Iraqi Journal of Science, 2017, Vol. 58, No.1C, pp: 574-582
DOI:10.24996/ijs.2017.58.1C.19

ISSN: 0067-2904

Auto Crop and Recognition for Document Detection Based on its Contents

Hasanen S. Abduallah*, Nesreen Waleed
Department Computer Science, University of Technology, Baghdad, Iraq

Abstract

An Auto Crop method is used for detection and extraction signature, logo and stamp from the document image. This method improves the performance of security system based on signature, logo and stamp images as well as it is extracted images from the original document image and keeping the content information of cropped images. An Auto Crop method reduces the time cost associated with document contents recognition. This method consists of preprocessing, feature extraction and classification. The HSL color space is used to extract color features from cropped image. The k-Nearest Neighbors (KNN) classifier is used for classification.

Keywords: Color Features, KNN Classifier, Document Image, Auto Crop Method, Signature, Logo, Stamp.

طريقة مقترحة لتميز محتويات الوثيقة بالاعتماد على القطع الالي

حسنين سمير عبد الله*، نسرين وليد عبد الواحد
قسم علوم الحاسبات، الجامعة التكنلوجيا، بغداد، العراق.

الخلاصة

يستخدم طريقة القطع الالي لتحديد و استخراج توقيع، شعار وختم من صور الوثيقة. ان هذه الطريقة تحسن من أداء النظام الأمني المعتمد على التوقيع والشعار والختم وكذلك هذه الطريقة تستخرج الصور من صوة الوثيقة الأصلية والحفاظ على محتوى المعلومات داخل الصور المستخرجه. طريقة القطع الالي يقلل من التكلفة الزمنية المرتبطة لتميز محتويات الوثيقة. تحتوي هذه الطريقة على معالجه اولية، واستخراج الخصائص والتصنيف. تم استخدام التحويل اللوني HSL لاستخراج الصفات لونية من الصورة التي تم قصها. المصنف KNN قد استخدام للتصنيف.

Introduction

Normally a document is a paper that contains a stamp, logo, printed text, handwritten text and signature etc. It's may be in complex or simple form. A wide variety of paper is converted into digital form through a scanner for efficient storage and intelligent processing due to the technological advances [1].The document image analysis distinguishes between a graphic component and the text in a document image and also to extract the information from them [2, 3]. A Region of Interest (ROI) means extraction of a specific area within the image for investigation more closely. To do this, geometries operations are needed to modify the spatial coordinates of the image, the geometry operations include crop, shrink, translate, zoom, enlarge and rotate. The image crop process is selected part of the image and then cutting it away from the rest of the image [4].

Auto Crop process (AC) is the fast procedure to extract the ROI with speed time and reducing the computational complexities [5]. For more specifically, extraction ROI from an image is considered

*Email:nesreen.waleed@gmail.com

the base for further image analysis and classification. In order to crop ROI any type of traditional segmentation process is applied. These identified regions (ROI) are analyzed for using them in a variety of domains such as document analysis and medical image etc. The image crop process is extracted manually by cropping image for further processing but the manual intervention leads to a lot of misinterpretation, if the ROI is not extracted accurately. The accuracy of ROI extraction is played an important role in deciding. Therefore, it is desirable to utilize the method that processes the obstacle to reduce the manual intervention. Usually, ROI is extracted through segmentation methods like threshold, edge detection and morphological operation.

A ROI extraction plays an important role in a security system based document image contents. Auto Crop process provides ROI with reduces the time complexity [5].

1. Related Works

The present work includes significant previous works related to document image recognition and the summarized related work as follows:

In 2013, B.V.Dhandra et al. presented method entitled "***Classification of Document Image Components***" [1]. This paper described method for Classification of Image Components in a document using shape feature, the rule based method and connected component labeling. The KNN classifier was used for classification document into three classes handwritten, printed and seal. This method gave an overall Accuracy of 91.057%.

In 2012, Patil, U. et al. proposed work entitled "***Word Level Handwritten and Printed Text Separation Based on Shape Features***" [6]. This method presents the system for discriminate the handwritten and printed text components based on shape features. The KNN classifier is used for classification. The experimental results have shown Accuracy rate of 98.57%.

