Document Layout Analysis

Outline

- Introduction
- Projection method
- Morphology method
- Database
- Deep learning method
- Performance evaluation

Introduction

- Problem description
- Motivation
  - Improve performance of OCR
  - Graphics recognition
  - Table detection
  - Browsing and navigation
- Physical and logical structure

Problem Description

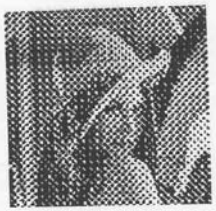


Figure 1: Example of figure inclusion.

This page is used to develop and algorithm for document page image segmentation. Page should contain mathematical equations, tables, section headings, graphics or line drawing and half-tone images alongwith regular text. The page is generated using LaTex word processing system. Then the page is printed on a HP-5MP laserjet printer, followed by scanned in by HP 4C scanner. Scanned in image may have tilt.

1 Math equations

Below we give an example of mathematical equation computing mean of a set of data points followed by the formula for computing variances of the same. Equations with or without equation number are given.

$$\overline{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

Following one is an example of equation array with equation number. However, other style of equations are not considered here because of lack of space. So the following equation is the example of the second type that is equation with number.

$$\begin{aligned} \sigma_x &= \frac{1}{n}\sum_{i=0}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n}\sum_{i=0}^{n} x_i^2 - \overline{x}^2 \end{aligned} \qquad (1)$$

Now we give an example of table that has column and row separators as well as captions and number. Finally we put a line drawing which is basically a normal probability density function. Hence this page illustrates all the basic entities we like to deal with in a document page image segmentation. Here is the end of this page.

Table 1: Example of table

Roll	Lit.	Sci.	Soc.	Math.
1	38	95	67	85
23	67	83	19	78
43	17	93	13	92
67	19	16	34	56

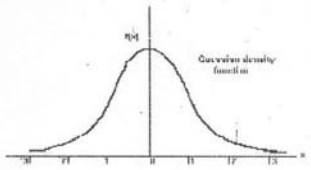


Figure 2: Example of Graphics inclusion.

Thanks for your attention. Resolution of scanning is 100 dpi (i.e., dots per inch).

Objective

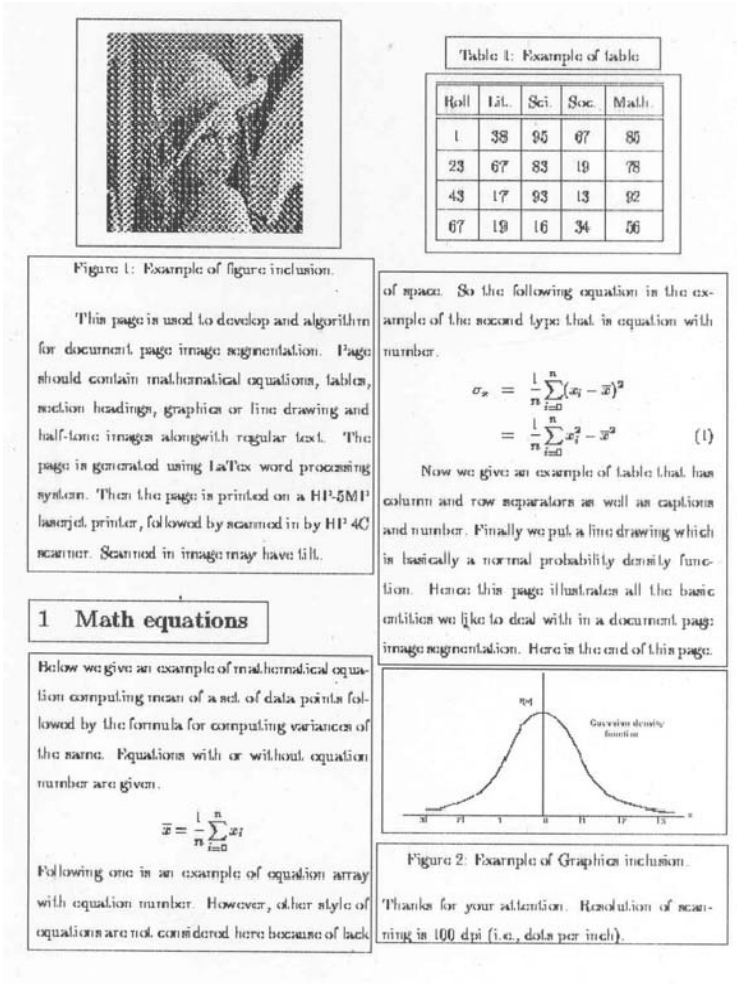


Table 1: Example of table

Roll	Lit.	Sci.	Soc.	Math.
1	38	95	67	85
23	67	83	19	78
43	17	93	13	92
67	19	16	34	56

Figure 1: Example of figure inclusion.

This page is used to develop and algorithm for document page image segmentation. Page should contain mathematical equations, tables, section headings, graphics or line drawing and half-tone images alongwith regular text. The page is generated using LaTex word processing system. Then the page is printed on a HP-5MP laserjet printer, followed by scanned in by HP 4C scanner. Scanned in image may have tilt.

1 Math equations

Below we give an example of mathematical equation computing mean of a set of data points followed by the formula for computing variances of the same. Equations with or without equation number are given.

$$\overline{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

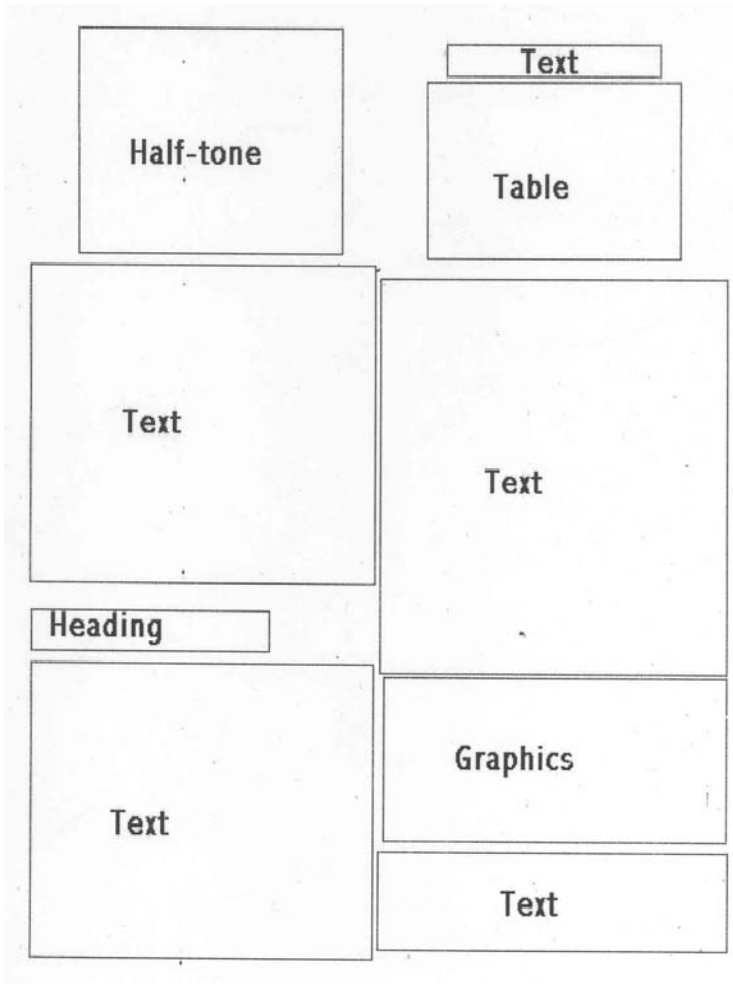
Following one is an example of equation array with equation number. However, other style of equations are not considered here because of lack of space. So the following equation is the example of the second type that is equation with number.

$$\begin{aligned} \sigma_x &= \frac{1}{n}\sum_{i=0}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n}\sum_{i=0}^{n} x_i^2 - \overline{x}^2 \end{aligned} \qquad (1)$$

Now we give an example of table that has column and row separators as well as captions and number. Finally we put a line drawing which is basically a normal probability density function. Hence this page illustrates all the basic entities we like to deal with in a document page image segmentation. Here is the end of this page.

Figure 2: Example of Graphics inclusion.

Thanks for your attention. Resolution of scanning is 100 dpi (i.e., dots per inch).



