Automation of Business Cards

Shreya Srivastava, Suryanshu Sahay, Deepti Mehrotra and Vikas Deep

Abstract Business card is shared as hardcopy, the data present in business card will be highly useful if it is available in digital format. The task of manually entering the details of all business cards is laborious and time-consuming. Document image analysis is used in this paper for automating this process. This will be accomplished by performing OCR and then using the text to extract the Meta data. One more important component of business card is the logo of the organization. The text extraction OCR will be done using the Tesseract API. After conversion of the image to text, the data will be saved in the database. The raw data will be saved in the database, which will later be segregated and stored in the appropriate fields. It is generally ignored in the process of saving text information, in this paper it is extracted and stored in database. For logo detection various techniques like Gabor Filter, Harris edge detection technique, MSER, etc., are compared to determine the best technique for acquiring the most accurate logo extraction. Gabor filter gives the best result is used for the extracting logo and storing in database. Java language on NetBeans IDE platform which use the Spring MVC framework is used to implement this work.

Keywords Business card · OCR · Tesseract · Logo detection

S. Srivastava (✉) · S. Sahay · D. Mehrotra · V. Deep
Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: shreya96srivastava@gmail.com

S. Sahay
e-mail: suryanshusahay19@gmail.com

D. Mehrotra
e-mail: mehdeepti@gmail.com

V. Deep
e-mail: Vikasdeep8@gmail.com

M. Kumar et al. (eds.), *Advances in Interdisciplinary Engineering*, Lecture Notes in Mechanical Engineering, https://doi.org/10.1007/978-981-13-6577-5_35

1 Introduction

Business card is an important document which requires preservation and efficient data handling. The limitation of the card being offline and as a result it requires manual work for preserving the data digitally [1]. The greatest advantage of a business card, its tangibility has somehow also become its greatest disadvantage. These are easily accessible yet fragile. This problem has been tried to be solved by automating this process. The automation of business cards will be done by extracting all the data and saving each field into the database using different algorithms and logic using document image analysis [2–5]. This data will be saved to different fields based on the field analysis [6].

The data field extraction will be done to extract Meta data like Name, Company, Phone, Fax, Email, and Address. The rest of the information will be saved in the also be saved. The main contribution to this work was to filter the text by reducing the noise of the obtained text and then obtain the information which is necessary, i.e., categorizing the data into mapped and unmapped data. Mapped data will then be used to segregate the information and classifying it on the basis of various algorithms. Regular Expressions were used for pattern matching which will be used to extract information like email address, phone numbers, etc., which follow a certain pattern. However there will be exceptions like '@' will also be for the twitter handle which might be interpreted as an email. Core-NER-NLP (Name Entity Recognizer-Natural Language Processing) which uses Stanford Dictionary for various predefined field extraction like name of a person, organization, location, etc. The data extracted from the text will then be saved in separate fields in the data base. As the tesseract OCR is not 100% efficient, there will be a modification option where the user can manually correct the text before the data extraction takes place [7]. The home page will contain upload, modify, list, search, and delete links. Upload tab will be used to upload an image of the business card. Modify will be used to change any text which has been wrongly read by OCR. List will display all the business cards in a single page. Any query based on the business card will be done using the Search tab. For omitting/removing any uploaded business card, an option of Delete will be present.

This database will be available on the browser using the JDBC and Spring MVC framework. This will be a service available for various platforms. The database will be tried for being converted to contacts which will be directly saved in the mobile device.

2 Methodology

This work uses following APIs and software's for feature extraction: Stanford Core NLP, Tesseract, Oracle Glassfish Server, NetBeans IDE, RegEx, and MATLAB. Stanford Core NLP uses algorithms that easies the way computers are programmed in order to fulfill the humanistic requirements. Tesseract API uses optical character