In 2012, Zagoris, K. et al. proposed work entitled "***Handwritten and Machine Printed Text Separation in Document Images using the Bag of Visual Words Paradigm***" [7]. This paper presents a method using the Bag of Visual Words (BoVW) model and Scale-Invariant Feature Transform (SIFT) features to separate the handwritten text from machine printed text. The classification was used Support Vector Machine. The method has reported 98.86% and 76.89% accuracy.

2. Theoretical Background

The proposed method has developed a number of different tools to be used for an Auto Crop method could be the following:

3.1 Gray Level Conversion

The gray level process means converting a color image into a gray level image .A gray level image is a two-dimensional array of dots which is called pixels. There is a pixel contains brightness pixels information only and no color information where the range value for brightness pixels [0,255]. Equation 1 is used to convert from RGB to gray image for each pixel:

$$\boldsymbol{Gray\ image} = (0.2989 * R) + (0.5870 * G) + (0.1140 * B) \qquad (1)$$

Where: R=read, G=green, B =blue color components [8].

3.2Noise Reduction

The document is obtained by the use of scanner device that it is usually necessary to eliminate the noise introduced during the acquisition process. The noise attached to a scanned document image has to be removed to avoid errors in the further processing steps. It is achieved by using a 3*3 median filter to generate the cleaned image [9].

3.3 Binarization Conversation

The document image is converted to binary images (image binarization) by using Otsu threshold method which is used to the reduction of a gray level image to a binary image. This method, a nonparametric and unsupervised technique of the automatic threshold selection for image segmentation, the Otsu method assumes that the image to be threshold includes two classes (like background and foreground) then computes the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal [10].

3.4 Edge Detection

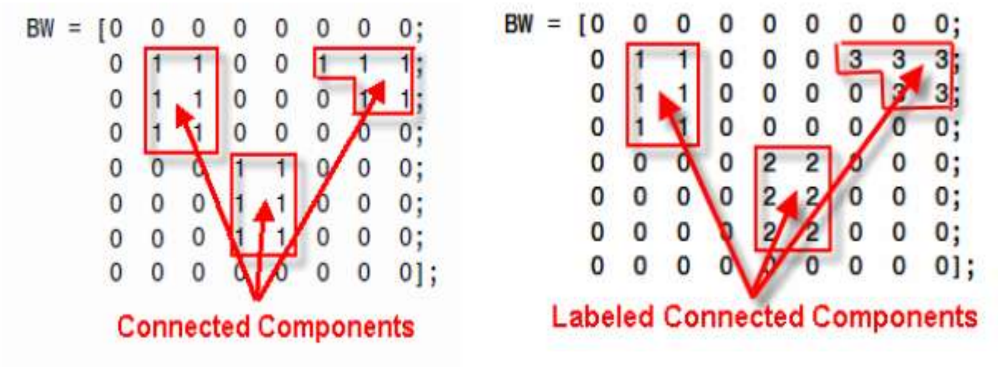
Edges are basic image features. It is carrying useful information about boundaries of the object. The edge detection is used for image analysis and object identification as well. The edge detection methods are used to find complex object boundaries when changes in brightness occur by marking potential edge points corresponding to places in an image, these edge points can be merged to form lines and object outlines. The Sobel operator is one of the best "simple" edge detection methods and used in this paper [4].

3.5 Morphological Operators

Morphological operators are used distinguish objects and structure of objects in the image. It simplifies a segmented image to ease the search for Region of Interest (objects of interest). This is done by filling small holes, smoothing out object outlines, eliminating small projections and with other similar techniques. The two principal morphological operations are dilation and erosion. Dilation allows objects in an image to expand, thus potentially filling in small holes and connecting disjoint objects in an image. Erosion shrinks objects in the image by etching away (eroding) their boundaries. The closing operator consists of dilation followed by erosion and can be used to fill in holes and small gaps in an image. The closing operator will be connecting small and adjacent objects in an image. The opening operator consists of an erosion followed by a dilation and can be used to remove all pixels in regions that are small to contain the structuring element [4].