版面理解技术

https://www.icst.pku.edu.cn/cpdp/yjfx/bmljjz/index.htm

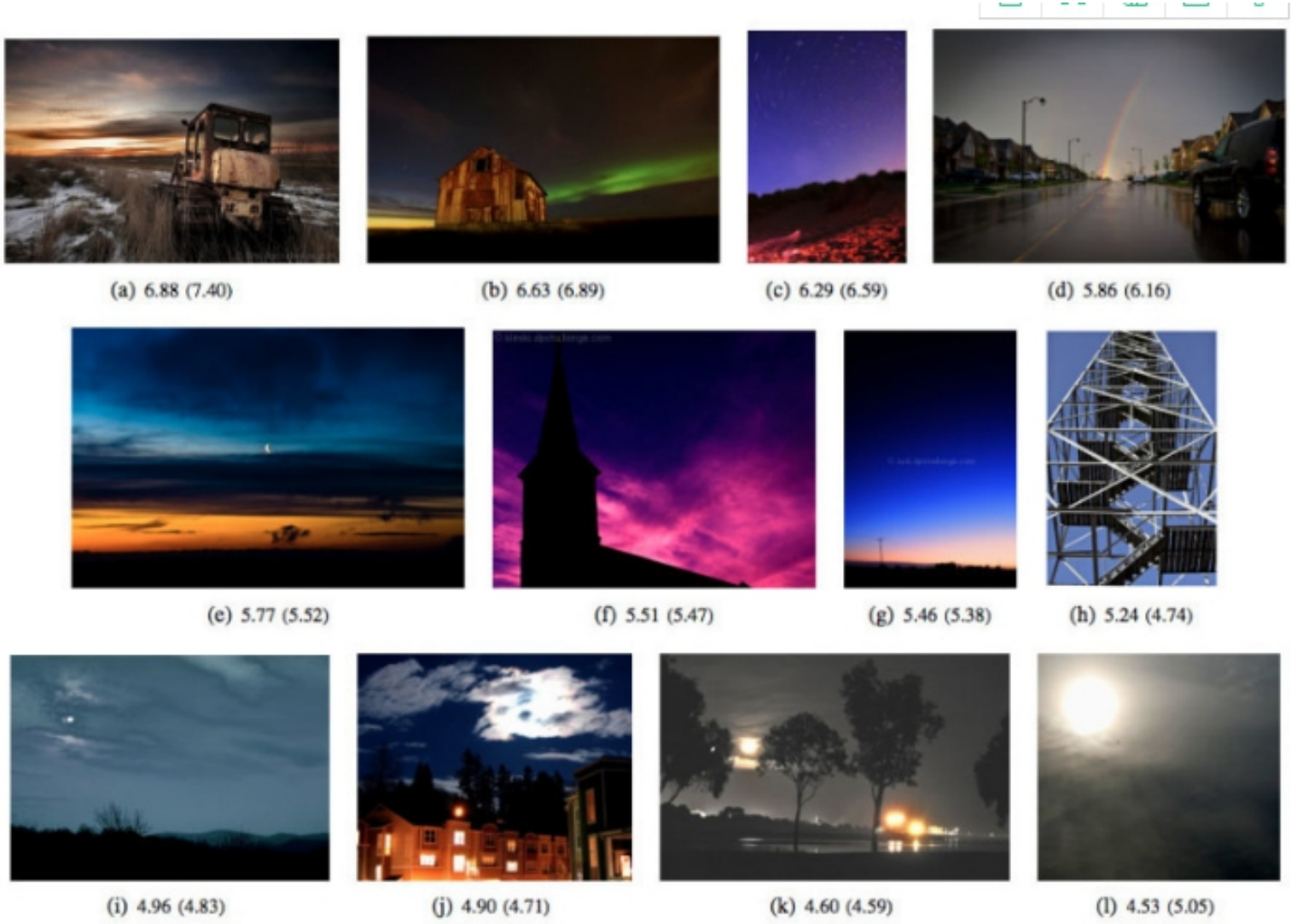
(a)

(b)

(c)

(d)

https://www.icst.pku.edu.cn/cpdp/yjfx/bmljjz/index.htm



Murray N, Marchesotti L, Perronnin F. AVA: A large-scale database for aesthetic visual analysis[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2408-2415.

Major Source of Document Pages

1. Books
2. Journals
3. Magazines
4. Newspapers
5. Forms and leaflets
6. Reports

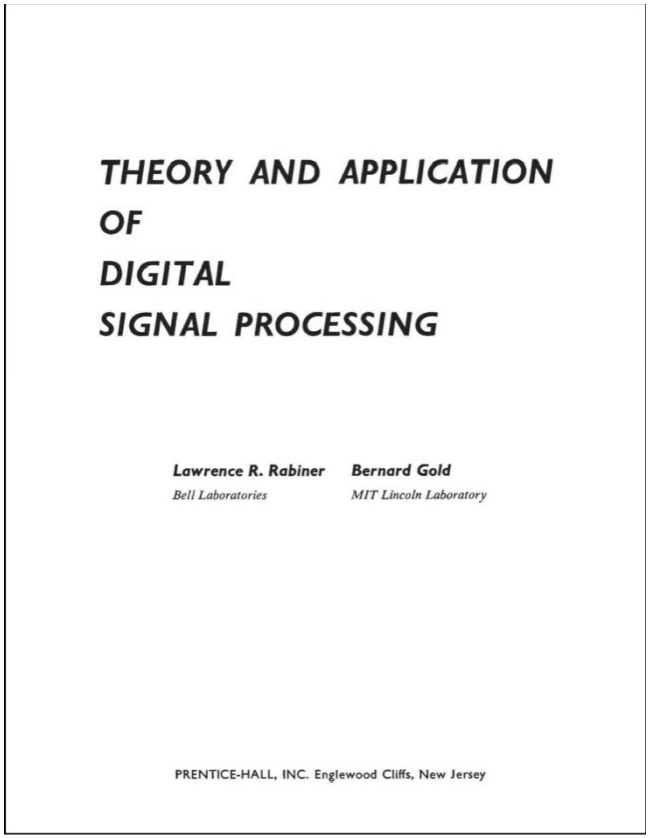
Types of document pages

Consider books and journals

- Title page
- Publisher’s page
- Table of Contents
- Text page
- Index page

Different types of pages

Title page

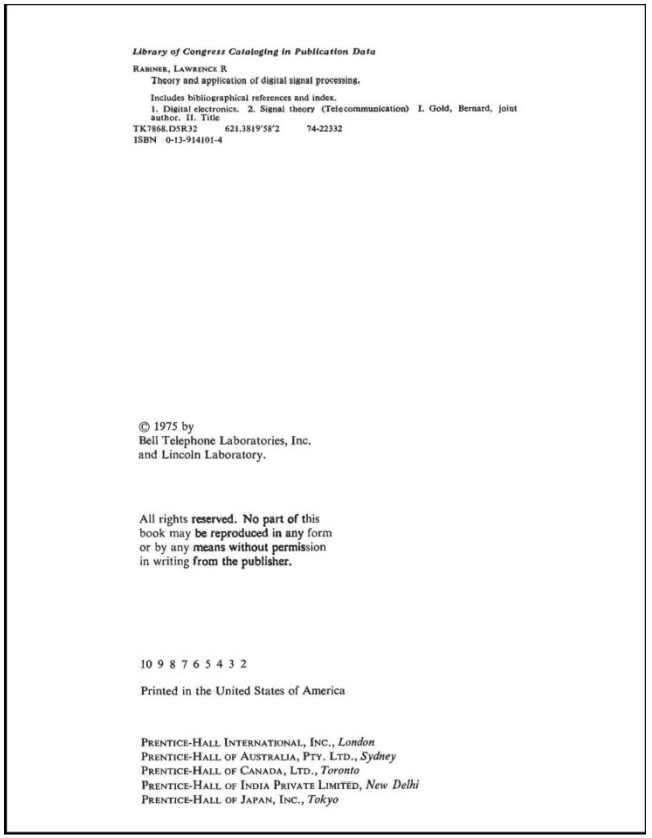


THEORY AND APPLICATION
OF
DIGITAL
SIGNAL PROCESSING

Lawrence R. Rabiner — *Bell Laboratories*
Bernard Gold — *MIT Lincoln Laboratory*

PRENTICE-HALL, INC. Englewood Cliffs, New Jersey

Publisher's page



Library of Congress Cataloging in Publication Data

RABINER, LAWRENCE R
Theory and application of digital signal processing.

Includes bibliographical references and index.
1. Digital electronics. 2. Signal theory (Telecommunication) I. Gold, Bernard, joint author. II. Title
TK7868.D5R32 621.3819'58'2 74-22332
ISBN 0-13-914101-4

10 9 8 7 6 5 4 3 2

Printed in the United States of America

PRENTICE-HALL INTERNATIONAL, INC., *London*
PRENTICE-HALL OF AUSTRALIA, PTY. LTD., *Sydney*
PRENTICE-HALL OF CANADA, LTD., *Toronto*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*

Different types of pages

Table of Content page

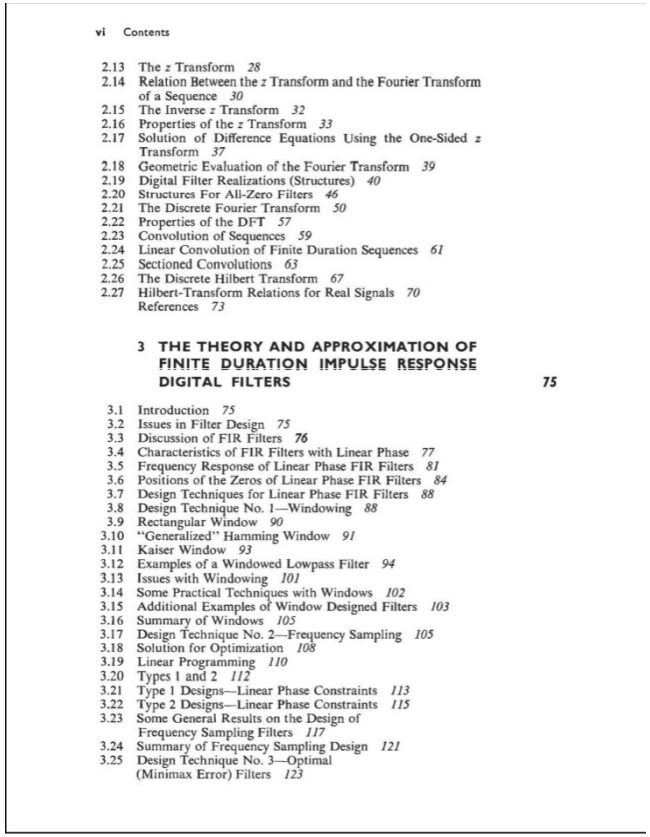
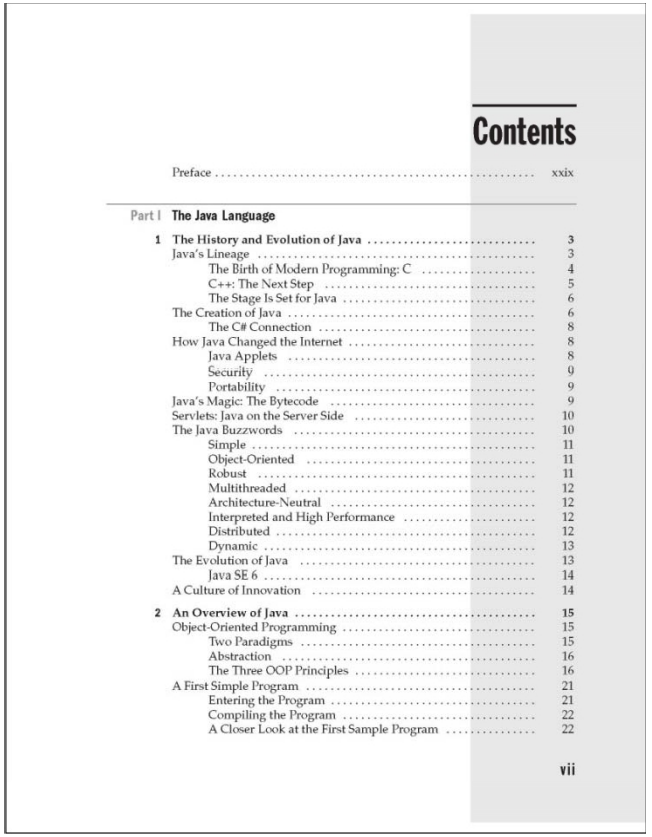


