# Extraction of Doodles and Drawings from Manuscripts

Chandranath Adak[1] and Bidyut B. Chaudhuri[2]

[1] Dept. of CSE, University of Kalyani, West Bengal-741235, India
adak32@gmail.com
[2] CVPR Unit, Indian Statistical Institute, Kolkata-700108
bbc@isical.ac.in

**Abstract.** In this paper we propose an approach to separate the non-texts from texts of a manuscript. The non-texts are mainly in the form of doodles and drawings of some exceptional thinkers and writers. These have enormous historical values due to study on those writers' subconscious as well as productive mind. We also propose a computational approach to recover the struck-out texts to reduce human effort. The proposed technique has a preprocessing stage, which removes noise using median filter and segments object region using fuzzy c-means clustering. Now connected component analysis finds the major portions of non-texts, and window examination eliminates the partially attached texts. The struck-out texts are extracted by eliminating straight lines, measuring degree of continuity, using some morphological operations.

**Keywords:** Connected Component, Document Image Analysis, Doodle Separation, Fuzzy C-Means Clustering, Manuscript Processing.

## 1 Introduction

In the field of *document image analysis* [1], the separation of texts and non-texts has gained interest since 1980. It is important, so that they can be sent to different systems/engines for processing. The texts are fed to OCR and non-texts are sent to graphics processing system. The texts may be printed, handwritten and mixture of both (hybrid). The existing methods [2-10] deal with different logos, diagrams, maps, engineering drawings and photographic images.

Here we separate the doodles and drawings from ancient manuscripts. For our experiment, we choose manuscripts of Leonardo da Vinci, Gustave Flaubert, Lewis Carroll, Rabindranath Tagore and Samuel Beckett. In the manuscript of Vinci, there are some engineering models and human figures. Some crossed lines and rectangular boxes are present in Flaubert's manuscript. Carroll's manuscript contains human figures and Beckett's manuscript is full of funny characters. In Tagore's manuscript, we find doodles of various shapes (e.g. real and imaginary animals, trees, human models, phantoms etc.).

The challenge of our work is that most of the doodles and drawings are touching the texts, and sometimes the doodles are formed with struck-out lines

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 515–520, 2013.

(ink-strokes) of irregular patterns. We approach to recover the texts behind those struck-out lines.

In our preprocessing stage, we use median filter to remove noise and fuzzy c-means clustering (FCM) to segment the image into object regions. We apply connected component (CC) analysis to separate non-touching doodles, examine windows to detach touching non-texts (doodles) and extract struck-out texts using straight line elimination, degree of continuity measurement and mathematical morphological operations.

## 2 Proposed Method

The proposed method consists of following steps:

1. The manuscript is scanned in RGB color ($I_{rgb}$, fig.1.a) and converted into gray-scale image $I_{gray}$ (fig.1.b).
2. Noise is removed from $I_{gray}$ using 3-by-3 neighborhood *median filtering* to get the image $I_{nf}$ (fig.1.c).
3. $I_{nf}$ is segmented (fig.1.d) by *FCM clustering* to obtain the ink-strokes and background pixels separated. The background pixels are converted to zero value (black), while the foreground ink-strokes are converted into one (white). Let the resulting image is $I_{bin}$.
   Steps 1-3 are basically preprocessing stage.
4. We generate all the connected components of white pixels from $I_{bin}$. The doodles are usually dense, well-connected, large sized components. We extract the larger connected components (doodles) by a threshold $T$ (fig.1.e). Now non-touching doodles (those do not touch any text) and texts are separated (fig.1.f-g).
5. For text touching with doodles, we note the basic features of text that it is elongated with curvy, thin, smaller lines with variation of degree of continuity. On $I_{bin}$ we take a 5×5 window, the pixel values of this window is zero/one due to binarization. If all pixels of a row/column/diagonal are one, then it is more likely that the window is part of doodle (fig.2.b-c).
   Steps 4,5 are repeated interactively for a satisfactory outcome.
6. Next we identify and delete the struck-out lines, which is done as follows:
   (a) Morphological *thin* and *shrink* [11] operations are used on the remaining image.
   (b) We eliminate the horizontal, vertical, diagonal straight lines by examining 5×5 window and checking *degree of continuity.*
   (c) After eliminating those straight lines, small pieces of lines may remain. We compute CC on the remaining image. If the number of pixels in a CC is less than a threshold, then that CC is converted into background.
   (d) In the remaining image, there is the skeleton of text that was struck-out. We find the edge of this skeleton using *Sobel* operator. We create dilated image $I_d$ by morphological *dilate* [11] operation. On $I_d$ we return the gray value of the original image $I_{gray}$ (fig.4.b,e).