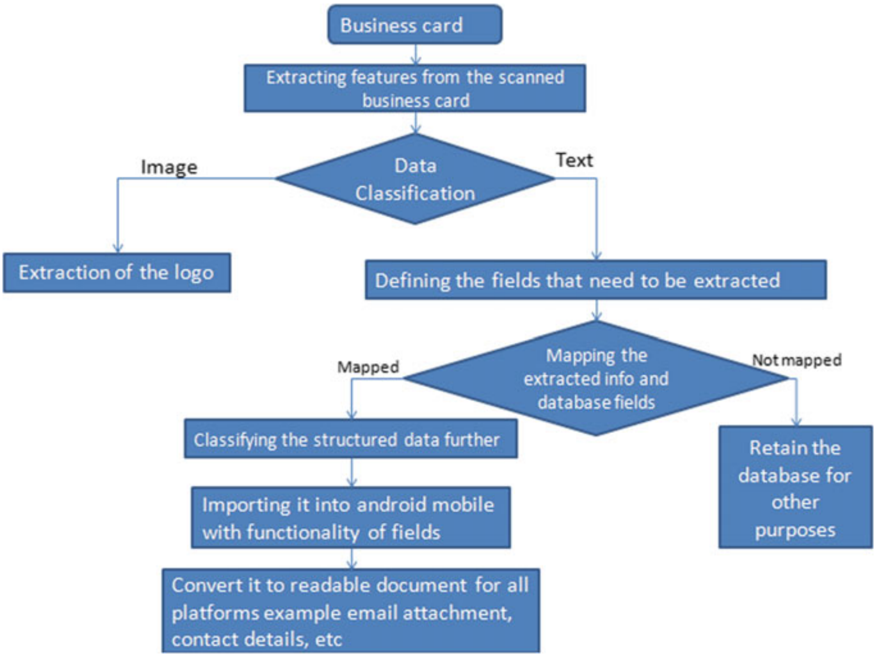
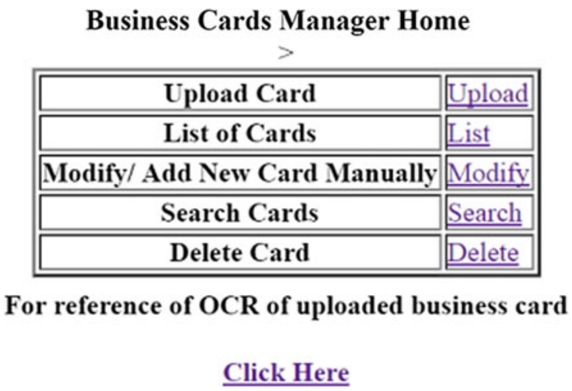

Business card
Extracting features from the scanned business card
Data Classification
Image
Text
Extraction of the logo
Defining the fields that need to be extracted
Mapping the extracted info and database fields
Mapped
Not mapped
Classifying the structured data further
Importing it into android mobile with functionality of fields
Convert it to readable document for all platforms example email attachment, contact details, etc
Retain the database for other purposes

Fig. 1 Data flow diagram

recognition for extracting text from scanned images. Oracle Glassfish Server serves the purpose of handling servlets and JSP. The following diagram describe the complete work flow of the research work.

Figure 1 as shown explains how the scanned business card will be classified into two parts: image and text. The image, if detected will be extracted and the classified as the logo of the respective institution. Else the text will be mapped and classified under the respective database fields. For instance, email ids, '@' will be used for identification and 10 digit numbers will be used for mobile numbers. There are six fields in which the extracted text will be saved and classified. They are Name, Company, Phone, Fax, Email, and Address.

This work has been created on the NetBeans IDE platform using Java language. Since this is a web application, Glass Fish server was used. Glass Fish is an open-source application server built for the java platform. For the architecture, Spring Web MVC was used. The sole purpose of this is to handle all the HTTP requests as well as responses. The Maven Repository was used for this work as it helped in automatically building the dependencies for any number of times. The Tesseract OCR API was used for reading the text from the images. The acquired text was then stored in the database. It then further processed and filtered into the respective fields. This was done by using RegEx, NER, and Stanford Core NLP.



Business Cards Manager Home

>

Upload Card	Upload
List of Cards	List
Modify/ Add New Card Manually	Modify
Search Cards	Search
Delete Card	Delete