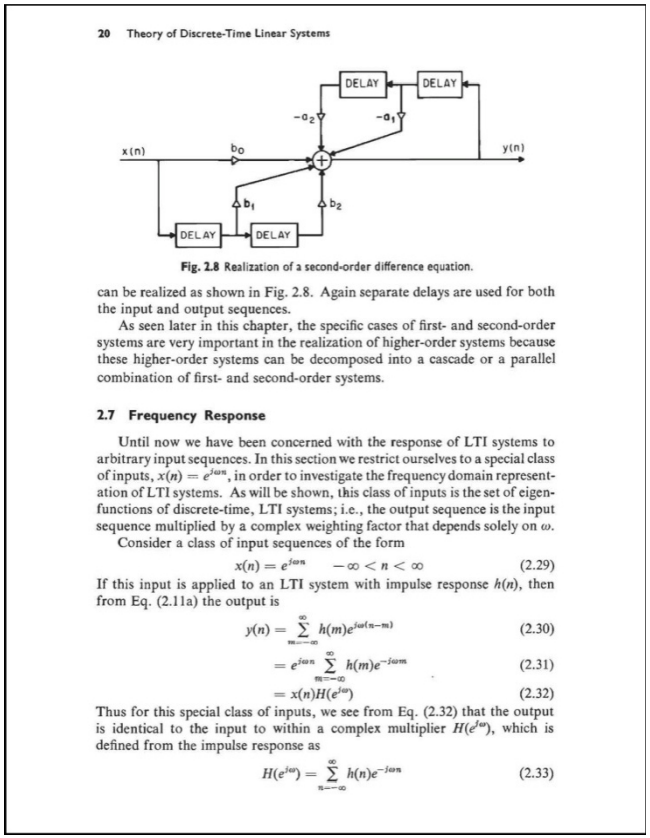
Table of Content page



Contents

Different types of pages

Text page-1



20 Theory of Discrete-Time Linear Systems

Fig. 2.8 Realization of a second-order difference equation.

can be realized as shown in Fig. 2.8. Again separate delays are used for both the input and output sequences.

As seen later in this chapter, the specific cases of first- and second-order systems are very important in the realization of higher-order systems because these higher-order systems can be decomposed into a cascade or a parallel combination of first- and second-order systems.

2.7 Frequency Response

Until now we have been concerned with the response of LTI systems to arbitrary input sequences. In this section we restrict ourselves to a special class of inputs, $x(n) = e^{j\omega n}$, in order to investigate the frequency domain representation of LTI systems. As will be shown, this class of inputs is the set of eigenfunctions of discrete-time, LTI systems; i.e., the output sequence is the input sequence multiplied by a complex weighting factor that depends solely on ω.

Consider a class of input sequences of the form

$$x(n) = e^{j\omega n} \qquad -\infty < n < \infty \tag{2.29}$$

If this input is applied to an LTI system with impulse response $h(n)$, then from Eq. (2.11a) the output is

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)e^{j\omega(n-m)} \tag{2.30}$$

$$= e^{j\omega n} \sum_{m=-\infty}^{\infty} h(m)e^{-j\omega m} \tag{2.31}$$

$$= x(n)H(e^{j\omega}) \tag{2.32}$$

Thus for this special class of inputs, we see from Eq. (2.32) that the output is identical to the input to within a complex multiplier $H(e^{j\omega})$, which is defined from the impulse response as

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n} \tag{2.33}$$

Text page-2

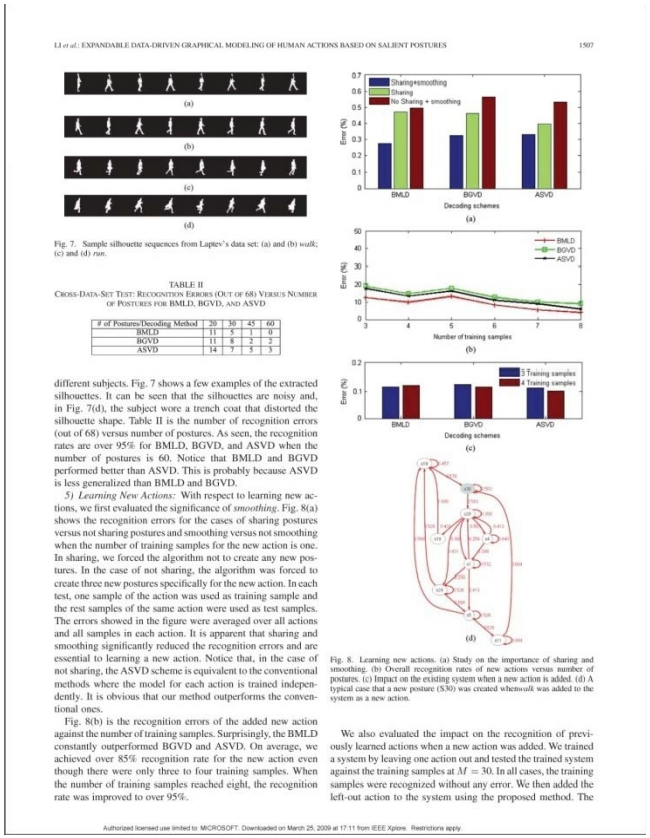


Fig. 7. Sample silhouette sequences from Laptev's data set: (a) and (b) *walk*; (c) and (d) *run*.

TABLE II
CROSS-DATA-SET TEST: RECOGNITION ERRORS (OUT OF 68) VERSUS NUMBER OF POSTURES FOR BMLD, BGVD, AND ASVD

# of Postures/Decoding Method	20	30	45	60
BMLD	11	5	1	0
BGVD	11	8	2	2
ASVD	14	7	5	3

different subjects. Fig. 7 shows a few examples of the extracted silhouettes. It can be seen that the silhouettes are noisy and, in Fig. 7(d), the subject wore a trench coat that distorted the silhouette shape. Table II is the number of recognition errors (out of 68) versus number of postures. As seen, the recognition rates are over 95% for BMLD, BGVD, and ASVD when the number of postures is 60. Notice that BMLD and BGVD performed better than ASVD. This is probably because ASVD is less generalized than BMLD and BGVD.

5) Learning New Actions: With respect to learning new actions, we first evaluated the significance of *smoothing*. Fig. 8(a) shows the recognition errors for the cases of sharing postures versus not sharing postures and smoothing versus not smoothing when the number of training samples for the new action is one. In sharing, we forced the algorithm not to create any new postures. In the case of not sharing, the algorithm was forced to create three new postures specifically for the new action. In each test, one sample of the action was used as training sample and the rest samples of the same action were used as test samples. The errors showed in the figure were averaged over all actions and all samples in each action. It is apparent that sharing and smoothing significantly reduced the recognition errors and are essential to learning a new action. Notice that, in the case of not sharing, the ASVD scheme is equivalent to the conventional methods where the model for each action is trained independently. It is obvious that our method outperforms the conventional ones.

Fig. 8(b) is the recognition errors of the added new action against the number of training samples. Surprisingly, the BMLD constantly outperformed BGVD and ASVD. On average, we achieved over 85% recognition rate for the new action even though there were only three to four training samples. When the number of training samples reached eight, the recognition rate was improved to over 95%.

Fig. 8. Learning new actions. (a) Study on the importance of sharing and smoothing. (b) Overall recognition rates of new actions versus number of postures. (c) Impact on the existing system when a new action is added. (d) A typical case that a new posture (S30) was created when *walk* was added to the system as a new action.

