rsv@gmail.com
Phone: 7550170483

Kunal Dilip Chandiramani
Vellore Institute of Technology, Chennai
Email: kunal.kc.chandiramani@gmail.com
Phone: 7972366632

INTRODUCTION

- Logos are very useful for the categorization of documents, especially in business and administrative documents. They allow us to quickly determine the source of the documents and accurately with low costs. In recent years, the explosion of the amount of digital documents poses challenges for the categorization and indexing of digital documents based on their origins.
- There are many challenges associated with performing efficient and accurate logo retrieval in document images. Documents are often binary images that preclude many texture based features. The binarization of the images adds noise that can distort the original logo.
- We present an approach for logo extraction and recognition for documents categorization based on Histogram of Gradients(HOG) features and Hue Saturation Value(HSV) based color features.
- First, we will read the invoice and extract the company logo from the document then we will use content based image retrieval technique to match the company logo to the logo in our database and retrieve six relevant images for categorization the documents.

METHODS AND MATERIALS

Pointers in python are used to extract the logo from the invoice and content based image retrieval system to recognize the logo. Two methods for image matching are developed.

Method 01: Color-HSV Based Features- This phase identifies the unique feature vector corresponding to the image features. Hue saturation value (HSV) color space is used for color feature extraction.

Method 02: HOG Based Features-The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image.
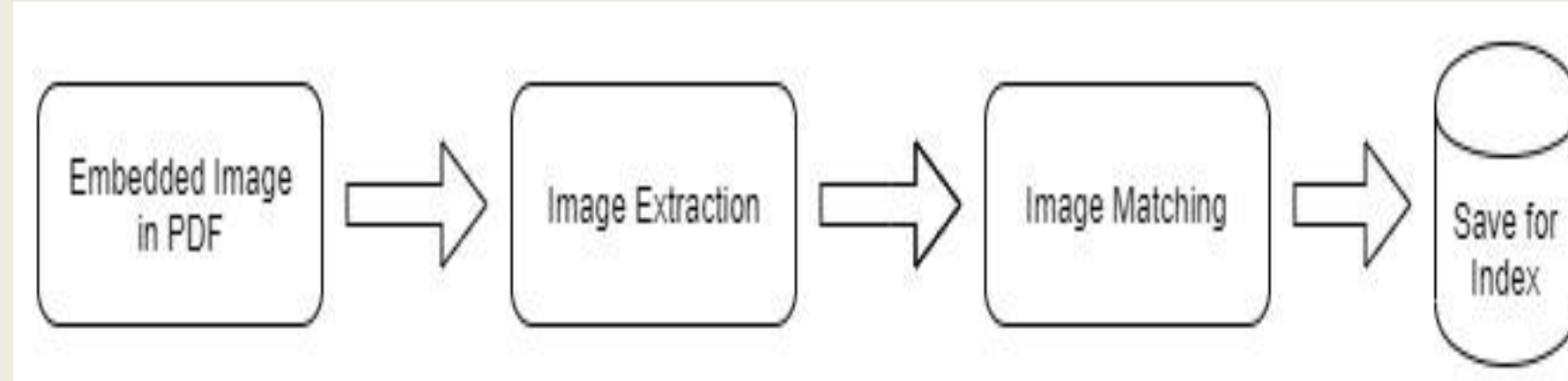
RESULTS

Our experiments were conducted on a set of 1060 Logo images of FlickrLogos-32 dataset downloaded as compressed bundles from Kaggle data feed .

