DETECTING HIGHLIGHTS AND NOTES ON PRINTED TEXT

Roee Sulimarski, Gal Gur-Arye, Avishai Adler

Technion - Israel Inst. Technology
Department of Electrical Engineering
Haifa 32000, Israel

Yaakov Navon

IBM Research Laboratory
Document Processing and Management Group
Haifa 31905, Israel

ABSTRACT

The extensive use of digital cameras in creating an electronic database of document images, leads to a major problem in managing the vast amount of information. Many images include user-added marks, such as highlights and handwritten notes, which can be used in Content Based Image Retrieval. This paper presents a novel method for determining the existence of user-added marks in a given image and classification of existing marks. Two different approaches are integrated for the detection of marks: one based on color-spaces and the other on anomaly detection in text-dominant images. The proposed method achieved very successful results.

Index Terms— document image processing, image segmentation, content-based retrieval , image retrieval

1. INTRODUCTION

There are many advantages to converting and storing documents in digital form, such as establishing a paperless office. Many methods of Document Image Analysis (DIA) convert document images to OCR-ed text, graphics and metadata for storage and retrieval [3], using layout decomposition [4]-[6] and other methods [7].

In recent years, creating large document image databases has become a feasible possibility. It is very simple to create such databases by using digital cameras and scanners. Storing documents in image form is beneficial, maintaining a high quality representation of the original hard copy, especially graphics and images. In order to use the information gathered, we need to organize the databases to enable efficient searching and retrieval. Yet, manually categorizing and properly indexing the document images is a tiring and time-consuming task.

In the past years, Content Based Image Retrieval (CBIR) has become popular for indexing general images [11],[12]. Images are indexed by analyzing their visual content, such as color, shape and texture. In specific applications, high-level content is used, such as human faces or fingerprints. Several CBIR systems have been built, implementing different types of queries. Query by example is a method in which the user inputs an example image for the search to be based upon. Another method is semantic retrieval: query by specific features such as shape or faces.

While reading journal papers, conference agendas or presentation handouts, readers tend to mark areas which are of interest to them, or add notes to the text. Traditional DIA methods do not utilize the presence of user-added marks. Thus, an advantage of storing documents in image form, is that it preserves marks, such as highlighted paragraphs or handwritten notes in the margins. These marks may be more meaningful to the user in retrieving the document, than trivial metadata such as the author or index terms of the document. They can be used in CBIR, enabling advanced queries for retrieval of regions of interest from documents stored in an image database.

This paper presents a method for automatic identification and classification of user-added mark in document images. Two different approaches are integrated for the detection of marks: one based on color-spaces and the other on anomaly detection in text-dominant images. Both methods take advantage of various features in document images, such as color, text density, edges, and statistical measures of connected components, to determine the physical layout of the image. Extracting text regions and graphics from the image, enables detection of the external user-added marks. Classification of the marks is done by a set of rules, based on the connected components data of each identified mark.

The method is currently demonstrated on documents with a well-defined structure, however we do not assume a priori knowledge of the layout itself, text size, etc. We assume text is the dominant feature in the image. The method is robust to skew, up to a reasonable degree.

2. INTEGRATED FRAMEWORK

The framework of our solution is divided into two stages: detection and classification. The first stage detects regions of interest in the document image, by two parallel approaches. A color-space approach is used to detect and segment color regions in the image. Assuming text regions to be dominant in the image, an anomaly detection approach is used to detect non-text regions in the image. The output of both processes is combined to create a “regions of interest” image.

This image includes graphics which are detected and then removed. Thus, the result of the first stage is an image containing anomalies which are suspected as user-added marks. The second stage of the proposed solution is classification of the anomalies in the image. This is done by a set of rules. A flowchart of the proposed framework is shown in Fig. 1.