3.6 Connected Component Labeling

A connected component in a binary image is a set of pixels that form a connected group. For sample, the binary image under has three Connected Components (CC). Connected component labeling is the process of identifying the CC in an image and assigning each one a unique label as shown in Figure- 1 [1].



```
BW = [0  0  0  0  0  0  0  0;
      0  1  1  0  0  1  1  1;
      0  1  1  0  0  0  1  1;
      0  1  1  0  0  0  0  0;
      0  0  0  1  1  0  0  0;
      0  0  0  1  1  0  0  0;
      0  0  0  1  1  0  0  0;
      0  0  0  0  0  0  0  0];
```

```
BW = [0  0  0  0  0  0  0  0  0;
      0  1  1  0  0  0  3  3  3;
      0  1  1  0  0  0  0  3  3;
      0  1  1  0  0  0  0  0  0;
      0  0  0  0  2  2  0  0  0;
      0  0  0  0  2  2  0  0  0;
      0  0  0  0  2  2  0  0  0;
      0  0  0  0  0  0  0  0  0];
```

Figure 1- Connected components and labeled connected components

3.7 Feature Extraction

Feature plays a very important role in the image processing. The main purpose of the feature is to extract the most important information from the image [11]. In this paper, the binary features are used. The color is an important feature that makes possible recognition of images. The color is used to show the difference between objects [12]. Usually, color images are represented as RGB (Red, Green and Blue) images.

For many applications are interested in color features want to combine information into the feature vector pertaining to the relationship between the color bands. These relationships are done by using the color transforms, the RGB color information is mapped transformed into a mathematical space that decouples the color information from the brightness information. This transformation is referred as a color transform or a color model.

Once this is done, the image information consists of two-dimensional color space and a one-dimensional brightness. Now, the two-dimensional color space contains information regarding the relative amounts of the different colors but does not contain any brightness information [4].

3.7.1 HSL Color Space

The Hue/Saturation/Lightness (HSL) color space decouples the image brightness from the color, where the hue is what we normally think of as “color” (like, green or orange), the saturation is a measure of how much white is in the color and the lightness is the brightness of the color. For transform RGB to HSL is used Equations 2, 3 and 4 respectively [4].

$$\boldsymbol{Hue} = \begin{cases} 0 & if\ max = \mathrm{min} \\ 60^0 * \dfrac{g-b}{max-min} + 360^0 & if\ max = r \\ 60^0 * \dfrac{gb-r-b}{max-min} + 120^0 & if\ max = g \\ 60^0 * \dfrac{r-g}{max-min} + 240^0 & if\ max = b \end{cases} \tag{2}$$

$$\boldsymbol{Saturation} = \begin{cases} 0 & if\ max = min \\ \dfrac{max-min}{max+min} = \dfrac{max-min}{2l} & if\ L \leq 1/2 \\ \dfrac{max-min}{2-(max+\mathrm{min})} = \dfrac{max-min}{2-2l} & if\ L > 1/2 \end{cases} \tag{3}$$

$$\boldsymbol{Lightness} = L = \frac{1}{2}(max - min) \tag{4}$$

The max and min value are, respectively, the largest and smallest of the RGB values.
Color features are computed by using Equations 5, 6 and 7 respectively [4].

- Mean can be computed by using Equation 5 as follows:

$$\boldsymbol{mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

- Variance can be computed by using Equation 6 as follows:

$$\boldsymbol{variance} = \frac{1}{N}\sum_{i=1}^{N} (x_i - mean)^2 \tag{6}$$

- Skewness can be computed by using Equation 7 as follows:

$$\textbf{skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - mean)^2\right)^{\frac{3}{2}}} \tag{7}$$

Where N mean the number of pixels within the image, xi is the pixel intensity in the channel.