We also evaluated the impact on the recognition of previously learned actions when a new action was added. We trained a system by leaving one action out and tested the trained system against the training samples at $M = 30$. In all cases, the training samples were recognized without any error. We then added the left-out action to the system using the proposed method. The

Different types of pages

Text page-3

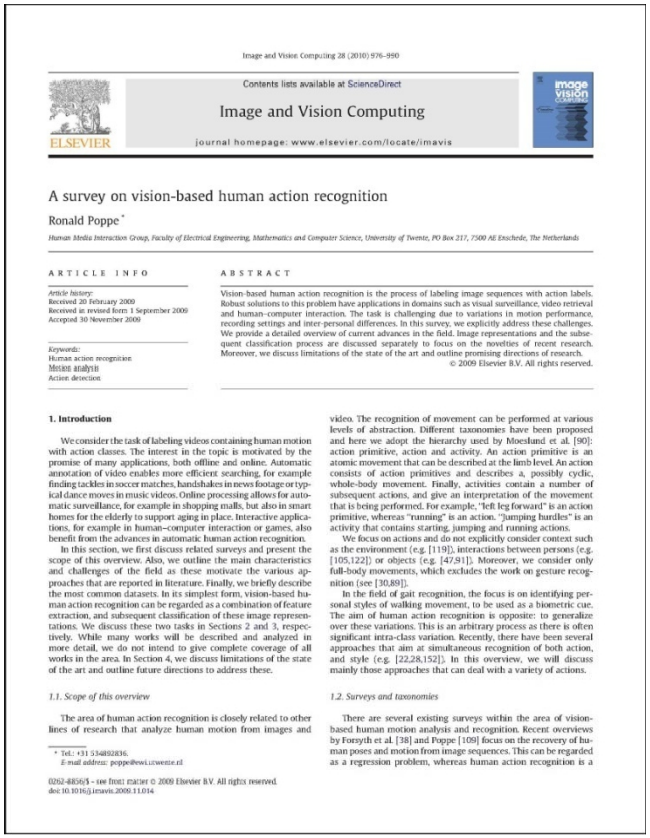


Image and Vision Computing 28 (2010) 976–990

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A survey on vision-based human action recognition

Ronald Poppe *

Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 20 February 2009
Received in revised form 1 September 2009
Accepted 30 November 2009

Keywords:
Human action recognition
Motion analysis
Action detection

ABSTRACT

Vision-based human action recognition is the process of labeling image sequences with action labels. Robust solutions to this problem have applications in domains such as visual surveillance, video retrieval and human–computer interaction. The task is challenging due to variations in motion performance, recording settings and inter-personal differences. In this survey, we explicitly address these challenges. We provide a detailed overview of current advances in the field. Image representations and the subsequent classification process are discussed separately to focus on the novelties of recent research. Moreover, we discuss limitations of the state of the art and outline promising directions of research.

1. Introduction

We consider the task of labeling videos containing human motion with action classes. The interest in the topic is motivated by the promise of many applications, both offline and online. Automatic annotation of video enables more efficient searching, for example finding tackles in soccer matches, handshakes in news footage or typical dance moves in music videos. Online processing allows for automatic surveillance, for example in shopping malls, but also in smart homes for the elderly to support aging in place. Interactive applications, for example in human–computer interaction or games, also benefit from the advances in automatic human action recognition.

In this section, we first discuss related surveys and present the scope of this overview. Also, we outline the main characteristics and challenges of the field as these motivate the various approaches that are reported in literature. Finally, we briefly describe the most common datasets. In its simplest form, vision-based human action recognition can be regarded as a combination of feature extraction, and subsequent classification of these image representations. We discuss these two tasks in Sections 2 and 3, respectively. While many works will be described and analyzed in more detail, we do not intend to give complete coverage of all works in the area. In Section 4, we discuss limitations of the state of the art and outline future directions to address these.

1.1. Scope of this overview

The area of human action recognition is closely related to other lines of research that analyze human motion from images and video. The recognition of movement can be performed at various levels of abstraction. Different taxonomies have been proposed and here we adopt the hierarchy used by Moeslund et al. [90]: action primitive, action and activity. An action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the movement that is being performed. For example, "left leg forward" is an action primitive, whereas "running" is an action. "Jumping hurdles" is an activity that contains starting, jumping and running actions.

We focus on actions and do not explicitly consider context such as the environment (e.g. [119]), interactions between persons (e.g. [105,122]) or objects (e.g. [47,91]). Moreover, we consider only full-body movements, which excludes the work on gesture recognition (see [30,89]).

In the field of gait recognition, the focus is on identifying personal styles of walking movement, to be used as a biometric cue. The aim of human action recognition is opposite: to generalize over these variations. This is an arbitrary process as there is often significant intra-class variation. Recently, there have been several approaches that aim at simultaneous recognition of both action, and style (e.g. [22,28,152]). In this overview, we will discuss mainly those approaches that can deal with a variety of actions.

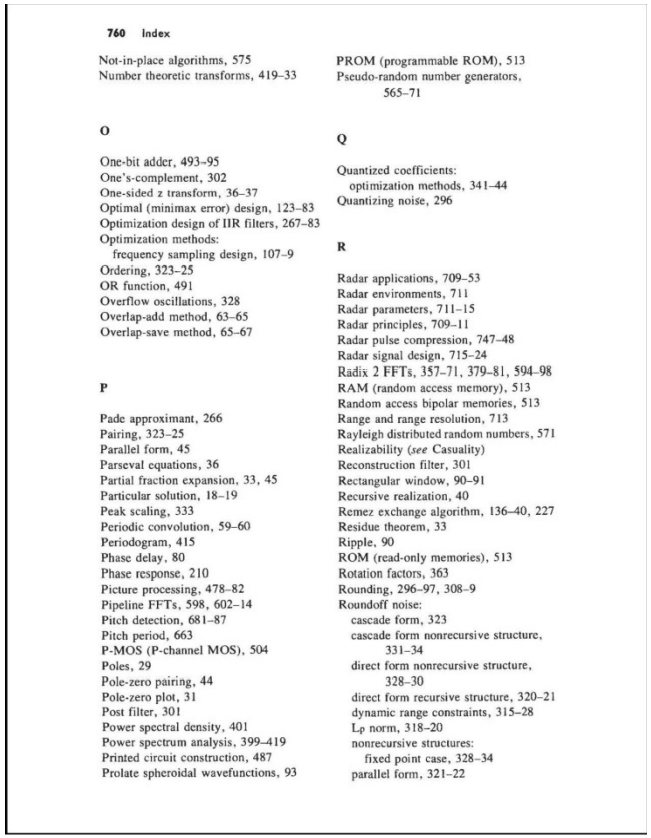
1.2. Surveys and taxonomies

There are several existing surveys within the area of vision-based human motion analysis and recognition. Recent overviews by Forsyth et al. [38] and Poppe [109] focus on the recovery of human poses and motion from image sequences. This can be regarded as a regression problem, whereas human action recognition is a