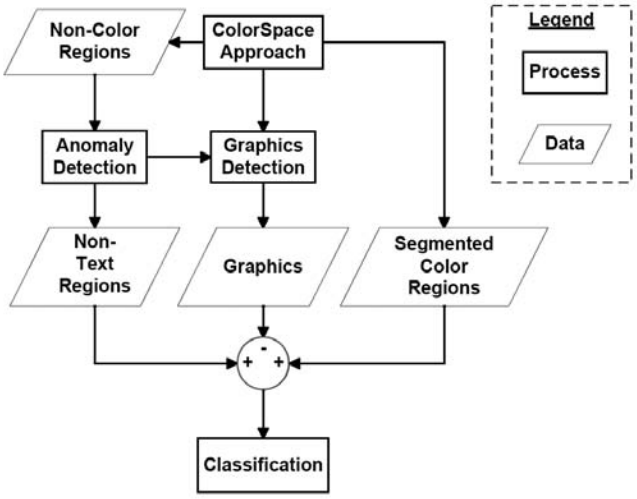


Fig. 1: Integrated Framework

The paper is organized as follows. In Sec. 3 we detail how color is used to detect highlights and colorful hand-written notes. In Sec. 4 we describe how typical text features are used to detect anomalies in document images. Sec. 5 details a method for detecting graphic regions in the image. Classification of the remaining anomalies in the image is detailed in Sec. 6. In Sec. 7 we display examples of images processed by our method.

3. COLOR-SPACE APPROACH

We assume the text and background to be black and white, respectively. Illuminance correction can be performed as a preprocessing stage to ensure this assumption. As opposed to text, the highlights are colored, as are some of the hand-written notes. This feature can be utilized in their detection. This also enables us to determine the color of the user-added marks, which can be used in advanced retrieval queries. Since graphics in document images will usually appear in color, the output of this phase will include color graphics. The graphics will be detected and extracted, as explained in Sec. 5.

We make several logical assumptions regrading the hue of different objects in the image:

1. A handwritten note is made in a single color.
2. A paragraph will usually be highlighted by up to two different colors, with possible overlapping of the colored marks.
3. Colorful graphics include a variety of hue values.

These assumptions assist in determining the possible classification of a colored region in the image.

3.1. Color Region Detection

Colored regions are detected using RGB color-space. This is done in a pixelwise manner. Grayscale pixels have similar values in all three color channels. Therefore, for each pixel, the mean of the three color channels is calculated. The pixel is considered grayscale if each color channel value is close to the mean, up to a pre-determined threshold. Subsequently, all grayscale pixels are removed from the image. Remaining colored pixels are grouped into connected components (CC), using a binary mask. Connected components, in a binary mask, are a set of pixels of the same logical value which are eight-connected.

This stage yields a segmentation of the document image into colored regions and non-colored regions. An image of the non-colored regions is used as an input to the anomaly detection approach, described in Sec. 4. An image of the color regions is used as an input to the next phase of the colorspace approach. In this phase, each of the color regions is analyzed.

3.2. Color Segmentation

Following the assumptions in Sec. 3, we wish to determine whether a certain color region has one or more hue values. This is done in two stages, using HSV color-space.

First, we determine whether the hue value of a color region can be classified as a highlight. For this purpose, a lookup table was created, based on hue and saturation statistics of highlight markers, from images collected under various lighting conditions. This lookup table can be updated for the user by providing an image of typical highlights used, acquired by either camera or scanner. Thus, there is no need to calibrate the equipment one uses. This stage can recognize if a color region contains only one hue value.

Secondly, we determine the number of colors in a given color region by segmentation. Color regions which do not correspond to typical highlight colors are segmented using K-Means [13], based on hue values. The K used is the number of colors expected to be found in a color region. Using $K = 2$, for example, means that the color region will be segmented into two different colors. We perform K-means segmentation for each color region separately, using $K \in \{1, 2, 3, 4\}$.

To determine K_{opt}, the optimal K for segmentation, we use the Mumford-Shah functional [8] for piecewise constant approximation. Define R_i as the pixels belonging to i-th segment of a given color region, and S_{R_i} as the area of this segment. Defining $h(x, y)$ as the hue value of a pixel at position (x, y), the mean of the hue values in a segment is calculated using

$$a_i = \frac{1}{S_{R_i}} \int \int_{R_i} h(x, y). \quad (1)$$

The length of the segments' contour is defined as Γ. The area of the entire document image is defined as S_I. The Mumford-Shah functional for piece-wise constant approximation is then

$$\sum_{i=1}^{K} \int\int_{R_i} (h(x,y) - a_i)^2 dxdy + \alpha|\Gamma|. \quad (2)$$

The parameter α is used as a tradeoff between the two terms of the functional. Using $\alpha = \sum_i S_{R_i}/S_I$ yields satisfying segmentation results.

