Content analysis in document images: A Scale Space Approach

Y. FATAICHA[1, 2], M. CHERIET[1, 3], J. Y. NIE[2], and C. Y. SUEN[3]

1. LIVIA Laboratory, École de Technologie Supérieure de Montréal, Québec, Canada
2. INCOGNITO laboratory, Université de Montréal, Québec , Canada
3. CENPARMI, Concordia University , Montréal, Québec , Canada

fyoussef@livia.etsmtl.ca cheriet@gpa.etsmtl.ca nie@iro.umontreal.ca suen@cenparmi.concordia.ca

Abstract

With the growing interest in automatic transformation of paper document to its electronic version, geometrical and logical structures have become an active research area for a decade. Nowadays, kernel scale space has been widely adopted as the most promising multi-scale image document analysis method. Yet still, traditional methods using scale space approach has its limitations: they are useful mostly on character extraction and they carry a large computational load. In view of these limitations, this paper proposes a new approach using scale space in order to analyse the composite document content. In the proposed method, scale space transform is used to decompose an image into different scaled objects where the scale value is used for detecting progressively finer objects: text, line drawing, logo, and image, with encouraging results on real-life data.

Keywords: Image document analysis, Segmentation, Identification, Kernel Scale Space, Content image analysis.

1. Introduction

Over the past 20 years, scientists have conducted extensive research on various aspects of document processing. In practice, much knowledge has to be acquired from documents, including technical reports, government files, newspapers, books, journals, magazines, letters, and bank checks, to name a few. Scientists have designed and developed numerous document image recognition algorithms. Since 1991, many papers reporting new achievements in document processing have appeared [4,6,11]. It is not trivial to automatically transform a paper document into an electronic one. Automatic extraction and identification of information space in the image by computer can be divided into two approaches, namely, top-down and bottom-up. The objective is to describe the components of the document and to identify them as texts, logos, line drawings, and images. The top-down ones can process only simple documents, which have specific format or contain some a priori information. Both methods fail to process the documents with complicated geometric structures. Conventional identification methods [2,5] cluster adjacent rectangular blocks and simply classify them into text and non-text objects.

This paper presents a new method using scale space to analyze the composite document content. The method consists of two steps: region extraction and object identification. Based on the hybrid combination of top-down and bottom-up techniques, the scale space segments the document into several blocks at different scales without a priori information. Identification is based on the object's shape and relationship among geometric properties such as width, length, gaps, density, and entropy. Moreover, experimental results demonstrated that applying scale space on document images at different scales gives a higher accuracy and a better discriminatory power.

This paper is organized as follows. Section 2 presents related works and our motivation. In section 3, a scale space document analysis method is proposed to specify the geometry of the homogeneous regions and then identify them. Experimental results using the document database from the University of Washington are presented in section 4. Finally, section 5 gives the conclusions and further research directions.

2. Related works and motivation

The performance of a document understanding system greatly depends on the soundness of page segmentation and labelling of different regions such as texts, tables, lines drawings, and illustrations. Previous works can be classified into three categories: top-down [2,3,4], bottom-up [3,6], and hybrid techniques [1,8,10]. Experimental results have been done on newspapers [2,6,7], and papers [10]. A brief survey of geometric page layout analysis methods is given in [6]. Other approaches use run length smoothing and adaptive threshold. In [5], the document is

segmented into the X-Y projection, which merges regions using histograms of generic layout object. Reference [3] presents a definition language to express a geometric format. Reference [1] proposes the use of grammar to recognize a document and transform the image to HTML. Reference [6] proposes a bottom-up technique based on the extraction of connected components to efficiently implement page segmentation and region identification. Many methods extract columns of text and non-text and a lot of a priori information is needed (specific structures, thresholds, parameters, etc.). There are some cases that make it difficult to segment correctly using only the projection measure because textual regions are not organized only in rectangular boxes and can be recovered together with other objects.

To avoid this problem, Cheriet [8] demonstrates that scale space theory is a solid mathematical/physical framework for multi-scale noisy character recognition. Our motivation is to be able to segment a document image to detect an object in a higher scale, and then, gradually look at it more closely in a lower scale. The other point of view is that the text regions can be easily distinguished from the logo regions by the property of being aligned horizontally or vertically. To find this property, we concentrate on extracting the visual feature of text, logo, and line drawing regions.