3.8 Image Classification

Image classification is a subject of pattern recognition in computer vision. It analyzes the numerical properties of various image features and organizes data into classes [13, 14]. In this paper, the K-Nearest Neighbor (KNN) classifier is used to classify features based on the closest between features test and trained samples in the feature space. The KKN classifier is used distance metric, such as Euclidean distance for the purpose of classification, the Euclidean distance can be computed by using Equation 8 as follows:

$$\boldsymbol{dist}(\boldsymbol{x1}, \boldsymbol{x2}) = \sqrt{\textstyle\sum_{i=1}^{n}(x1_i - x2_i)^2} \tag{8}$$

Where X_1, X_2 are feature vectors.

The KNN is computed the Euclidean distance between the feature of a test image and training features. The class label of a training feature is assigned to a test feature, if the distance between a test feature and a training feature is a minimum among all the distances [15].

4 Proposed Auto Crop Method

The Auto Crop method is a fast operation to select the Regions of Interest (ROI). A signature, logo and stamp (or any one) images were extracted from the document image in this work depending on the Auto Crop method.

It reduces the time cost and computational complexities associated with extraction ROI and keeping the content information of the ROI. This method, the signature, logo and stamp are detected and extracted from original document image for father process. It assumes that any part of document image except signature, logo and stamp parts of the document image as noise. The Auto Crop method includes different steps for extraction the ROI as shown in Figure- 2