The functional is calculated for $K \in \{1,2,3,4\}$, and K_{opt} is chosen as the one yielding the minimal value of the functional. This determines the number of hue values in a given color region.

If $K_{\text{opt}} = 4$ for a given color region, this region is very colorful and is therefore classified as graphics. This follows directly from our assumptions. A highlighted paragraph will be segmented into three colors at most. Therefore, all regions with more than three colors are graphics.

4. ANOMALY DETECTION APPROACH

Text is the main feature of document images. Document images usually have a structured layout into paragraphs, sentences, headlines, etc. Determining the layout of text in the image, will enable to detect the non-text regions, which are treated as anomalies. Aside from graphics, these regions will include handwritten notes. Many methods exist for text detection in images, especially in DIA literature. Our approach is based on simple features: edge density and texture, described using a local STD measure. The input to this stage is an image of non-colored regions, created in Sec. 3.1.

4.1. Text Detection

Following the basic assumption that text is the dominant component of a document image, along with the fact that text regions have a structured layout, drives our main approach to text detection and extraction. We wish to find typical features to distinguish text regions from the rest of the image. We use a two-phase approach for text detection. In the first phase the Canny edge detection algorithm [1] is used in order to create an edge image. We assume text in the image to have a high density of edges [9]. Therefore, local edge density is calculated in a coarse representation of the edge image and typical statistics are then used to mark "suspicious areas" which contains high edge density values. These suspicious areas are divided into CCs, as smooth filled regions. Hand-written notes may also be contained in suspicious areas, especially if these notes are close to text paragraphs. Text regions are usually extracted as entire paragraphs or columns, whereas handwritten notes will usually be extracted as small areas, or will be contained within a text region.

Text has a periodic texture, which can be quantified using a local STD measure. In the second phase, a local STD image is formed for every suspicious area. Defining $I[m,n]$ as the intensity value of a pixel and W as the width of a square window, the local STD for each pixel $\sigma[m,n]$ is calculated using:

$$\mu[m,n] = \frac{1}{W^2} \sum_{i,j=\frac{-W}{2}}^{\frac{W}{2}} I[m+i, n+j] \quad (3)$$

$$\sigma^2[m,n] = \frac{1}{W^2} \sum_{i,j=\frac{-W}{2}}^{\frac{W}{2}} I^2[m+i, n+j] - \mu^2[m,n]. \quad (4)$$

Each suspicious area is assigned two values: μ_i^{STD} and σ_i^{STD}, which are the mean and the STD of the local STD values of each area, respectively. Define S_i as the area of the i-th suspicious region. A weighted average is used to calculate typical values for local STD of text regions, defined as μ_{text} and σ_{text}:

$$\mu_{text} = \sum_i \frac{S_i}{\sum_i S_i} \mu_i^{\text{STD}} \quad ; \quad \sigma_{text} = \sum_i \frac{S_i}{\sum_i S_i} \sigma_i^{\text{STD}} \quad (5)$$

Following our assumptions that text is a dominant textured feature in the image, we have found typical values for determining text regions. A local STD image of the entire document image is calculated, and the values μ_{text} and σ_{text} are used to extract all text regions. All pixels for which

$$\mu_{text} \leq \sigma[m,n] \leq \mu_{text} + \sigma_{text} \quad (6)$$

are classified as text and removed from the image. Handwritten notes may be damaged during this process if they correspond to this value, yet they will be reconstructed in the following stage.

4.2. Handwriting Reconstruction

The result of the previous stage requires solving two issues. The first is removal of "leftover" text, which appears as noise or speckles in the image. The second is reconstruction of handwritten notes [10], which was damaged in the Text Extraction stage. The difference between these two objects is that remaining text speckles have low pixel density and notes have high pixel density. Using an averaging window, areas of high density are detected and restored using morphological reconstruction [2].

5. PRINTED GRAPHICS DETECTION

The output images of the color-space stage and the anomaly detection stage are different, as the former is performed on colored regions and the latter is performed on non-colored regions. Graphics detection is performed separately on each output image, resulting in a binary mask corresponding to detected printed graphics. Both masks are merged in order to ensure that all printed graphics in the image have been detected.