* Tel.: +31 534892836.
E-mail address: poppe@ewi.utwente.nl

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.11.014

Index page

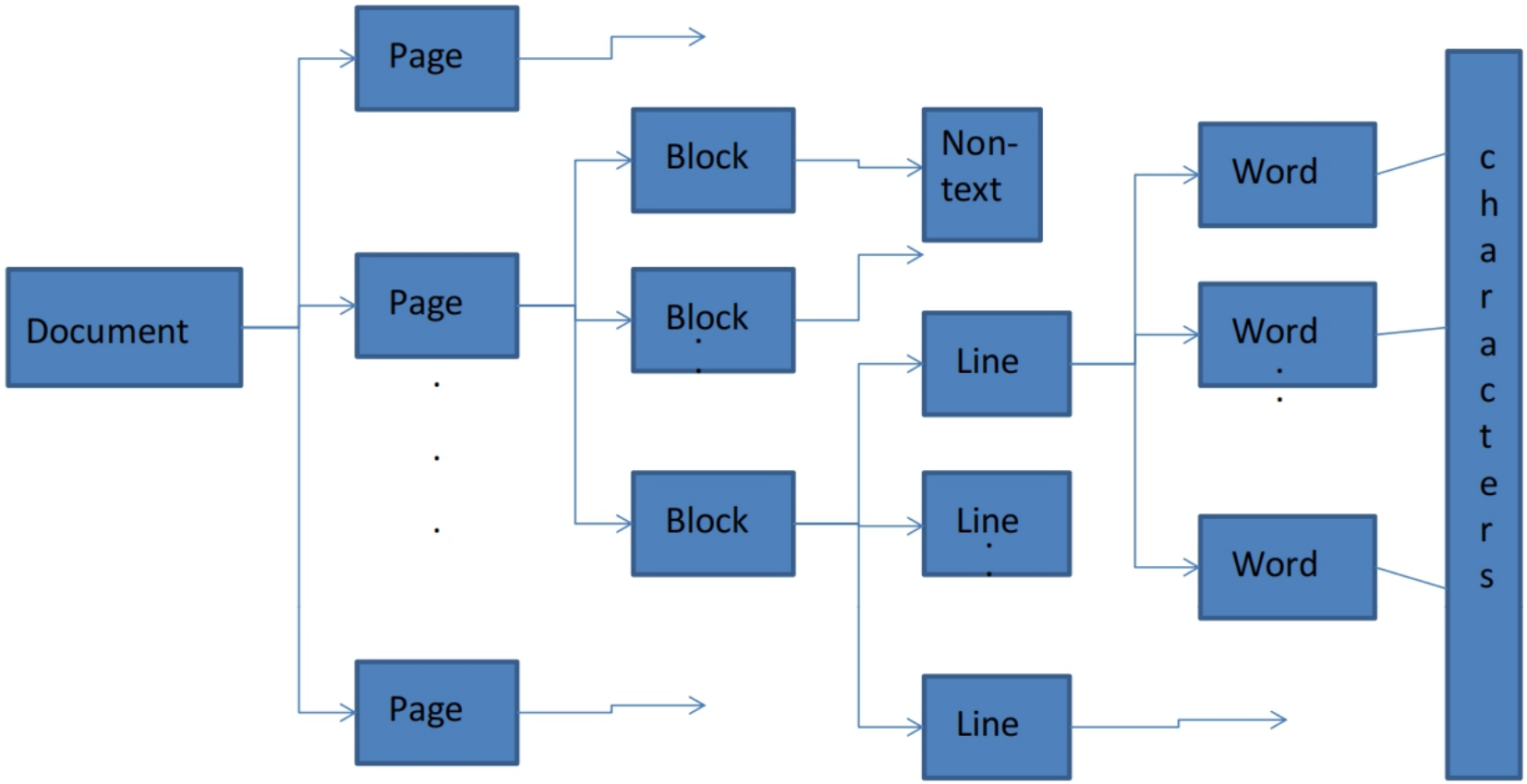


Different types of pages

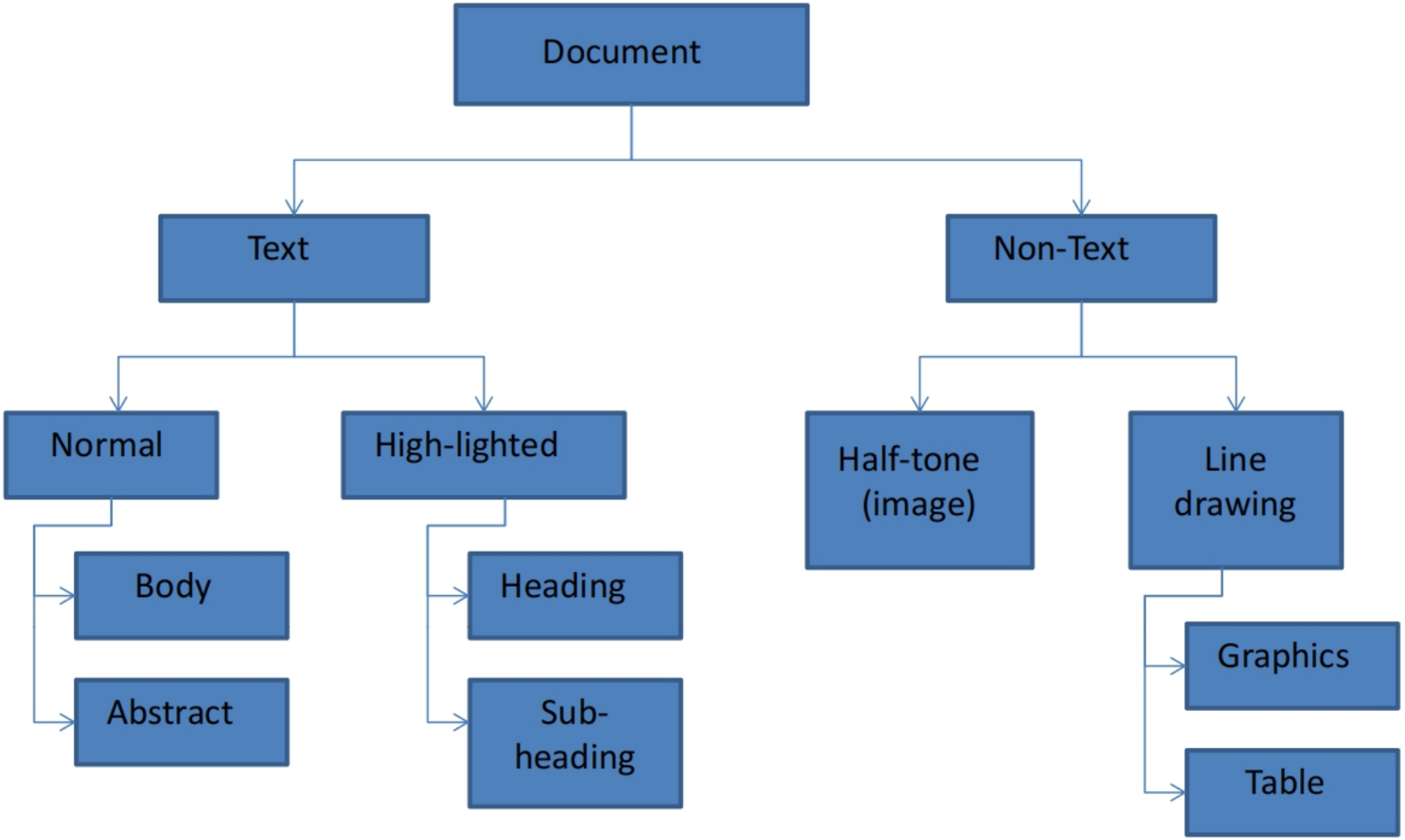
Fig. 1. Document Layouts: (a) Regular, (b) Manhattan-based, (c) Non-Manhattan, (d) Multi-column Manhattan, (e) Arbitrary Complex, (f) Overlapping horizontally and diagonally.

Binmakhashen G M, Mahmoud S A. Document layout analysis: a comprehensive survey[J]. ACM Computing Surveys (CSUR), 2019, 52(6): 1-36.

Geometrical / Physical structure



Logical structure



Entities of Document Page

- Text
  - Body text
    - Line → Word → Character
  - Heading
- Non-text
  - Half-tone
  - Table
  - Graphics or line drawing

Entities of Document Page

Figure 6: Example real documents a their corresponding segmentation. Top: DSSE-200. Middle: ICDAR2015. Bottom: SectLabel. Since these documents ar not in PDF format, the simple post-processing in Sec. 5 can not be applied. One may consider exploiting a CRF [13] to refine the segmentation, but that is beyond the main focus of this paper. Segmentation label colors are: **figure**, **table**, **section heading**, **caption**, **list** and **paragraph**.

Entities of Document Page

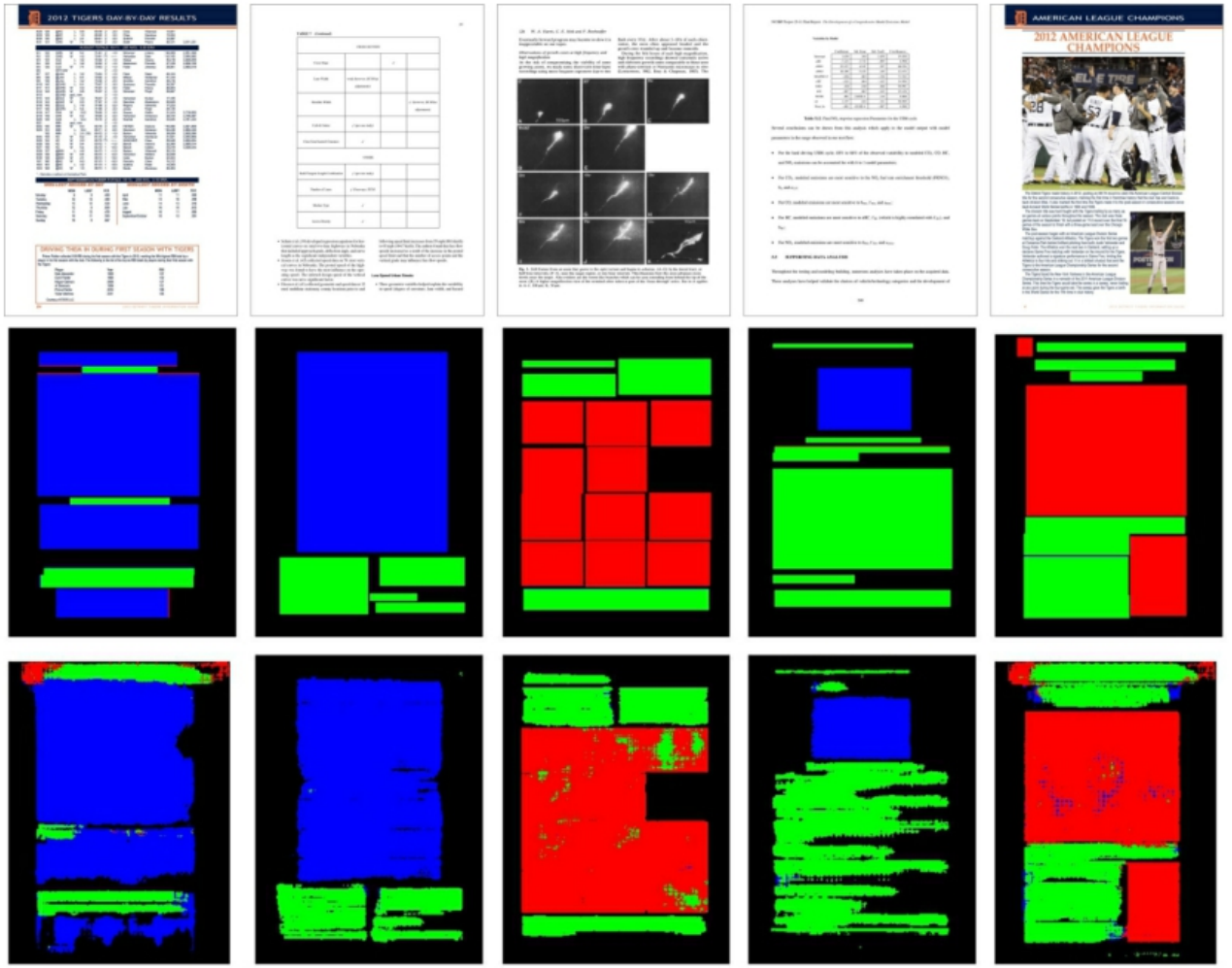


Figure 9: Example real documents and their corresponding segmentation. Top: original. Middle: ground-truth. Bottom: predictions. Segmentation label colors are: figure, table, text and background.