   Steps 4,5,6 are repeated for a suitable separation of texts and non-texts.
7. We combine these all non-touching, touching and struck-out texts to obtain the total text portion. The remaining portions are considered as non-texts.

## 3 Experimental Result and Discussion

To assess the stability and correctness of the proposed method, the results are obtained from different ancient manuscripts.

In *fig.1*, we handle the trivial non-touching case, where the non-texts (doodles) are not connected with the texts.

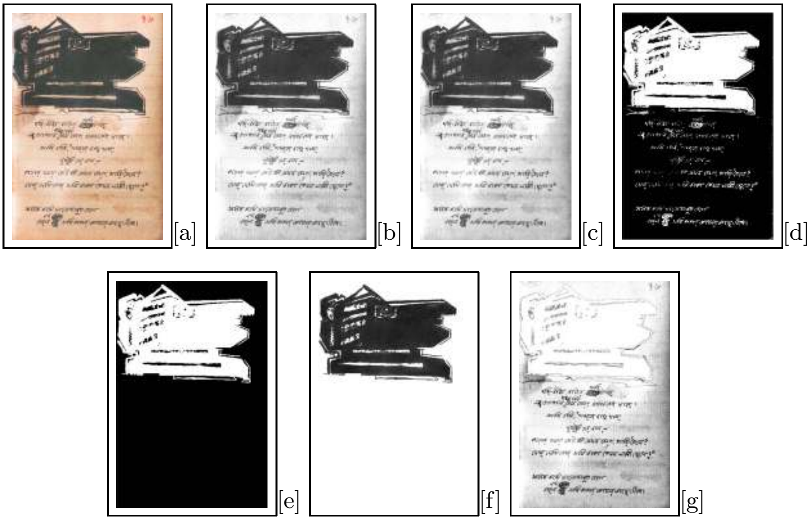

**Fig. 1.** *Non-touching case:* Manuscript of Tagore's poem ("Nutan Srota"), *Parisesh*, 19 August, 1927: (a) original, (b) gray, (c) noise free, (d) segmented (binary), (e) CC (large), (f) gray value of CC (doodle), (g) texts in gray

In *fig.2*, we separate the touching texts and non-texts by the window examination. *Fig.2.c* shows the doodle in binary after eliminating the major portion of touching text.

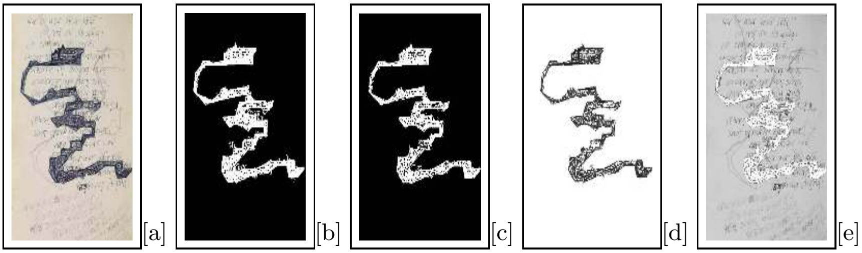

**Fig. 2.** *Touching case:* Manuscript of Tagore's song, *Geetabitan*, 1929: (a) original, (b) larger CC: before window examination, (c) CC: after window examination, (d) CC in gray (doodle), (e) texts in gray

