Intelligent content based Logo Retrieval System for Industrial Application

Kunal Dilip Chandiramani, Rishabh Verma
School of Computer Science Engineering
Vellore Institute of Technology,
Chennai, India

***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 will use 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.**

1. Introduction

Logos (as well as seals) 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. First, documents are often binary images that preclude many texture based features. Second, the binarization of the images adds noise that can distort the original logo. Third, scanned document images are typically high resolution images ranging from 2 to 5 megapixels and logos can be comprised of less than 1% of the document's surface areas. Fourth, all of the current approaches rely heavily on training on the logos of interest. In a typical retrieval scenario, however, the logos being queried for are not known ahead of time.

There are some problems that every logo detection method should deal with. There are often different versions of the same logo. This includes logos with or without text, differently colored logos or logos that were changed over time. Logos can be placed on different surfaces made of different materials, including textile, paper, computer screens, glass, and metals. These materials look differently and some of them create an additional obstacle by being reflective. Logos are often placed on a curved or another non-planar surface, for example on water bottles or cans, hats, balls or just on a curved piece of textile. Such features as a position of the logo and illumination also play a significant role in how easy or hard it is to detect a logo.

In this paper, 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.

This paper is organized as follows. In section II., we will discuss about the existing literature. In section III., we report our method to extract company logo and recognize it. Finally, we present the experimental results in section IV. and draw the conclusions in the last section.

2. Literature Review

In recent years, there have been a number of papers about the related topics of logo detection, recognition, and retrieval for document images. Given a document image, logo detection can be defined as the problem of finding a logo's boundary on the page without regard to class. Logo recognition (or matching) on the other hand is the problem of determining which class a given logo belongs to. Logo retrieval can be viewed as a combination of the two problems where one wants to simultaneously detect and recognize a logo across a dataset given some query image.

Extracting images from PDF documents has been a topic of research interest. Images that include complex vector graphic elements, text, and other pictorial graphic elements are particularly challenging. Xu[1] proposed a method to segment graphics embedded in a PDF document using a layer based document analysis method. Shao[2] proposed a method to recognize and classify diagrams in vector-based PDF documents. Lin[3] proposed a method to identify mathematical formulas using rule-based and

learning-based methods. The solutions reported in these articles are specific to their problem domain, and software tools and algorithms are either not available or easily replicable. We aim to use off-the-shelf tools to develop reliable algorithms for extraction and labeling of complex graphics from PDF documents.

As a consequence, many research works in the field of logo recognition have been carried out. In particular, Doermann et al. [4] use a combination of text, shape, and global and local affine invariants for logo recognition. Meanwhile, Zhu et al. [5] present an approach using a multi-scale boosting strategy to detect and extract logo(s) in document images. At a coarse image scale, a Fisher classifier provides an initial classification. Then, each logo candidate region is further classified at a finer image scales by a cascade of simple classifiers.

Jain and Doermann [6] present an approach for logo retrieval without segmentation. SURF features are used for logo retrieval; and they propose an indexing technique to group feature vectors and a filter method based on the properties of the features orientation and their geometric characteristics. Meanwhile, Rusinol and Llados [7] introduce a method for organizing and indexing logos based on describing logos by a variant of the shape context descriptor.

In another paper of Rusinol and Llados [8], they propose a logo spotting method where the logo image and the query documents image are described by a set of SIFT features describing key-points. A bag-of-word model is further used for matching. In order to filter the matching key-points and consider only the key-points belonging to the logo in the query document, they consider clusters of key points.

3. Implementation

In this section we describe our methods for extraction and labeling of logos extracted from invoices in PDF format. The system can be categorized into four main phases (1)Invoice Pre-processing (2)Logo Detection (3)Logo Matching (4)Store for indexing as shown in Figure.01.

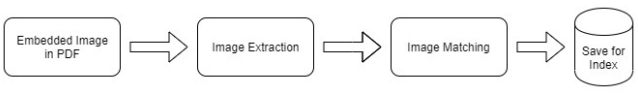

Figure 1: Various Phases in this system

Some of these PDF documents also contain additional images such as borders, publisher banners, and empty spacers (blank white graphics) that are of little interest to us. Additionally, as PDF documents can be embedded with images and graphics in different formats, format conversion is performed to convert the extracted images to the JPEG format. Each extracted image is then compared to the labeled low-resolution images provided in the flickr logo data set as part of the article bundle. The extracted image is labeled when a match is found. We have developed two methods for image matching: one is based on Hue saturation value based features, and the other uses Histogram of Gradients based features. We discard the remaining images for which no match is found using our method.