Entities of Document Page

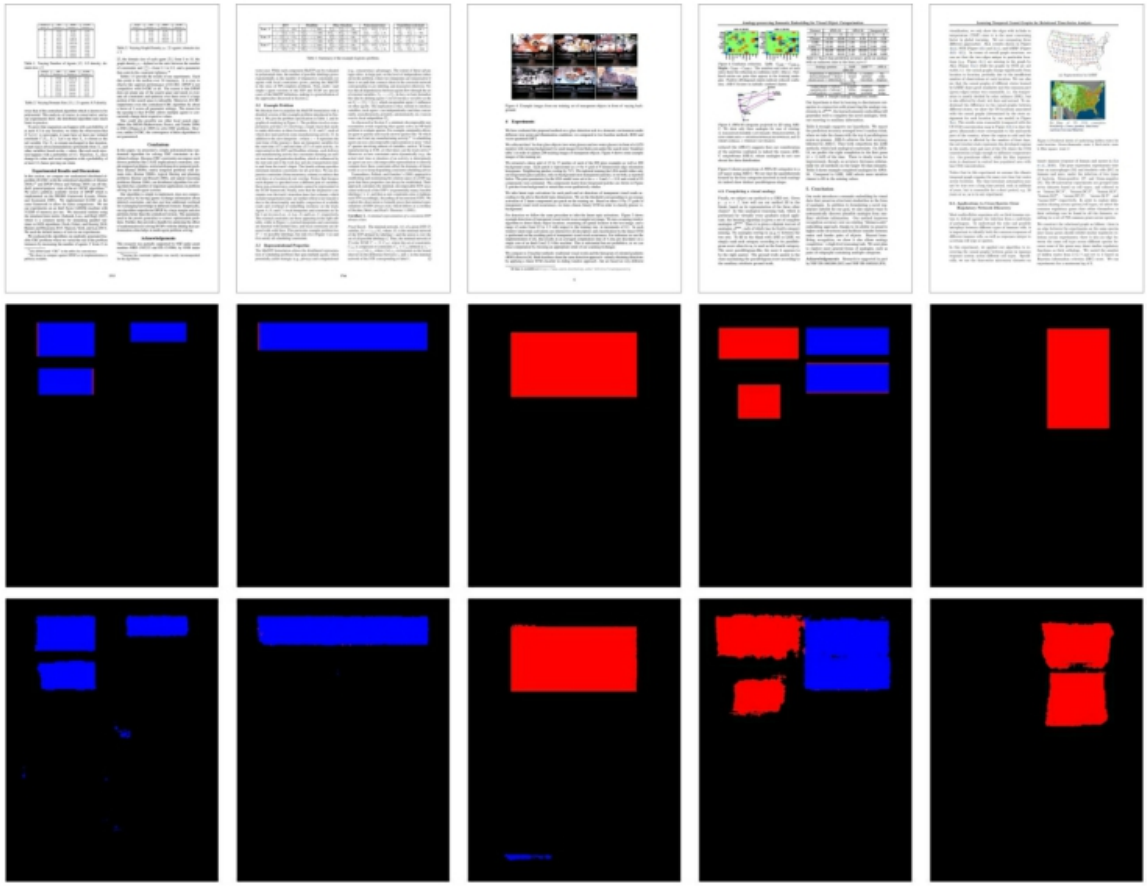
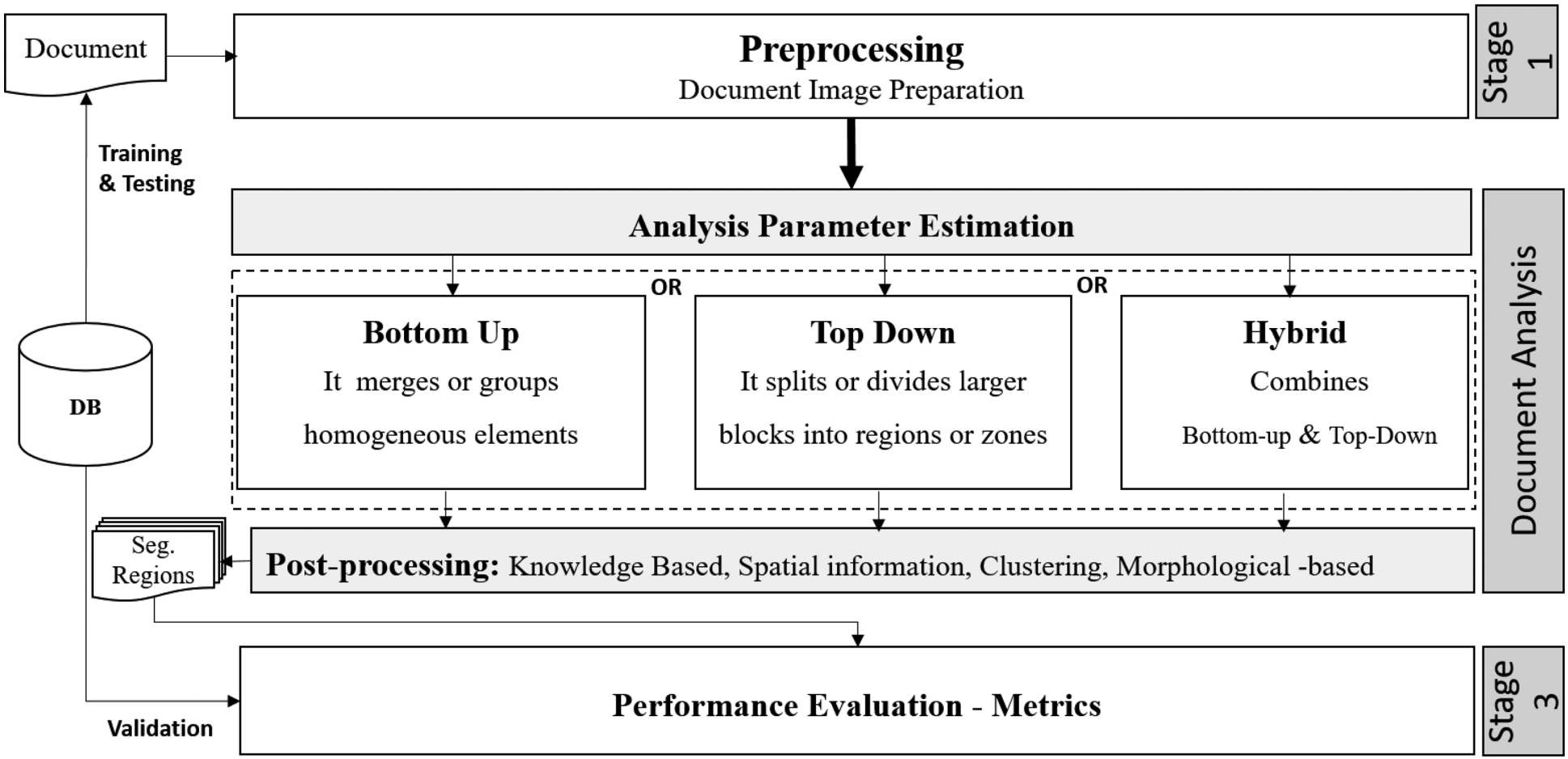


Figure 11: Example real documents and their corresponding segmentation. Top: original. Middle: ground-truth. Bottom:predictions. Segmentation label colors are: figure, table and non-text.

General Document Layout Analysis (DLA) framework

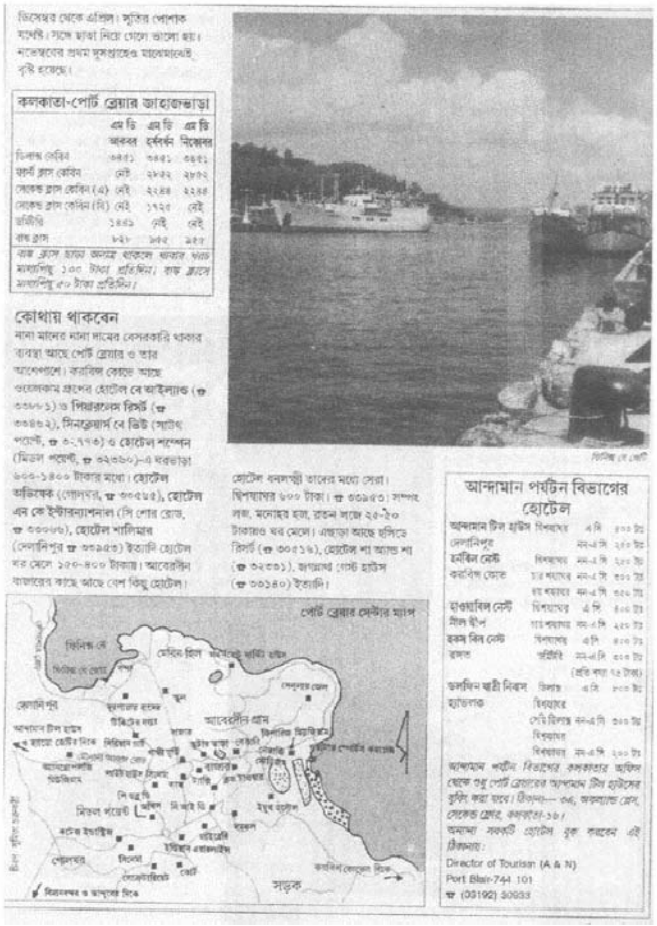


Zone / block detection

- One of the simple way is ***Projection method.***
- Algorithm
  - Take horizontal (or vertical) projection of foreground pixels. (may be implemented as pixel count)
  - If there exists a characteristic change in projection profile, put a horizontal (resp. vertical) separator.
  - Take horizontal and vertical direction alternately.

    Continue, until above condition is satisfied.
- Works well for ***structured document***, usually the pages of technical journals, books, etc.