Graphics usually have varying texture, as opposed to highlights, which have a uniform intensity. Also, graphics in document images are usually contained within large rectangular frames. Both texture and size are used to detect graphic regions in the image. In addition, we wish to use a measure based on shape. To ensure robustness to skew, a measure of solidity is used. Solidity is defined as the ratio between the number of pixels in a given region and the number of pixels in the minimal convex hull enclosing the region.

Detection of textured areas in the image is performed using a local STD measure, in the value channel of the HSV image. A local STD image is created, using Eq. 3-4. An empirical threshold was found for determining high local STD values.

The textured region is then compared to a binary mask of the original document image. A textured region is classified as graphics, if it corresponds to a CC in the mask which matches the morphological assumptions about graphics.

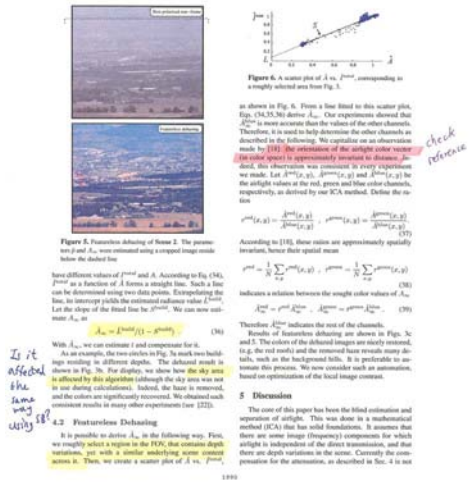
6. CLASSIFICATION

The remaining objects in the document image are user-added marks, which we wish to classify as either highlights or handwritten notes. A binary mask of the image is created and divided into CCs. Classification is performed using a set of rules based on characteristic CC values, such as height, width, effective area, etc. We assume highlights to be higher than the height of a text letter and to be large in size, compared to an individual handwritten letter. Handwriting does not have any uniform features, aside from size. Each letter is usually treated separately.

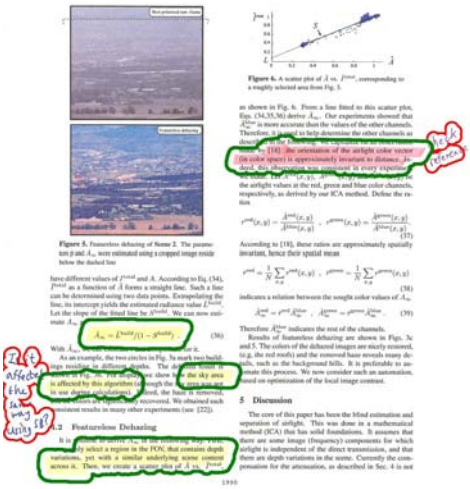
7. RESULTS

The method described in this paper was used successfully on dozens of document images. In our experiments, we used various images of documents, including both colored and grayscale graphics. These images were marked using highlights of different colors, and notes were written in the margins, both in color and in black ink. Images were scanned to dimensions of roughly 1700x2400 pixels. In addition to determining whether the document contains user-added marks, these marks are classified and outlined in the document itself. Thus, we have information on the existence of marks, and their color and location in the document. These identified marks can be separated from the image and presented to the user as regions of interest.

Examples of our results are shown in Figs. 2-3. In our implementation of the algorithm, user added-marks that were detected and classified, are automatically outlined in the output image. Detected highlights are outlined in green and detected handwritten notes are outlined in red. In Fig. 2a, the document page contains several highlighted areas. Square graphics objects appear at the top left of the page and a colored graph figure appears on the top right. Handwritten notes were added in the left and right margins of the text. Text regions were all detected, including the equations, which are irregular in the page. In Fig. 2b, we can see that the highlights have been fully detected. The comments were completely outlined, despite their proximity the text. In Fig. 3, we present only the output image. The original document page contains many graphics objects. These graphics are a mixture of color and grayscale. The text regions take up less than half a page. Also note that the page itself is rotated. There are two highlighted area, one containing two overlapping colors. During the processing of the image, all text regions were detected, including both isolated captions accompanying the figures and equations, which are irregular in the document. We can see that the highlights have been fully detected. The comments were completely outlined, despite their proximity



(a) Original Document Image



(b) Output Image: Detected highlights are outlined in green, Detected handwritten notes are outlined in red

Fig. 2: Example of Result

to both text and graphics. The rotation angle did not affect the detection and classification of the marks in the image.