3.1. Database

The dataset FlickrLogos-32 contains photos showing brand logos and is meant for the evaluation of logo retrieval and multi-class logo detection/recognition systems on real-world images. The dataset consists of total 8240 images and there are 32 different logo brands by downloading them from Flickr. All logos have an approximately planar surface. There are 32 classes: Adidas, Aldi, Apple, Becks, BMW, Carlsberg, Chimay, Coca-cola, Corona, DHL, Erdinger, Esso, Fedex, Ferrari, Ford, Foster, Google, Guiness, Heineken, HP, Milka, Nvidia, Paulaner, Pepsi, Ritter Sport, Shell, Singha, Starbucks, stellaArtois, Texaco, Tsingtao and UPS.

3.2. Working

In this section of the paper we will elucidate the detail working of our system starting from reading the query invoice to identifying the company it belongs to. First, we have used pointers in python to extract the logo from the invoice and used a content based image retrieval system to recognize the logo. Next, We have developed two methods for image matching: one is based on Hue saturation value based features, and the other uses Histogram of Gradients based features. We discard the remaining images for which no match is found using our method. The detail description about the two methods used for image matching is as follows:

3.2.1. Method 01: Color-HSV Based Features. This phase identifies the unique feature vector corresponding to the image features. In this paper, hue saturation value (HSV) color space is used for color feature extraction. The RGB color components in a digital images are directly related to the amount of light hitting on the objects. Therefore the object discrimination process with respect to those components are sometimes become very difficult. Because of that, HSV color space is often used in this work.

3.2.2. 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. This method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. The steps involved in the implementation of HOG Features on an image is as follows:

- Gradient computation: The first step of calculation in many feature detectors in image pre-processing is to ensure normalized color and gamma values.
- Orientation binning: The second step of calculation is creating the cell histograms. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel based on the values found in the gradient computation.
- Descriptor blocks: To account for changes in illumination and contrast, the gradient strengths must be locally normalized, which requires grouping the cells together into larger, spatially connected blocks.
- Block normalization: Dalal and Triggs explored four different methods for block normalization.

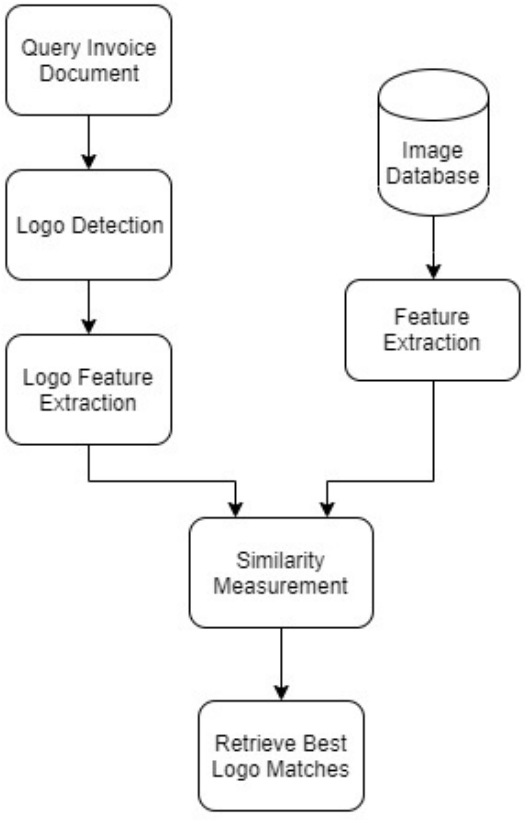

Figure 2: Working of this System

4. Experimental Results

In this section we briefly describe performance evaluation method and experimental result of our proposed method. Our experiments were conducted on a set of 1060 Logo images of FlickrLogos-32 dataset downloaded as compressed bundles from Kaggle data feed (https://www.kaggle.com/hsankesara/flickr-image-dataset). The dataset FlickrLogos-32 contains photos showing brand logos and is meant for the evaluation of logo retrieval and multi-class logo detection/recognition systems on real-world images. The evaluation metric used to calculate the accuracy of the system is as follows:

4.1. Evaluation Metrics

4.2. Efficiency Results

We have used two methods for image matching: one is based on Hue saturation value based features, and the other uses Histogram of Gradients based features. The results for both the methods are as follows:

4.2.1. Method 01: Color-HSV Based Features. To check the accuracy of the system, this method was experimented on multiple company invoices. The query invoice given to the system is shown in Figure.3 and the results of intelligent logo retrieval system using Hue Saturation value as feature vector as as shown in Figure.4.