3. Proposed Scale Space document analysis method

Given a grey scale image, the proposed method to segment a document image without a priori information, is described as follows:

1- Construct a scale space structure and get multiscale images of the input image (Section 3.1).
2 - Design system and extract information regions (Section 3.2).
3 - Calculate surface, entropy, and construct model document (Section 3.3).
4 - Identify kinds of objects extracted, based on shapes and entropy (Section 3.4).

3.1. Construction of scale space structure

This section provides the background theory of scale space and discusses the use of scale space for extracting the objects in a composite document without a priori information. The Gaussian kernel is established as the unique scale-space operator to change scale [12]. This operator is used to reduce noise at different levels, and an object is coarser at a larger scale and finer at a smaller scale. Then, the Laplacian operator is used to locate the information region. The standard deviation σ of the Gaussian kernel is used to scale the view and break the image into sub-images. We used the formula developed in [8] to model discretely the observed noise:

$$g_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}}$$

where g is the kernel operator to be determined, σ is the standard deviation and (x,y) indicates the pixel in the image.

The Laplacian of gaussian $L \circ G$ is defined as:

$$L \circ G_\sigma = \nabla^2 g_\sigma(x, y) = \left(\frac{x^2 + y^2}{\sigma^4} - \frac{1}{\sigma^2} \right) e^{\frac{-(x^2+y^2)}{2\sigma^2}}$$

This operator is usually used to detect the local shape in an image. The curvature obtained with a second order scalar property, is positive (convex) for the information region and negative (concave) for the background.

3.2. System design

Our algorithm operates in two phases (*Fig. 1*): bottom-up and top-down. In the bottom-up phase, the original image is convoluted by $L \circ G$ at different scales. Top-down technique then validates and identifies the objects obtained.

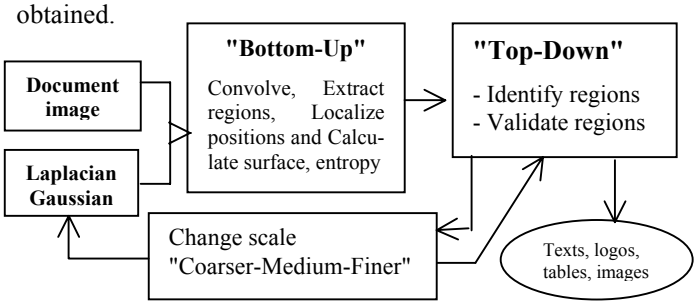

Figure 1. Diagram of scale space approach to document analysis

Algorithm: Extract and Identify content in document.

- **For scale σ** = coarser to finer
  Construct $L \circ G$ and **convolve** with grey level image;
  Extract information regions in image;
  Delete objects in a coarser scale when pixels do not exist at current scale;
  Extract extreme points, calculate surface, and entropy of each remaining object. Entropy is calculated on the values of pixels at a finer scale (considered as information without noise);
- **Matching** shape of objects with forms like line, square, rectangle, ellipse, etc.
- **Labelling** objects with terms like " text block, line text, logo, drawing line, illustration"

3.3. Document Model construction

We are concerned here with the question of how to generate a document representation. We represent the transformed image of a page as a quadruple

$\mathbf{P = (N, \mathfrak{I}, \mathfrak{R}_\sigma, \Psi_\sigma)}$ where:

- $\mathbf{N} = \bigcup_{i=1}^{c} n_i \left(\bigcup_{j=1}^{m} n_{ij} \left(\bigcup_{k=1}^{f} n_{ijk}\right)\right)$ is a list of objects and decompositions at each scale. Parameters c, m, and f are the numbers of remaining objects at respectively coarser, medium and finer scales.
  In figure *2,* $N=(D_1(O_1(O_2(O_{21}O_{22})O_3)\ \ldots)$.

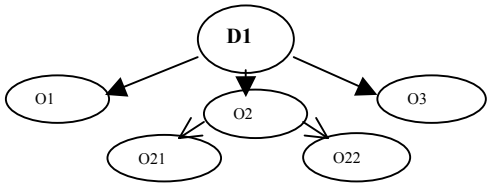

Figure 2. Decomposition of document D_1 at different scales.