For reference of OCR of uploaded business card

Click Here

Fig. 2 Home page of the application

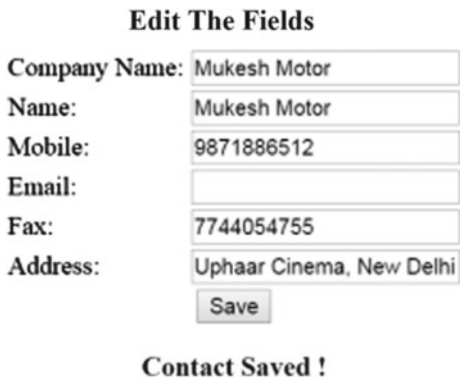
RegEx stands for regular expression. It is used for searching strings which follow a definite pattern. For instance an email id will always have '@' in between the string or a mobile number will have a '+' followed by ISD code and then the remaining 10 digits. There are exceptions in this as well, like for instance '@' might also be present in a twitter handle so.com was used for further classification. Phone numbers are also present in various formats like XXX-XXX-XXXX or + XX-XXXXXXXXXX, etc., which were also dealt with different user cases [8]. NER or Name Entity Recognition is generally used to extract information like names, addresses, percentages, or various quantities. Here it was used to identify the Name of the contact mentioned in the business card. Stanford Core NLP (Natural Language Processing) is an implementation of NER, it is a predefined dictionary which is used to identify and extract information having three major classes (Location, Person, and Organization) this was used to identify the address of the office or institution as duly mentioned in the business card.

The MySQL database was administered and managed by phpMyAdmin. Xampp was used to manage cross-platform operations of majorly database and php.

The class HomeController was created to manage the map the browser requests. The DAO class is used to as a logical interface between the database and the MVC model. The POJO files manage uploading of the business card as well as creating contacts in the database. Utility files include all the text extraction techniques for each field, viz., email, fax, and phone number.

Figure 2 shows the home page of the application. As mentioned, there are options to list all the uploaded cards from the database, manually edit the stored information (see Fig. 3) as the OCRed text is not 100% efficient, search for the business cards using the primary key from a database, and also delete any card.

Dependencies were built separately are done in MVC framework. The main advantage of building dependencies is to eliminate the need to the work of identifying and specifying them. These dependencies are added automatically. It is particularly useful when a large dependency tree is formed as it then becomes difficult to keep a track



Edit The Fields

Company Name: Mukesh Motor

Name: Mukesh Motor

Mobile: 9871886512

Email:

Fax: 7744054755

Address: Uphaar Cinema, New Delhi

Contact Saved !

Fig. 3 Manually editing the fields

on each one of them. The few dependencies that have been built include JAR files like tess4j for OCR extraction, jdbc for java and database connectivity, Stanford-corenlp for the predefined dictionary, spring expression which is used for querying, spring-tx which manages transactions.

For the logo part, MATLAB was used. MATLAB was preferred over Java because the scanned image requires to be processed and since Java Image processing slows down the application, therefore it is best to use the MATLAB in place of Java for image processing. Java Image Processing is useful when the image is only of a few bytes (few hundred dpis), but when larger sized images are being considered for processing, it is best to use platforms which are best suited to process the images as per the requirements. Many different techniques were used in order to identify the logo from business cards. The techniques involved were edge detection, Gabor filter, SURF and SIFT algorithms [9, 10]. All of these techniques are based on the phenomenon of feature extraction. The feature extraction is a method of extracting the important features from the image. These are useful in various fields like object recognition or identifying a particular set of textures or for image retrieval [11, 12].

Edge detection: it is a technique to identify the image boundaries. This will be useful in detecting the logos as it will extract the image present in the image itself. Boundary detection is done by comparing the image brightness and checking its discontinuities [13–15].

SIFT: Scale-Invariant Feature Transformation is one of the basic algorithms used to detect logos. It is very useful as it is invariant to mostly all types of features like scaling, translation, rotation, or even the illumination. It basically converts the image into vectors which are then used for image description.

SURF: Speeded-Up robust Features is an algorithm derived through SIFT for image detection it is better than SIFT in terms of computability and distinctiveness (Fig. 4a, b).

Harris features: This technique is used to detect the corners and identify them (Fig. 5).