Projection Method: An Example



কলকাতা-পোর্ট ব্লেয়ার জাহাজভাড়া

	এসি ফার্স্ট ক্লাস	এসি টু টায়ার	এসি থ্রি টায়ার
ডিলাক্স বেডিং	৩৪৫১	৩৪৫১	৩৪৫১
ফার্স্ট ক্লাস বেডিং	নেই	২৮৫২	২৮৫২
সেকেন্ড ক্লাস বেডিং (এ)	নেই	২২৪৪	২২৪৪
সেকেন্ড ক্লাস বেডিং (বি)	নেই	১৭২৮	নেই
ডরমিটরি	১৪৫৩	নেই	নেই
ডেক ক্লাস	৮৫৮	৮৫৮	৮৫৮

ডেক ক্লাসে ছাড়া অন্যত্র থাকলে অতিরিক্ত খরচ মাথাপিছু ১০০ টাকা প্রতিদিন। ডেক ক্লাসে মাথাপিছু ৫০ টাকা প্রতিদিন।

কোথায় থাকবেন

নানা মানের নানা দামের বেসরকারি থাকার ব্যবস্থা আছে পোর্ট ব্লেয়ার ও তার আশেপাশে। করবিন কোভে আছে ওয়েলকাম গ্রুপের হোটেল বে আইল্যান্ড (☎ ৩৩৬৮১) ও পিয়ারলেস রিসর্ট (☎ ৩৩৪৮২), সিনক্লেয়ার্স বে ভিউ (সাউথ পয়েন্ট, ☎ ৩-২৭৭৩) ও হোটেল শম্পেন (মিডল পয়েন্ট, ☎ ৩২৩৪৮)—এ ধরনটা ৬০০-১৪০০ টাকার মধ্যে। হোটেল অভিষেক (গোলঘর, ☎ ৩০৫৮৫), হোটেল এন কে ইন্টারন্যাশনাল (সি পোর্ট রোড, ☎ ৩৩০৩৬), হোটেল শালিমার (নেপালিপুর ☎ ৩৩৯৫৩) ইত্যাদি হোটেল ঘর মেলে ১২০-৪০০ টাকায়। আবেরডীন বাজারের কাছে আছে বেশ কিছু হোটেল।

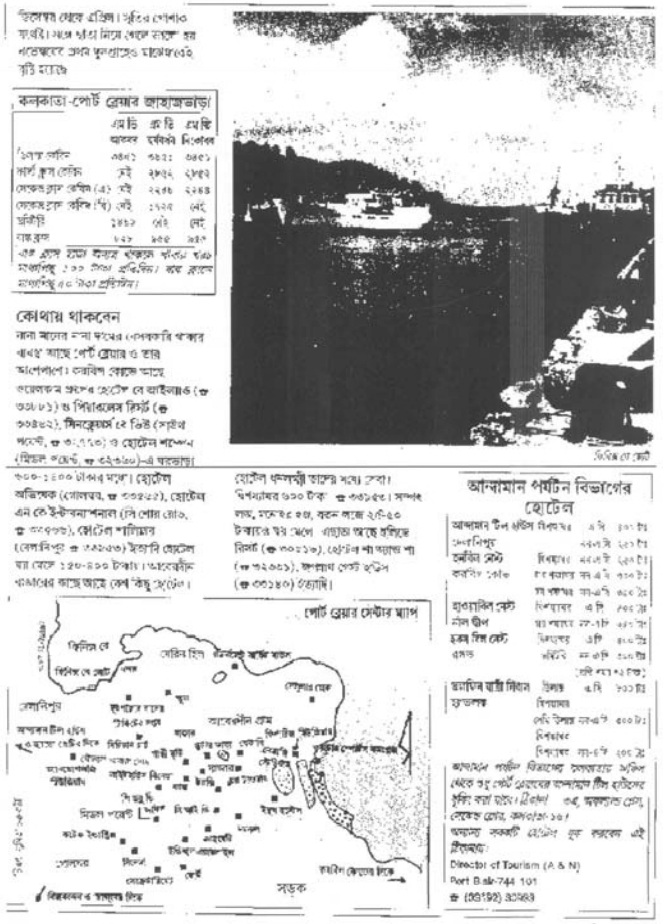
হোটেল বন্দরপ্রী তালের মধ্যে সেরা। দ্বিশয্যার ৬০০ টাকা। ☎ ৩৩৯৫৩। সাগর লজ, মনোহর হল, রতন লজে ২৫-৫০ টাকায়ও ঘর মেলে। এছাড়া আছে হলিডে রিসর্ট (☎ ৩০২১৬), হোটেল শা অ্যান্ড শা (☎ ৩২৩৬১), জগন্নাথ গেস্ট হাউস (☎ ৩৩১৪০) ইত্যাদি।

আন্দামান পর্যটন বিভাগের হোটেল

আন্দামান টিল হাউজ	দ্বিশয্যার	এসি	৫০০ টা
নেপালিপুর		নন-এসি	২৫০ টা
হর্ণবিল নেস্ট	দ্বিশয্যার	নন-এসি	২৫০ টা
করবিন কোভ	২য় শয্যার	নন-এসি	৩০০ টা
	৪র্থ শয্যার	নন-এসি	৩০০ টা
হ্যাণ্ডাবিল নেস্ট	দ্বিশয্যার	এ সি	৪০০ টা
নীল ঊীপ	৩য় শয্যার	নন-এসি	২৫০ টা
ডরম বিল্ড নেস্ট	দ্বিশয্যার	এসি	৪০০ টা
ধরমশালা	প্রতিটি	নন-এসি	৫০০ টা
			(প্রতি শয্যা ১২ টাকা)
ফলবিন ঘাটী নিবাস	ডিলাক্স	এ সি	৮০০ টা
হ্যাডডাক	দ্বিশয্যার		
	সেট রিলাক্স	নন-এসি	৩০০ টা
	দ্বিশয্যার		
	দ্বিশয্যার	নন-এসি	২০০ টা

আন্দামান পর্যটন বিভাগের কলকাতার অফিস থেকে শুধু পোর্ট ব্লেয়ারের আন্দামান টিল হাউজের বুকিং করা যাবে। ঠিকানা— ৩৬, জওহরলাল নেহরু রোড, কলকাতা-১৬।

অন্যান্য সরকারি হোটেল বুক করবেন এই ঠিকানায়:
Director of Tourism (A & N)
Port Blair-744 101
☎ (03192) 30333



কলকাতা-পোর্ট ব্লেয়ার জাহাজভাড়া

	এসি ফার্স্ট ক্লাস	এসি টু টায়ার	এসি থ্রি টায়ার
ডিলাক্স বেডিং	৩৪৫১	৩৪৫১	৩৪৫১
ফার্স্ট ক্লাস বেডিং	নেই	২৮৫২	২৮৫২
সেকেন্ড ক্লাস বেডিং (এ)	নেই	২২৪৪	২২৪৪
সেকেন্ড ক্লাস বেডিং (বি)	নেই	১৭২৮	নেই
ডরমিটরি	১৪৫৩	নেই	নেই
ডেক ক্লাস	৮৫৮	৮৫৮	৮৫৮

ডেক ক্লাসে ছাড়া অন্যত্র থাকলে অতিরিক্ত খরচ মাথাপিছু ১০০ টাকা প্রতিদিন। ডেক ক্লাসে মাথাপিছু ৫০ টাকা প্রতিদিন।

কোথায় থাকবেন

নানা মানের নানা দামের বেসরকারি থাকার ব্যবস্থা আছে পোর্ট ব্লেয়ার ও তার আশেপাশে। করবিন কোভে আছে ওয়েলকাম গ্রুপের হোটেল বে আইল্যান্ড (☎ ৩৩৬৮১) ও পিয়ারলেস রিসর্ট (☎ ৩৩৪৮২), সিনক্লেয়ার্স বে ভিউ (সাউথ পয়েন্ট, ☎ ৩-২৭৭৩) ও হোটেল শম্পেন (মিডল পয়েন্ট, ☎ ৩২৩৪৮)—এ ধরনটা ৬০০-১৪০০ টাকার মধ্যে। হোটেল অভিষেক (গোলঘর, ☎ ৩০৫৮৫), হোটেল এন কে ইন্টারন্যাশনাল (সি পোর্ট রোড, ☎ ৩৩০৩৬), হোটেল শালিমার (নেপালিপুর ☎ ৩৩৯৫৩) ইত্যাদি হোটেল ঘর মেলে ১২০-৪০০ টাকায়। আবেরডীন বাজারের কাছে আছে বেশ কিছু হোটেল।

হোটেল বন্দরপ্রী তালের সেরা। দ্বিশয্যার ৬০০ টাকা। ☎ ৩৩৯৫৩। সাগর লজ, মনোহর হল, রতন লজে ২৫-৫০ টাকায়ও ঘর মেলে। এছাড়া আছে হলিডে রিসর্ট (☎ ৩০২১৬), হোটেল শা অ্যান্ড শা (☎ ৩২৩৬১), জগন্নাথ গেস্ট হাউস (☎ ৩৩১৪০) ইত্যাদি।