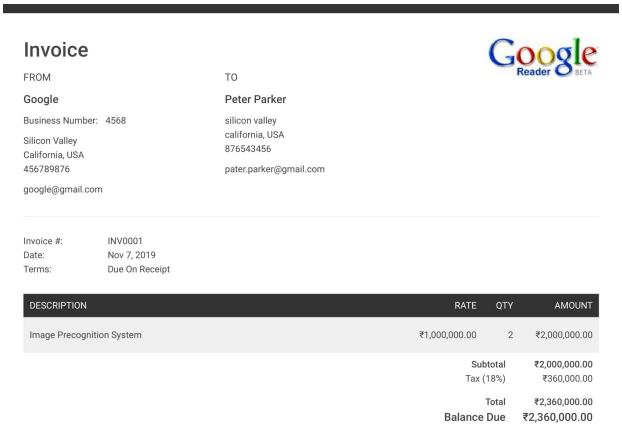

Figure 3: Query Invoice

Figure 4: Retrieved Images for method 01.

TABLE 1: Confusion Matrix for Method 01.

	Relevant	Not-Relevant
Retrieved	3	0
Not-Retrieved	0	1057

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} = \frac{3}{3} = 1.0 \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} = \frac{3}{3} = 1.0 \quad (2)$$

4.2.2. Method 02: HOG Based Features. To check the accuracy of the system this method was experimented on multiple company invoices. The results of intelligent logo retrieval system using Hue Saturation value as feature vector as as shown in ..

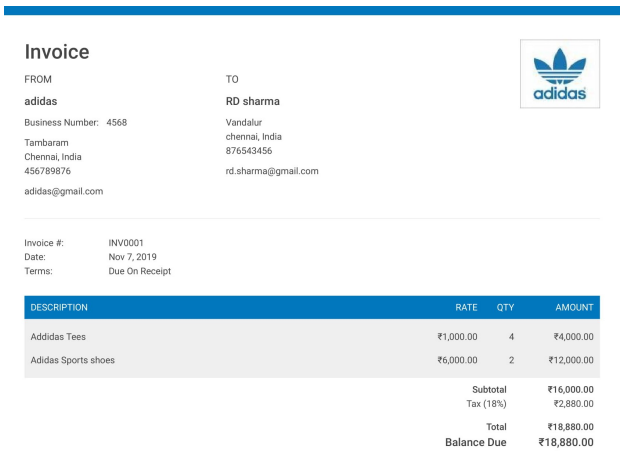

Figure 5: Query Invoice

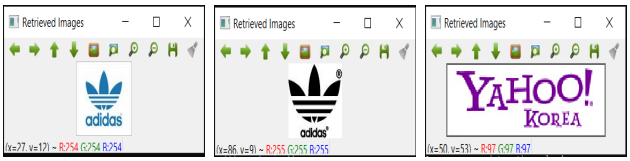

Figure 6: Retrieved Images for method 02.

TABLE 2: Confusion Matrix for Method 02.

	Relevant	Not-Relevant
Retrieved	2	1
Not-Retrieved	1	1057

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} = \frac{2}{3} = 0.67 \quad (3)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} = \frac{2}{3} = 0.67 \quad (4)$$

Acknowledgments

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr. Geetha S, School of Computer Science and Engineering, for her consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We express our thanks to our Program Chair Dr. Justus S. (B.Tech CSE) for his support throughout the course of this project. We also take this opportunity to thank all the faculty of the school for their support and their wisdom imparted to us throughout the course. We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

Conclusion

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: one method uses Hue saturation value based features in the system, and the other uses Histogram of Gradients based features. 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).

[3] Lin, X., Gao, L., Tang, Z., Lin, X., Hu, X., "Mathematical formula identification in PDF documents," 2011 International Conference on Document Analysis and Recognition, 1419-1423 (2011)

[4] D. Doermann, E. Rivlin I. Weiss. Logo Recognition Using Geometric Invariants. International Conference on Document Analysis and Recognition (ICDAR), pp. 894-897, 1993.

[5] G. Zhu and D. Doerman. Automatic Document Logo Detection. ICDAR, pp. 864–868, 2007.

[6] R. Jain and D. Doermann. Logo Retrieval in Document Images. 10th IAPR International Workshop on Document Analysis Systems (DAS), pp. 135-139, 2012.

[7] M. Rusinol and J. Llados. Efficient Logo Retrieval Through Hashing Shape Context Descriptors. DAS 2010, pp. 215-222, 2010.

[8] M. Rusinol and J. Llados. Logo Spotting by a Bag-of-words Approach for Document Categorization. ICDAR, pp. 111-115, 2009.