To measure the accuracy of the system , two famous accuracy measures known as precision and recall are used. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved.
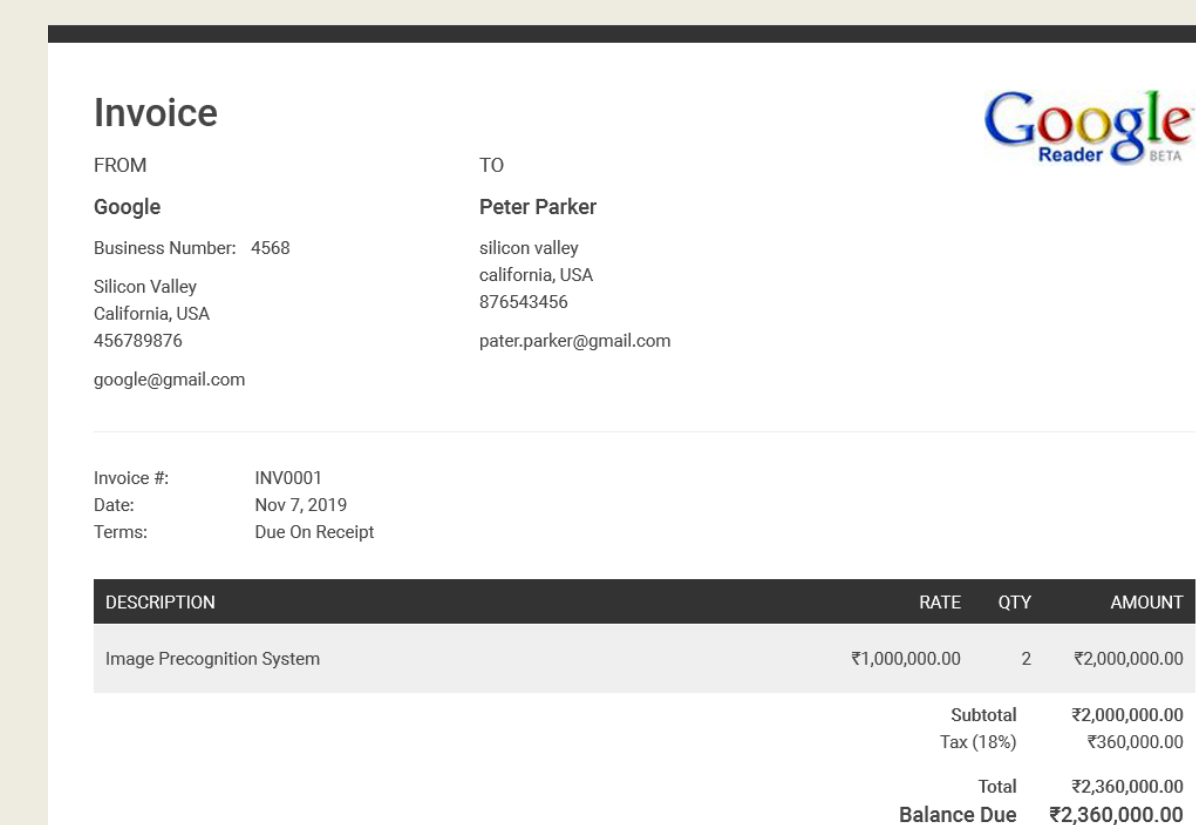
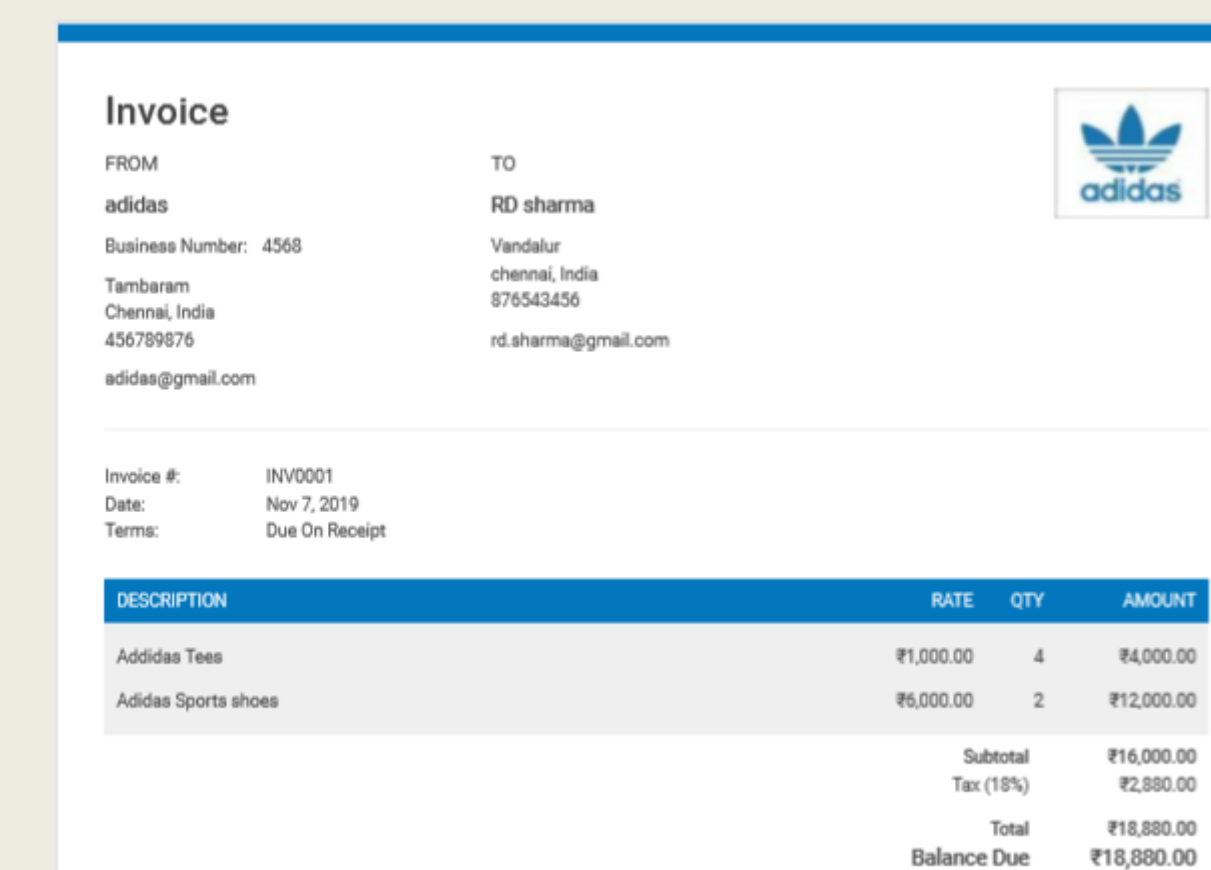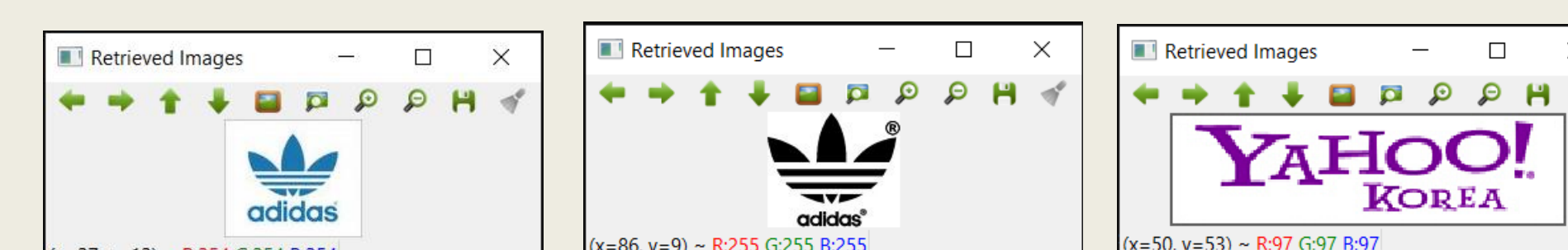**Figure 1.** Query Image.