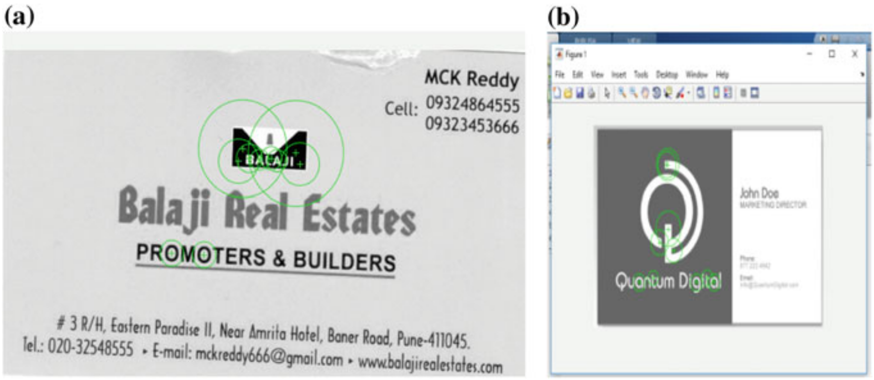

(a)
MCK Reddy
Cell: 09324864555
09323453666
BALAJI
Balaji Real Estates
PROMOTERS & BUILDERS
# 3 R/H, Eastern Paradise II, Near Amrita Hotel, Baner Road, Pune-411045.
Tel.: 020-32548555 • E-mail: mckreddy666@gmail.com • www.balajirealestates.com
(b)
Figure 1
File Edit View Insert Tools Desktop Window Help
Quantum Digital
John Doe
MARKETING DIRECTOR

Fig. 4 **a** Implementation of SURF. **b** Implementation of SURF

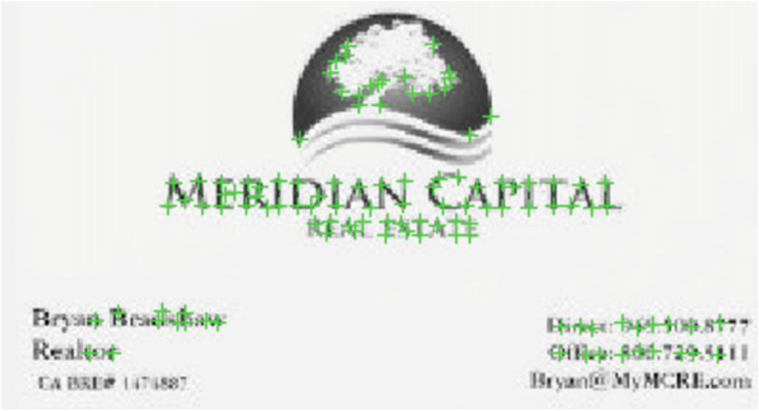


Fig. 5 Harris corner detection

MSER: Maximally Stable External Regions is used for image detection. It deals with the correspondences among the different image components (Fig. 6).

Gabor Filter: It is a filter used to identify the textures present. It serves the purpose of identifying the regions where a certain frequency is present. The main disadvantage of this filter is that it is over reliant on the orientation of the image Fig. 7. The following code snippet was used for implementing the Gabor filter:

Comparing all the features and results, it was Gabor Filter which was found to be most accurate and result-oriented. Others techniques like MSER and Harris corner detection identified the texts also. Thus making it difficult to extract the logo from them.