- $\mathfrak{I} = \bigcup_{i=1}^{c} f_i \left(\bigcup_{j=1}^{m} f_{ij} \left(\bigcup_{k=1}^{f} f_{ijk}\right)\right)$ is a list of features corresponding to objects. Vector f_{ijk} contains surface, entropy, and the points of extreme Cartesian coordinates. For each localized region X, a horizontal projection histogram $H_p(X)$ is calculated after convolution. The size of H_p is the width of object X. Once this is done, we calculate horizontal entropy E_h that measures homogeneity for a given histogram [13]. $E_p(x)$ is calculated as follows:

$$E_p(x) = -\sum_{i=1}^{width} \frac{H_p(x)[i]}{surface} \ln\left(\frac{H_p(x)[i]}{surface}\right)$$

  where surface is a number of pixels in object X at a finer scale.
- $\mathbf{\mathfrak{R}_\sigma = \{r(n_i,n_j)\ /n_i,n_j \in N\}}$ is a set of relations between n_i and n_j at scale σ. $R(n_i,n_j)$ = {left, right, up, down, around, up-left, up-right, down-left, down-right} is determined by the extreme cartesian values of the objects.
- $\mathbf{\Psi_\sigma = \{l(n_i)\ /n_i \in N\}}$ is a set of logical labels assigned to information region n_i at scale σ. $l(n_i)$ = {Text, logo, line drawing, image} is determined by the shape, the surface, the entropy and the extreme cartesian values of the objects.

3.4. Region identification and validation

To contribute to document analysis, we consider the following criteria to match the shape of the extracted objects for identifying the extracted regions as texts, logos, line drawings, and images. The heuristics for region identification are:

✓ logo identification.
- shape of an object at high scale is oval, stretched in height or width (Surface $(n_i) \cong \pi*(length/2)^2*(width/2)^2$) and;
- entropy measures distribution histogram gives maximal values. The frequency distribution of pixels is homogeneous.

✓ Text identification.
- shape of an object is tuned best to a rectangle stretched in width (surface $(n_i) \cong$ length x width)
- length / width ratio is very small..
- density of a text area in an original image is relatively higher than an equation.
- a region has a horizontal periodicity and is composed of some peaks in the fine scale. Its entropy is small.

✓ Line drawing identification. We identify lines in a small scale, as objects with limited height or width. The ratio of width to height or height to width is small (surface (n) ≅ width or length). The entropy is near zero.

✓ Table identification. Horizontal lines similar in length, and with the text regions between them, are considered as a table.

✓ Image identification. The rest of the regions are regarded as image regions.

For validation purpose, which is undertaken, we extract more features in order to validate the identity of the actual region.

4. Experiments

Fifty images of the Media Team document database (Washington University) are used in our experiments. They come from various fields (journal papers, business cards, correspondences, and forms). The figures below show that the standard deviation can easily distinguish symbols, text lines and drawing lines. For the segmentation of the image in figure 3a, a coarser scale emphasized the logo only (Fig. 3b), a medium scale emphasized the logo and the text, and a fine scale will distinguish the formats stretched horizontally corresponding to text (Fig. 3d) or line drawing. For the illustrations, the dispersion of the black and white area is not uniform and causes the creation of multiple free form objects (Fig. 3c).

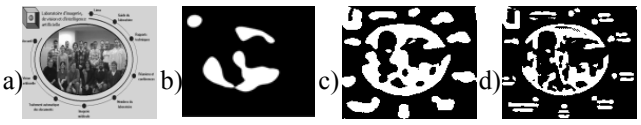

Figure 3. a) Original image segmented at b) scale 30 c) scale 10 d) scale 5

The two tables below show the results obtained by a successive decomposition of the image of documents.

Table 1 gives the results obtained for the image shown in figure 3b. At this scale, only logos or some portions of the image with high density are presented, and all objects, lines, columns, and surfaces have been detected correctly.

Table 1. Example of result of processing the image in figure 3b

Image/σ	Object	Lines	Columns	Surface	Type
1/30	1	40 to 96	20 to 76	2556	Logo
	2	94 to 169	206 to 356	5639	---
	3	206 to 393	81 to 373	23171	---
Total	Logos: 1	Text blocks: 0		Line draws: 0	

The entropy characteristic distinguishes text from logo and image. To validate the regions obtained in figure 3d, we calculate entropy of the objects obtained. The entropy of 19 text zones obtained has a mean value of 1.9 and a standard deviation of 0.24. But logo and image have a value greater than 5, and Logo has an elliptic shape on a large scale.

Table 2 presents global experimental results per document class. We observe that there is some over and under segmentation due to blank space in a region. The entropy calculated on the finer image eliminates rectangular zones confused with text regions.