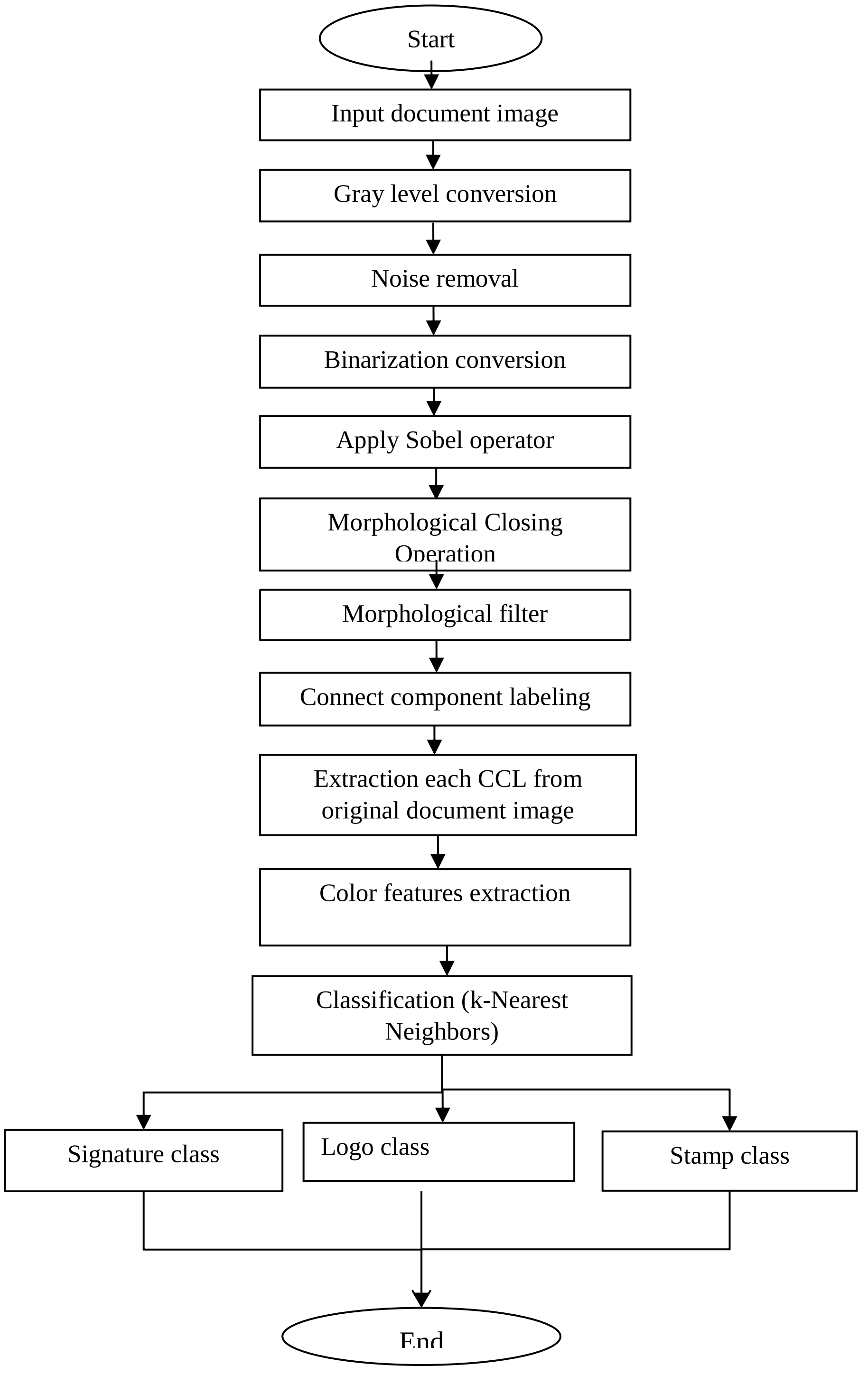

Start
Input document image
Gray level conversion
Noise removal
Binarization conversion
Apply Sobel operator
Morphological Closing Operation
Morphological filter
Connect component labeling
Extraction each CCL from original document image
Color features extraction
Classification (k-Nearest Neighbors)
Signature class
Logo class
Stamp class
End

Figure 2- Flowchart of auto crop method

5 Dataset

The dataset is a significant part to test the proposed method. The documents are proposed (created and collected) through the work for the performance evaluation of the proposed method. The document dataset contains as shown below:

- The documents that contain signature and logo.
- The documents that consist of stamp and logo.
- The documents that have stamp and signature.

- The documents that include signature, logo and stamp.

The total number of document images that are used in this paper is 200 color document images.

6 Preprocessing of the Proposed Method

Preprocessing stage contains several sub-processes that make the input document image as uniform as possible to facilitate the Auto Crop process. The preprocessing steps include the following processes.

6.1Gray Level Conversion and Noise Removal

The document image is acquired using a scanner device. The scanned document image is a color image. It is converted into a gray level then noise removal by media filter to eliminate the noise introduced during the acquisition process. The noise attached to a scanned document image has to be removed to avoid errors in the further processing steps.

6.2 Document Image Binarization

The document images are converted to binary images by using Otsu threshold. The Otsu threshold calculates the optimum threshold separating foreground and background.

6.3 Edge Detection

Edges are carrying useful information about boundaries of the object. The Sobel operator is used to find complex object boundaries when a change in brightness in the document, this step is necessary to different between objects in the document image.

6.4 Morphological Closing Operator

Morphological closing operators are used distinguish objects in the document image and segmentation image to facilitate the search for ROI. The closing operator consists of dilation followed by erosion. It can be used to fill in holes and small gaps in an image as well as connecting small and adjacent objects in an image.

6.5 Morphological filter

The document image contains handwritten text, printed text, lines, signature, logo and stamp etc. A signature, logo and stamp are often having the greater area as well as higher height and width then another element in the document image. Therefore, a morphological filter is used to removal any element has lower height, width and area. The morphological filter consists of serial steps which are erosion, opening, dilation and closing.

6.6 A ROI Extraction

The resulting of the Morphological filter is document image contains signature, logo and stamp. Then, Label each segment of the document image through Connected Component Labeling (CCL). The coordinate points of the CCL (signature, logo and stamp) in the binary image are used to extract signature, logo and stamp from the original document image.

7 Feature Extraction

Feature extraction is a significant step for image classification. The color features are extracted from each cropped images (signature, logo and stamp). To extract the features of cropped images we are using HSL color space. The color has considered as one of the based. The image (RGB color information) is transformed into HSL color model to decouple the color information from the brightness information. The HSL color space has three channels. The H and S channels will be taken because contains the color information, the Lightness information (L) is ignored because of it sensitive to less difference in lighting conditions of the input image and this is causing create color features differently. Each H and S channels are divided into four partitions. Then, each partition computes Mean, Color Variance and Color Skewness features.