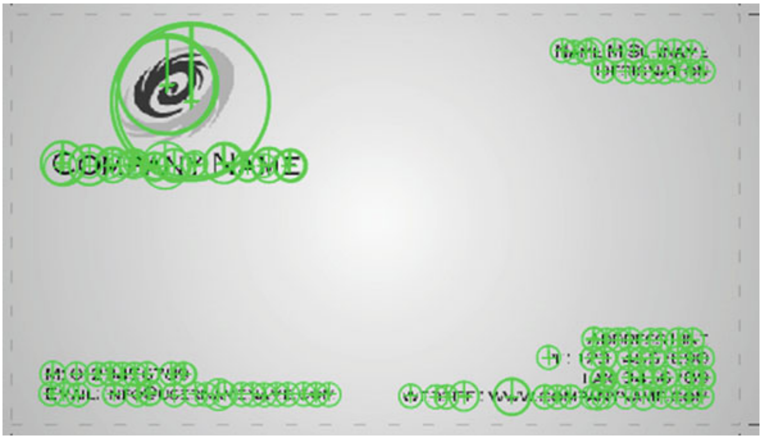


Fig. 6 Implementation of SURF

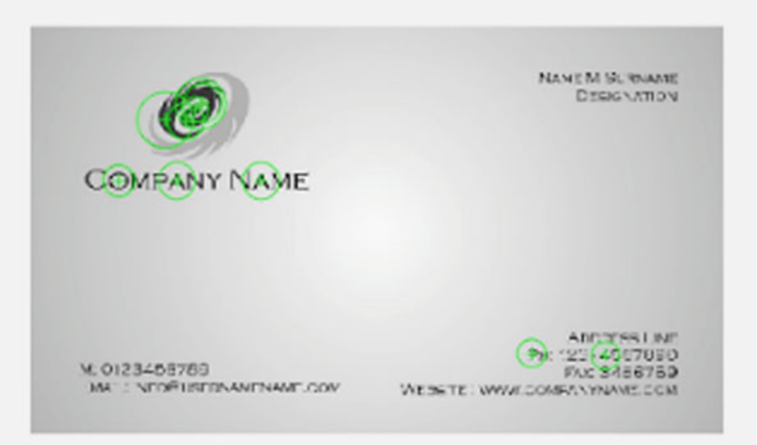
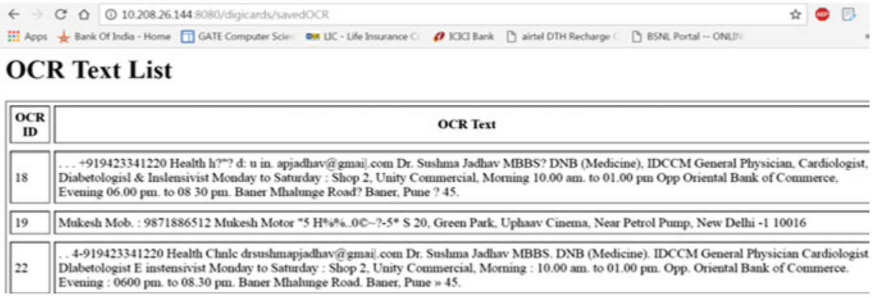


Fig. 7 Implementation of Gabor Filter

3 Results and Discussions

The business card scanned was used for text as well as logo extraction. As explained below the following results were obtained:

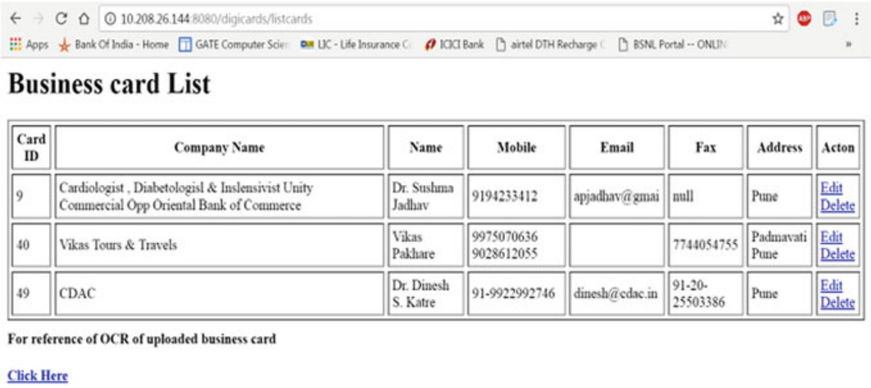
The OCR performed text was retrieved in the database because data other than the structured data might also be useful and hence can be of some purpose as shown in Fig. 8. Figure 9 shows how information has been mapped to the respective fields and thus the data has been retrieved in the database as well. Data has been stored in each category differently.



OCR Text List

OCR ID	OCR Text
18	. . . +919423341220 Health h?"? d: u in. apjadhav@gmail.com Dr. Sushma Jadhav MBBS? DNB (Medicine), IDCCM General Physician, Cardiologist, Diabetologisl & Inslensivist Monday to Saturday : Shop 2, Unity Commercial, Morning 10.00 am. to 01.00 pm Opp Oriental Bank of Commerce, Evening 06.00 pm. to 08 30 pm. Baner Mhalunge Road? Baner, Pune ? 45.
19	Mukesh Mob. : 9871886512 Mukesh Motor "5 H%%..0©~?-5* S 20, Green Park, Uphaav Cinema, Near Petrol Pump, New Delhi -1 10016
22	. . 4-919423341220 Health Chnlc drsushmapjadhav@gmail.com Dr. Sushma Jadhav MBBS. DNB (Medicine). IDCCM General Physician Cardiologist Dlabetologist E instensivist Monday to Saturday : Shop 2, Unity Commercial, Morning : 10.00 am. to 01.00 pm. Opp. Oriental Bank of Commerce. Evening : 0600 pm. to 08.30 pm. Baner Mhalunge Road. Baner, Pune » 45.