On average, a $\mathbf{MATLAB}^{\mathrm{TM}}$ implementation of the algorithm takes 40 seconds to process a 1700x2400 image, using a 2.4 GHz Multi-Core PC. This implementation is not optimized.

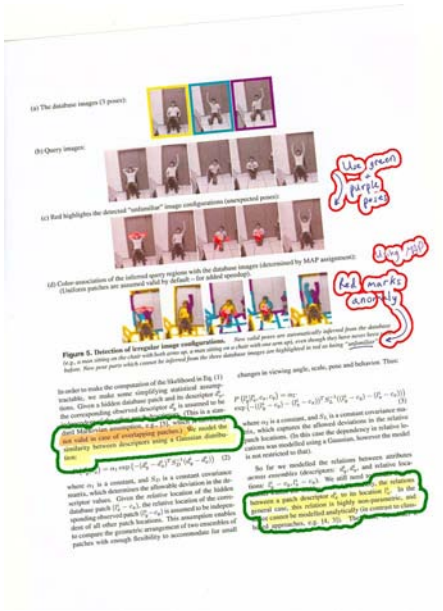


Fig. 3: Example of Result: Output Image.
Detected highlights are outlined in green, Detected handwritten notes are outlined in red

8. SUMMARY AND CONCLUSIONS

A framework for document image retrieval based on user-added marks has been presented. The detection stages utilizes various features in document images, such as color, texture and edges, to locate anomalies in the image. Classification is based on simple, logical rules. The algorithm is robust to skew. A preprocessing stage to the proposed solution should include illuminance correction and typical text size detection for scaling. The algorithm was tested on dozens of document images, achieving impressive results.

This algorithm can be implemented in CBIR systems for retrieval of document images. Queries can be by example, providing an image of the desired highlight color. Another option is semantically specifying the color of the desired mark or its location in the image.

9. ACKNOWLEDGMENTS

This work was performed in the Signal and Image Processing laboratory, Dept. of EE, Technion IIT. The authors are grateful to Avi Rosen for his help which allowed us to carry out this project.

10. REFERENCES

[1] Canny, J., "A Computational Approach To Edge Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 679-714, 1986.

[2] Vincent, L., "Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms," *IEEE Transactions on Image Processing*, Vol. 2, No. 2, pp. 176-201 April, 1993.

[3] G. Nagy, "Twenty Years of Document Image Analysis in PAMI," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 38-62, Jan. 2000.

[4] J.L. Fisher, S.C. Hinds, and D.P. D'Amato, "A Rule-Based System for Document Image Segmentation," *Proc. 10th Int'l Conf. Pattern Recognition*, pp. 567-572, June 1990.

[5] K.Y. Wong, R.G. Casey, and F.M. Wahl, "Document Analysis System," *IBM Journal Res. Dev.*, vol. 26, no. 6, pp. 647-656, 1982.

[6] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1,162-1,173, 1993.

[7] I. A . Fletcher and R. Kasturi, "Robust Algorithm for Text String Separation from Mixcd Text/Graphics Images," *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 10, no. 6, pp. 910-918, June 1988.

[8] D. Mumford and J. Shah. "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems," *Comm. Pure Appl. Math.*, 42, pp 577-684, 1989.

[9] P. Clark, M. Mirmehdi, "Finding Text Regions Using Localised Measures," *Proc. 11th BMVC*, pp. 675-684, 2000.

[10] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, New York, 1993.

[11] Y. Rui, T. S. Huang and S. Chang, "Image Retrieval: Past, Present, and Future," *Journal of Visual Communication and Image Representation*, pp. 1-23, 1997

[12] D. Doermann, "The Indexing and Retrieval of Document Images," *Computer Vision and Image Understanding*, vol. 70, pp. 287-298, 1998.

[13] G. A. F. Seber, *Multivariate Observation*, Wiley, New York, 1984.