8 Classification

In this paper, the K-Nearest Neighbour Classifier (KNN) classifier is used for classification process. The KNN is determined based on the Euclidean distance. The number of classes is three (signature class, logo class and stamp class). In this classifier, the features of a cropped image are compared against the features of all referenced images. The Euclidean distance is used for the computation of distances between the features of a cropped image and all reference images. The class label of a reference image is assigned to cropped image, if the distance between a cropped image and a reference image is minimum among all the distances. Thus a cropped image is classified as a member of any one of the three classes. The same procedure is followed in classifying all the remaining cropped images. Algorithm 1 describes a sequence of steps in this proposed method.

Algorithm 1: The Auto Crop Proposed Method

```
Input: Image size (N*M).
Output: Signature, logo and Stamp extracted parts of image.
Begin
 Step1:  Read a document image.
 Step2: Convert a color document image into a gray level image.
 Step3: Convert into a binary image by Otsu threshold method.
 Step4:  Apply Edge Detection by using Sobel operator.
 Step5:  Apply Morphological closing operation.
 Step6:  Apply Morphological filter to noises removal.
 Step7: Each connected components is labeled.
 Step8: The coordinate points of the bounding box in the binarized image are used to
        extract CCL from an original document image.
 Step9: The color features are extracted from each cropped images.
 Step10: The KNN classifier is applied to determine signature, logo and stamp.
End
```

9 Proposed Method Results

The evaluation of the proposed method is considered as an important part that measures method performance. The effectiveness of the proposed method is measured by using Accuracy to offer a clear overall indicator of proposed method performance. This method is applied on document image for extracted signature, logo and stamp from original document image with less time processing as shown in Table- 2. Figure- 3 shown segmentation signature, logo and stamp from original document image. Table-1 shows the evaluation measurement of proposed method. Table- 3 describe comparing proposed system with literature survey.

Table 1- The processing time of proposed method

Processing	Processing Time (Sec)
Preprocessing	3.4321
Feature extraction	0.4532
Classification	2.7832
Total Processing Time	6.6684

Table 2- Results of the proposed method

Dataset	Identification
Signature and logo	100%
Stamp and logo	98.65%
Stamp and signature	98%
Signature, logo and stamp	97.76%
Average	98.602%

From Table -2, the database contains stamp and logo was obtained 98.65% because the stamp not clear, therefore, it is consider noise. While the dataset contains signature, logo and stamp was obtained 97.78% because overlapping between signature and stamp.

Table 3- Comparison proposed method with related works

B.V.Dhandra et al	91.057%
Patil, U. et al.	98.57%
Zagoris, K.et al.	98.86%
Proposed method	98.602%