Fig. 8 The OCR performed text in the database



Business card List

Card ID	Company Name	Name	Mobile	Email	Fax	Address	Acton
9	Cardiologist , Diabetologisl & Inslensivist Unity Commercial Opp Oriental Bank of Commerce	Dr. Sushma Jadhav	9194233412	apjadhav@gmai	null	Pune	Edit Delete
40	Vikas Tours & Travels	Vikas Pakhare	9975070636 9028612055		7744054755	Padmavati Pune	Edit Delete
49	CDAC	Dr. Dinesh S. Katre	91-9922992746	dinesh@cdac.in	91-20-25503386	Pune	Edit Delete

For reference of OCR of uploaded business card

Click Here

Fig. 9 Business card list

As the scanned cards may vary, they will have different formats like .jpg, .pdf, etc. Hence, ghost jar as well as imagio jar was used for different formats of the uploaded image. This application was run on the glassfish server as all the connectors and backend APIs would work properly and efficiently on this server, therefore, making user-friendly GUI. Thus, the application successfully returns the relevant extraction of data using tesseract and other APIs and techniques.

The logo extraction was done based on the comparison of different techniques; following results were obtained in the end (Fig. 10).

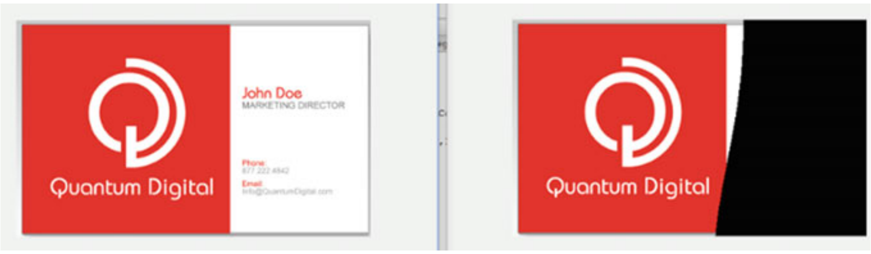


Fig. 10 Logo extraction using Gabor Filter

4 Conclusion

The work is tried and tested on business cards of different types but there was noise which was prevalent in every business card. This was reduced by specifically applying algorithms. As a result only important data was extracted in the database. Since every business card is different therefore it is difficult to extract the information from each business card with absolute efficiency. This can be improved with increase in the dictionary and refinement of algorithms of tesseract, so that the noise can be reduced and other important information can also be retrieved example some business cards may have websites as well as their designation and branches of their company. Therefore, such card's OCR text field extraction would filter out such important details as noise whereas these can also be extracted and saved under more information tab so that correct and detailed data is provided to the user. Along with the important data, the logo was also extracted which being an added feature will help in simpler and easier classification of the contacts. The Gabor Filter was the most relevant technique and as a result successful implementation was done.

References

1. LaForge L, Carlson D, Korver K, System for creating and reading digital business cards, forms, and stationery. U.S. Patent Application No. 10/055,011
2. Carton C, Lemaitre A, Coüasnon B (2015) Automatic and interactive rule inference without ground truth. In: 2015 13th international conference on document analysis and recognition (ICDAR). IEEE
3. Karatzas D et al (2016) Human-document interaction systems—a new frontier for document image analysis. In: 2016 12th IAPR workshop on document analysis systems (DAS). IEEE
4. Niyogi D, Srihari SN, Govindaraju V (1996) Analysis of printed forms. In: Handbook on optical character recognition and document image analysis. World Scientific Publishing Co., Singapore
5. Harvey R, Oliver G (2016) Digital curation. ALA Neal-Schuman
6. Rusinol M, Benkhelfallah T, Poulaind'Andecy V, Field extraction from administrative documents by incremental structural templates
7. Mithe R, Indalkar S, Divekar N (2013) Optical character recognition. Int J Recent Technol Eng 2.1:72–75

8. Zhu G, Bethea TJ, Krishna V (2007) Extracting relevant named entities for automated expense reimbursement. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM
9. Wang H, Chen Y (2009) Log detection in document images based on boundary extension of feature rectangles. IEEE
10. Park JH, Jang IH, Kim NC, Skew correction of business card images acquired in PDA
11. Aksoy S, Haralick RM (1998) Textural features for image database retrieval. Content-based access of image and video libraries. IEEE
12. Conners RW, Harlow CA (1976) Some theoretical considerations concerning texture analysis of radiographic images. In: Proceedings of IEEE conference on decision and control, pp 162–167
13. Flickner M (1993) The QBIC project: querying images by content using color texture and shape. SPIE storage and retrieval of image and video databases
14. Ma WY, Manjunath BS (1997) NETRA: a toolbox for navigating large image databases. In: Proceedings of ICIP
15. Pentland A (1994) Photobook: content-based manipulation of image databases. SPIE storage and retrieval of image and video databases II, pp 34–47, 1994 Feb