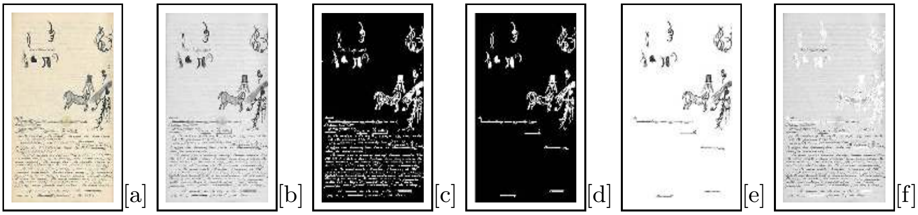

**Fig. 3.** Beckett's manuscript, *Watt I*, 1941: (a) original, (b) noise free gray, (c) segmented (binary), (d) CC (large), (e) CC in gray (doodle), (f) texts in gray

Our proposed method also works for small doodles on Beckett's manuscript. It is shown in *fig.3*. Some bold struck-out ink-strokes are marked as non-texts.

We approach to extract the struck-out texts on Tagore's manuscript shown in *fig.4.a,c*. Total recovery is not possible by our proposed method. After extracting the struck-out text, we combine this with the remaining texts.

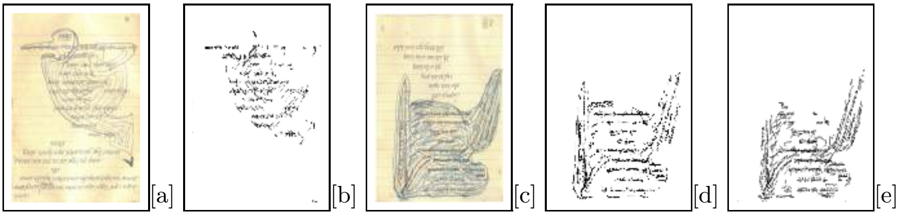

**Fig. 4.** *Struck-out case:* Tagore's manuscript: (a) original (*Tasher Desh*, 1933), (b) struck-out texts in gray scale, (c) original (*Nabajatok*, 4 May 1939), (d) struck-out texts in binary, (e) struck-out texts (with changing threshold) in gray

In *fig.5*, we deal with Vinci's manuscript and separate texts and drawings.

*Fig.6* shows complex (where human cannot read the struck-out texts properly without a great effort) case analysis of our proposed method.

We take total 115 manuscripts of different sizes, out of which there are 24 non-touching, 63 touching, 23 struck-out and 5 complex cases.

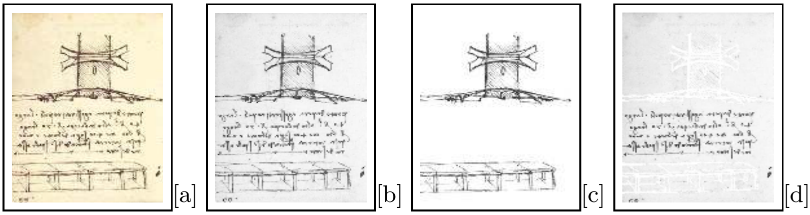

**Fig. 5.** Vinci's manuscript, *Golden Horn Bridge design*, 1502: (a) original, (b) noise free gray, (c) drawings in gray, (d) texts in gray

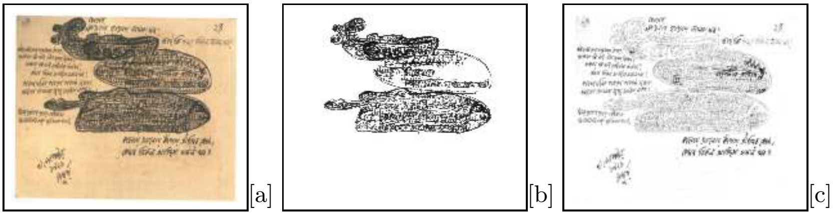

**Fig. 6.** *Complex case:* Manuscript of Tagore's song ("Chokher chaoar haoay dolay mon"), *Geetabitan*, 9 Sep. 1926: (a) original, (b) doodle in gray, (c) texts in gray

For performance analysis of our proposed method, we calculate precision (P), recall (R) and F-measure (F) considering their standard definition.

In table 1, we show the average P, R and F for different cases. To calculate the average, we use the arithmetic mean of P, R and F for all images under each case.