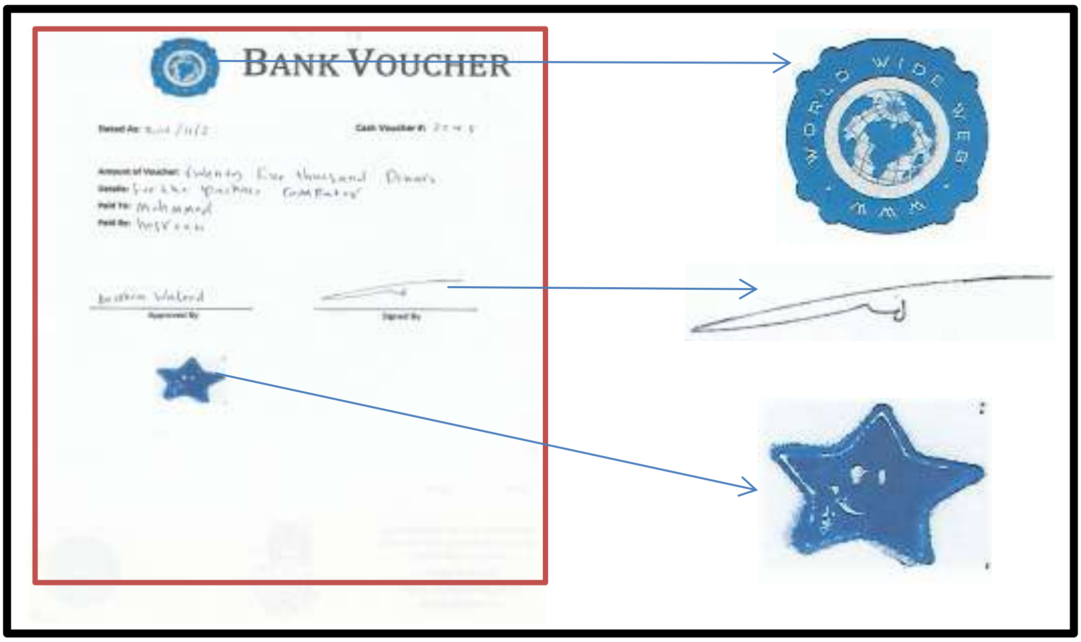


Figure 3- Segmentation signature, logo and stamp from original document

10 Conclusions

In This paper, the proposed method for detection and extraction signature, logo and stamp from original document image with less processing time and keeping the content information of the signature object without losing of any pixel of image. This method is worked well on document images that not are containing overlapping between signature and stamp as well as not clear stamp. The cropped images from the proposed method are used for the biometric system. Accuracy of 98.602% was achieved from Auto Crop method.

References

1. Dhandra, B.V. Shridevi, S. and Rashmi, T. **2013**. Classification of Document Image Components. *International Journal of Engineering Research and Technology*, **2**(10): 1429-1439.
2. Shazia, A., Mehraj-Ud-Din, Dar and Aasia, Q. **2010**. Image Processing- A Review. *International Journal of Computer Applications*, **10**: 35–40.
3. Raval, A., Jalwani, A. and Karathiya, M. **2012**. Document Image Analysis. *International Journal of Advanced Research in Computer Science and Software Engineering* (*IJARCSE*), **2**(5): 346-349.
4. Umbaugh, S.E. **2010**. *Digital image processing and analysis: human and computer vision applications with CVIPtools.* Second Edition. CRC Press, USA.
5. Gautam, C.M., Sharma, S. and Verma, J.S. **2012**. A GUI for Automatic Extraction of Signature from Image Document. *International Journal of Computer Applications*, **54**(15): 13-19.
6. Patil, U. and Begum, M.**2012**. Word level handwritten and printed text separation based on shape features. *International Journal of Emerging Technology and Advanced Engineering*, **2**(4): 590-594.

7. Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B. and Papamarkos, N. **2012**. Handwritten and machine printed text separation in document images using the bag of visual words paradigm. International Conference on Frontiers in Handwriting Recognition (*ICFHR*), pp: 103-108.

8. Saha, S., Basu, S., Nasipuri, M. and Basu, D.K. **2010**. A Hough transform based technique for text segmentation. *Journal of Computing*, **2**: 134- 141.

9. Güler, İ. and Meghdadi, M. **2008**. A different approach to off-line handwritten signature verification using the optimal dynamic time warping algorithm. *Digital Signal Processing*, **18**(6): 940-950.

10. Otsu, N.**1979**. A Threshold Selection Method from Gray-Level Histograms. *IEEE TRANSACTIONS ON SYSTREMS, MAN, AND CYBERNETICS,* **9**(1): 62-66.

11. Kumar, G. and Bhatia, P.K. **2014**. A detailed review of feature extraction in image processing systems. *In Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on*, pp: 5-12.

12. Tripathi, G. **2014**. Review on color and texture feature extraction techniques. *International Journal of Enhanced Research in Management and Computer Applications*, **3**(5): 77-81.

13. Umamaheswari, J. and Radhamani, G. **2012**. An amalgam approach for DICOM image classification and recognition. *World Acad Sci Eng Technol*, **6**: 807-812.

14. Nandita Chasta and Manish Tiwari**. 2016**. Optimized Image Classification based on Universal Image Distance and Support Vector Machines. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE,)* Volume5.

15. Han, J., Pei, J. and Kamber, M. **2011**. *Data mining: concepts and techniques*. Third Edition, Elsevier.