Table 2. Experimental results: number of objects detected vs. objects in database for 4 values of σ

σ	Journals (18 images)	Business (10 images)	Corresp. (10 images)	Draws (2 images)	Forms (10 images)
30	Logos 3	Logos 9	Logos 0	Logos 3	Logos 4
15	Text 3 Line 0	Text 20 Line 0	Text 7 Line 0	Text 2 Line 0	Text 3 Line 0
10	Text 4 Line 0	Text 39 Line 0	Text 9 Line 0	Text 9 Line 0	Text 10 Line 0
5	Text 49 Line 6	Text 71 Line 1	Text 49 Line 3	Text 42 Line 10	Text 83 Line 2
1		Line 16			Line 90
Total in dbase	Logos 3 Texts 64	Logos11 Texts 57 Lines 5	Logos 0 Texts 45 Lines 0	Logos 3 Texts 47 Lines 27	Logos 4 Texts 79 Lines 48

These observations show that these measures are somewhat subjective and require thresholding to obtain better results, e.g. by adjusting the values of σ and the thresholds for space between blocks detected. We can note some over-segmentation cases due to the presence of broken lines in the image.

5. Conclusion

This article presents an approach to detect the objects contained in the composite images of documents. We first used Gaussian to remove the noise to create a multi-scale space and then the Laplacian to locate existing information regions in the image. After that, we used a top-down approach to refine the contents of the image on successive scales by re-using all the relevant information of the preceding stages. This technique has the advantage of directly exploiting the information contained in a hierarchy. It works well in general cases but can be improved by taking into account some specific cases. In addition, it is not reasonable to try to treat all the cases. We conclude from this study that, a good approach of detection and analysis must be able to analyze and to synthesize the behavior to be adopted in new situations.

References

[1] Seong-Whan Lee, and Dae-Seok Ryu (Nov. 2001) ″Parameter-Free Geometric Document Layout Analysis,″ IEEE Transactions on Pattern Analysis And Machine Intelligence, p. 1240-1256.

[2] D. Wang and S. N. Srihari (1989) ″Classification of newspaper image blocks using texture analysis,″ Computer Vision, Graphics, and Image Processing , v 47, p. 327--352.

[3] Y.Y. Tang, C.D. Yan, M. Cheriet, and C.Y. Suen (1997) ″Automatic analysis and understanding of documents,″ Handbook of Pattern Recognition and Computer Vision, p. 625-654.

[4] R. Ingold, and D. Armangil (1991) ″A top-down document analysis method for logical structure recognition,″ in Proc. Fourth Int. Conf. On Document Analysis and Recognition, St. Malo, France, p. 41-49.

[5] M. Krishnamoorthy and G. Nagy and S. Seth and M. Viswanathan (1993) ″Syntactic Segmentation and Labelling of Digitized Pages from Technical Journals,″ IEEE Computer Vision, Graphics and image processing, vol. 47, p. 327-352.

[6] A.K. Jain, Bin Yu (1998)) ″Document Representation and Its Application to Page Decomposition,″ IEEE Transactions on Pattern Analysis And Machine Intelligence, p. 294-308.

[7] Rolf Ingold, Olivier Hitz, Lyse Robadey (2000) ″Segmentation de documents à structure complexe,″ Cifed 2000 – Lyon – France.

[8] M. Cheriet, (1999) ″Extraction of Handwritten Data from Noisy Gray-level Images Using a Multi-Scale Approach,″ IJPRAI, Vol.13, No. 5, p. 665-685.

[9] J. Babaud, A. P. Withkin, M. Baudin, and R. O. Duda, (1986) "Uniqueness of the Gaussian kernel for scale-space filtring," IEEE Trans. PAMI, Vol. 8, no. 1, p. 26-33.

[10] Kyong-Ho Lee, Yoon-Chul Choy, and Sung-Bae Cho (Nov. 2000) "Geometric Structure Analysis of Document Images: A Knowledge-Based Approach," IEEE Trans. PAMI, Vol. 22, no. 11, p. 1224-1240.

[11] S. M. Kerpedjiev (1997) ″Automatic extraction of information structures for document,″ in Proc. Fourth Int. Conf. On Document Analysis and Recognition, Ulm- Germany, p. 32-40.

[12] T. Lindeberg (1994) ″Scale-Space Theory in computer Vision,″ Kluwer Academic Publishers, Boston.

[13] Duong et al (2001) ″Extraction of text areas in printed document images,″ Atlanta, Georgia, p. 157-165.