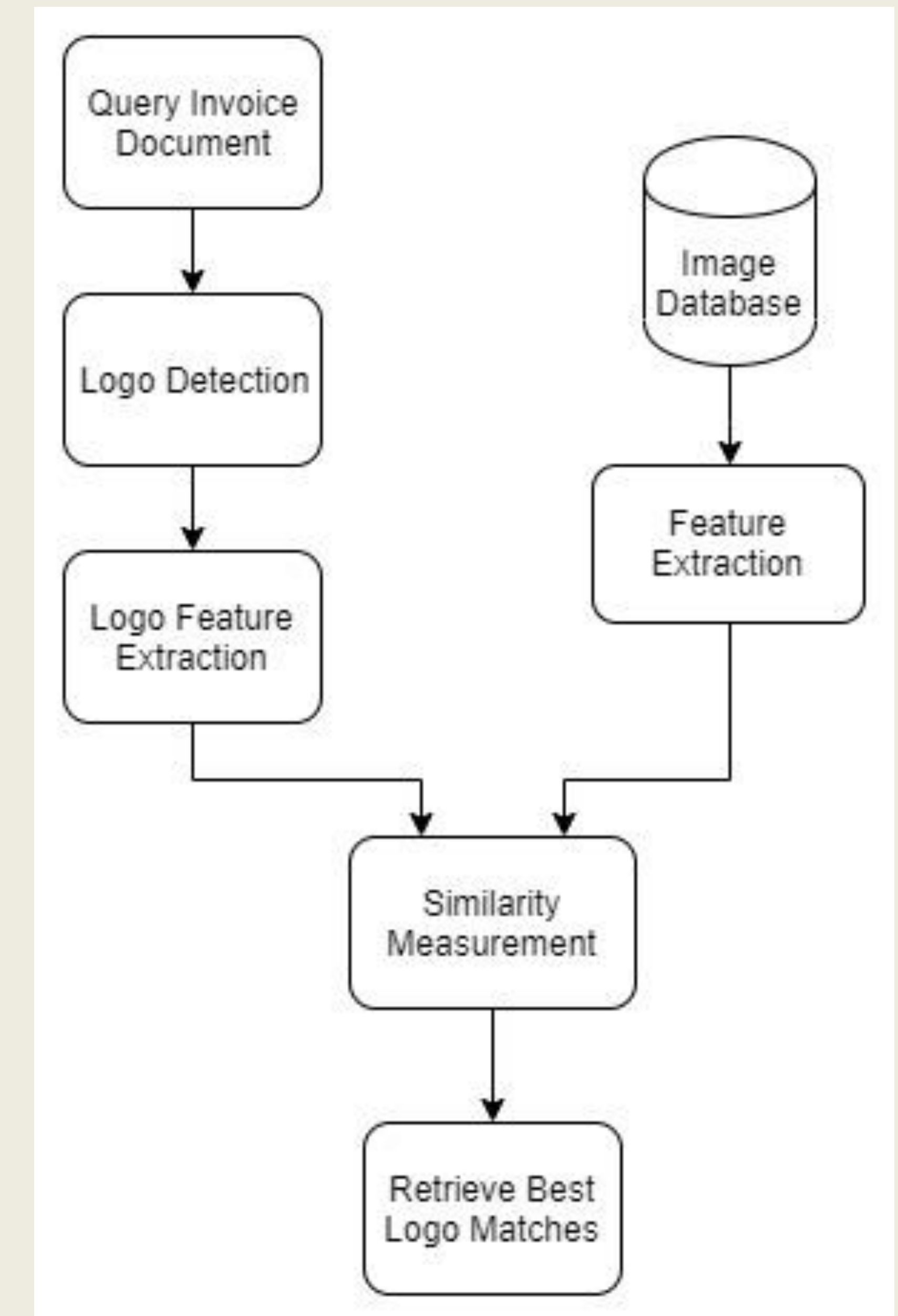Result for Color-HSV Based Model

Figure 1. Query.

Result for Color-HSV Based Model

	Relevant	Irrelevant
Retrieved	3	0
Not Retrieved	3	1057

Table 1. Confusion Matrics for Color-HSV Baed model.

	Relevant	Irrelevant
Retrieved	2	1
Not Retrieved	3	1057

Table 2. Confusion Matrics for HOG Baed model.

FLOWCHART

CONCLUSIONS

In this paper we have proposed an efficient invoice categorization system by extracting logo from these invoices and using content based image retrieval system to recognize it. To develop this system we have used two different types of content based image retrieval method to recognize the logo after extracting it from the query invoice: We discard the remaining images of the invoices for which no match is found using our method. Although both the method performed well for the system but using HOG based features for logo detection has a problem, Histogram of gradient based method is not scale invariant. In histogram of gradients based method the system calculates the features based on the size of the image i.e if the size of an image is big the length of the feature vector is large while if the size of the same logo image is small the length of the feature vector is small. Because of this issue in HOG based method for logo detection, Hue saturation value based method performs better for intelligent logo retrieval system since it is scale invariant.

REFERENCES

1. Xu, C., Tang, Z., Tao, X., Shi, C., "Graphic composite segmentation for PDF documents with complex layouts," Document Recognition and Retrieval XX, Proceedings of SPIE 8658, 86580E1-86580E10 (2013)

2. Shao, M., Futrelle, R., P., "Recognition and classification of figures in PDF documents," in Graphics Recognition. Ten Years Review and Future Perspectives. LNCS, 239-251 (2006).