The *true positive (TP), true negative (TN), false positive (FP)* and *false negative (FN)* are defined as follows:
TP: # pixel in the actual doodle portion in an image (correctly classified),
TN: # pixel in the actual text portion (correctly classified),
FP: # pixel in text incorrectly labeled as doodle (unexpected result),
FN: # pixel in doodle incorrectly marked as text (missing result).

**Table 1.** Performance analysis of doodles and drawings extraction

| Case Study | Non-touching | Touching | Struck-out | Complex |
|---|---|---|---|---|
| P % | 99.80 | 98.71 | 60.31 | 20.78 |
| R % | 99.28 | 99.15 | 59.52 | 84.92 |
| F % | 99.54 | 98.93 | 59.91 | 33.39 |

The non-touching and touching cases show 99.54% and 98.93% F-measure respectively, the struck-out case shows 59.91% F-measure, but for complex case it is only 33.39% , we have found less number of struck-out and complex cases, mostly they are touching.

To the best of our knowledge, a preliminary work on such separation was done by Chaudhuri et al. [12] on simple doodles of Tagore's manuscript and they obtained 92.17% accuracy (90.14% F-measure). We did not get any other reference/work to do further comparative study.

## 4 Conclusion

A text and non-text (doodle) separator from ancient manuscript has been presented in this paper. The proposed algorithm works well for handwritten/printed texts and it is independent of scripts. This algorithm can be extended to some other applications in document image analysis, such as image restoration from ancient documents, wills and testaments, newspapers, magazines, articles etc. Though it works for touching and non-touching cases, it does not work well for complex cases, and more modification is required to explore struck-out texts. Our system is semi-automatic, so our next venture will be to make the system automatic and more accurate.

## References

1. Nagy, G.: Twenty Years of Document Image Analysis in PAMI. IEEE Trans. on PAMI 22(1), 38–62 (2000)
2. Luo, H., Agam, G., Dinstein, I.: Directional Mathematical Morphology Approach for Line Thinning and Extraction of Character Strings from Maps and Line Drawings. In: Proc. ICDAR 1995, pp. 257–260 (1995)
3. Kasturi, R., Bow, S.T., El-Masri, W., Shah, J., Gattiker, J.R., Mokate, U.B.: A System for Interpretation of Line Drawings. IEEE Trans. on PAMI 12(10), 978–992 (1990)
4. Dori, D., Liu, W.: Vector-Based Segmentation of Text Connected to Graphics in Engineering Drawings. In: Perner, P., Rosenfeld, A., Wang, P. (eds.) SSPR 1996. LNCS, vol. 1121, pp. 322–331. Springer, Heidelberg (1996)
5. Lu, Z.: Detection of Text Regions from Digital Engineering Drawings. IEEE Transactions on PAMI 20(4), 431–439 (1998)
6. Fletcher, L.A., Kasturi, R.: A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. IEEE Tran. on PAMI 10(6), 910–918 (1988)
7. He, S., Abe, N.: A Clustering-Based Approach to the Separation of Text Strings from Mixed Text/Graphics Documents. In: IEEE Proc. of ICPR 1996, pp. 706–710 (1996)
8. Adak, C.: Unsupervised Text Extraction from G-Maps. In: Proc. Int. Conf. on Human Computer Interactions (ICHCI 2013), India (August 2013)
9. Roy, P.P., Lladós, J., Pal, U.: Text/Graphics Separation in Color Maps. In: Proc. Int. Conf. on Computing: Theory and Applications (ICCTA 2007) (2007)
10. Garg, R., Hassan, E., Chaudhury, S., Gopal, M.: A CRF Based Scheme for Overlapping Multi-Colored Text Graphics Separation. In: Proc. ICDAR 2011, pp. 1215–1219 (2011)
11. MATLAB R2012a (7.14.0.739), MathWorks Inc., http://www.mathworks.com
12. Chaudhuri, B.B., Borah, S., Saraf, A., Goyal, A., Kumari, A.: Separation of Text from Non-Text Doodles of Poet Rabindranath Tagore's Manuscripts. In: Proc. Nat. Conf. on Comp. and Comm. Systems (NCCCS 2012), pp. 1–5 (November 2012)