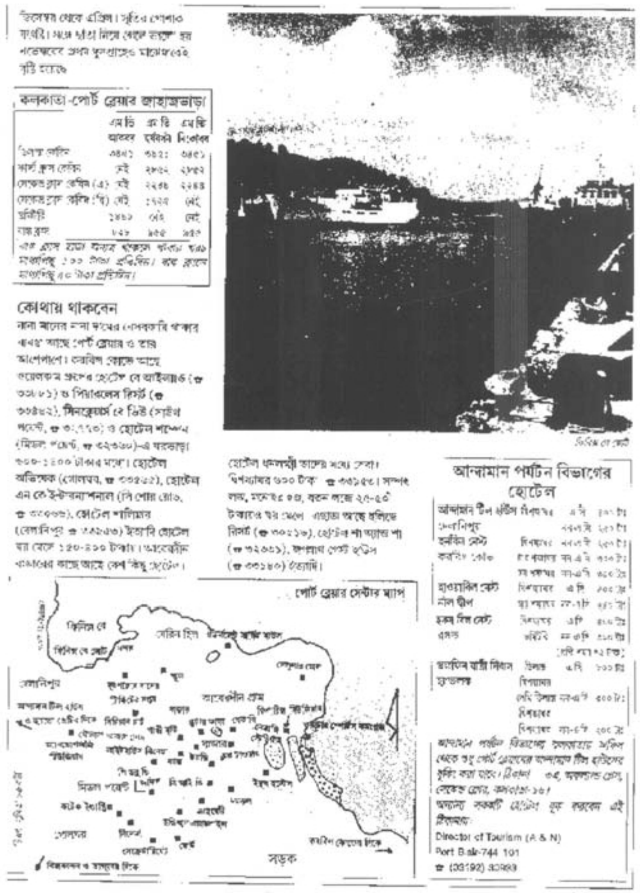
আন্দামান পর্যটন বিভাগের হোটেল

আন্দামান টিল হাউজ	দ্বিশয্যার	এসি	৫০০ টা
নেপালিপুর		নন-এসি	২৫০ টা
হর্ণবিল নেস্ট	দ্বিশয্যার	নন-এসি	২৫০ টা
করবিন কোভ	২য় শয্যার	নন-এসি	৩০০ টা
	৪র্থ শয্যার	নন-এসি	৩০০ টা
হ্যাণ্ডাবিল নেস্ট	দ্বিশয্যার	এ সি	৪০০ টা
নীল ঊীপ	৩য় শয্যার	নন-এসি	২৫০ টা
ডরম বিল্ড নেস্ট	দ্বিশয্যার	এসি	৪০০ টা
ধরমশালা	প্রতিটি	নন-এসি	৫০০ টা
			(প্রতি শয্যা ১২ টাকা)
ফলবিন ঘাটী নিবাস	ডিলাক্স	এ সি	৮০০ টা
হ্যাডডাক	দ্বিশয্যার		
	সেট রিলাক্স	নন-এসি	৩০০ টা
	দ্বিশয্যার		
	দ্বিশয্যার	নন-এসি	২০০ টা

আন্দামান পর্যটন বিভাগের কলকাতার অফিস থেকে শুধু পোর্ট ব্লেয়ারের আন্দামান টিল হাউজের বুকিং করা যাবে। ঠিকানা— ৩৬, জওহরলাল নেহরু রোড, কলকাতা-১৬।

অন্যান্য সরকারি হোটেল বুক করবেন এই ঠিকানায়:
Director of Tourism (A & N)
Port Blair-744 101
☎ (03192) 30333

Example (contd.)



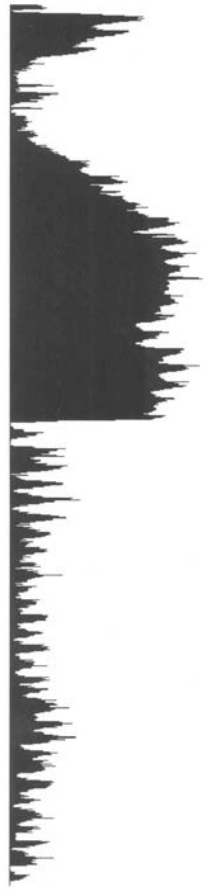
কলকাতা-পোর্ট ব্লেয়ার জাহাজভাড়া

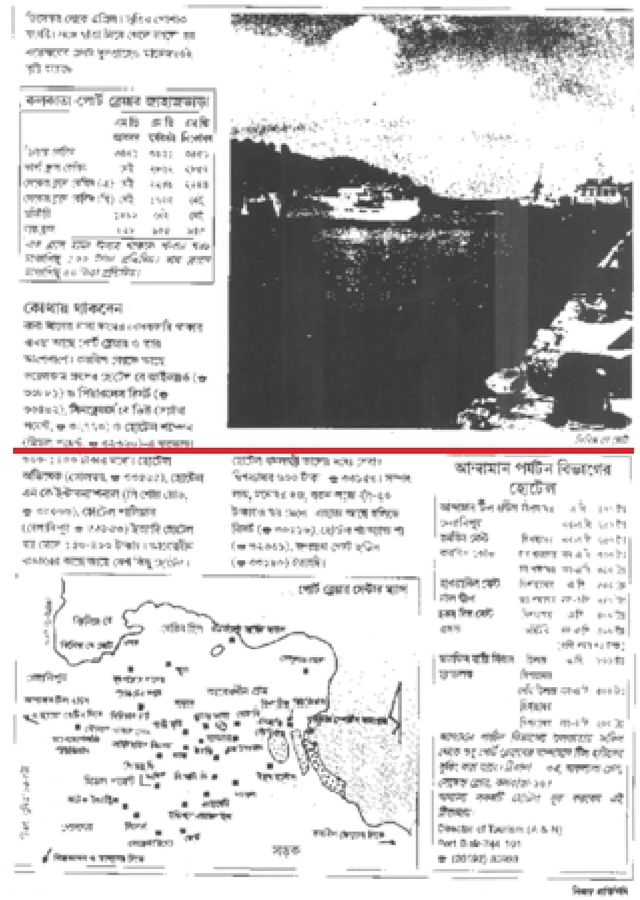
কোথায় থাকবেন

আন্দামান পর্যটন বিভাগের হোটেল

পোর্ট ব্লেয়ার সেন্টার ম্যাপ

Director of Tourism (A & N)
Port Blair-744 101





কলকাতা-পোর্ট ব্লেয়ার জাহাজভাড়া

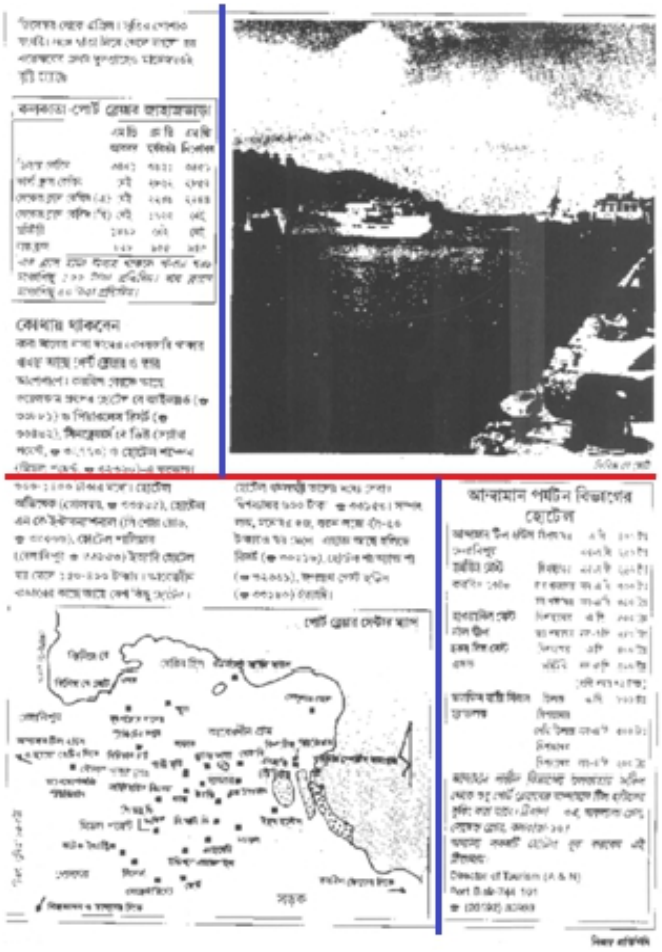
কোথায় থাকবেন

আন্দামান পর্যটন বিভাগের হোটেল

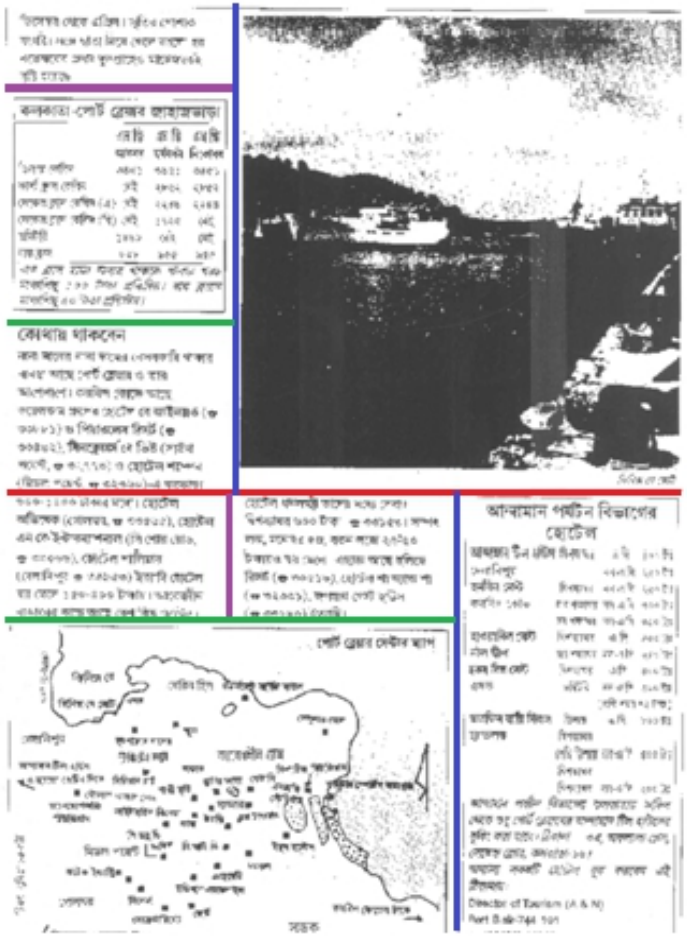
পোর্ট ব্লেয়ার সেন্টার ম্যাপ

Director of Tourism (A & N)
Port Blair-744 101

Example (contd.)



Example (contd.)



Problems of Projection method

- Cannot say what each block contains until further analysis.
  - Extract ***features*** from a zone
  - Recognize the zone content using a ***classifier***
- Results are highly dependent even on small skew in the scanned page.

Mathematical Morphology

- Mathematical morphological operators are good choice.

Objects

- All characters, figures, drawing, i.e., black components against white background

Structuring element

- Regular geometric figures:
  - mostly *line segment, square, circle,* etc.

Morphological Operations

Set theoretic operations (including union, intersection, etc.):

1. Dilation

2. Erosion

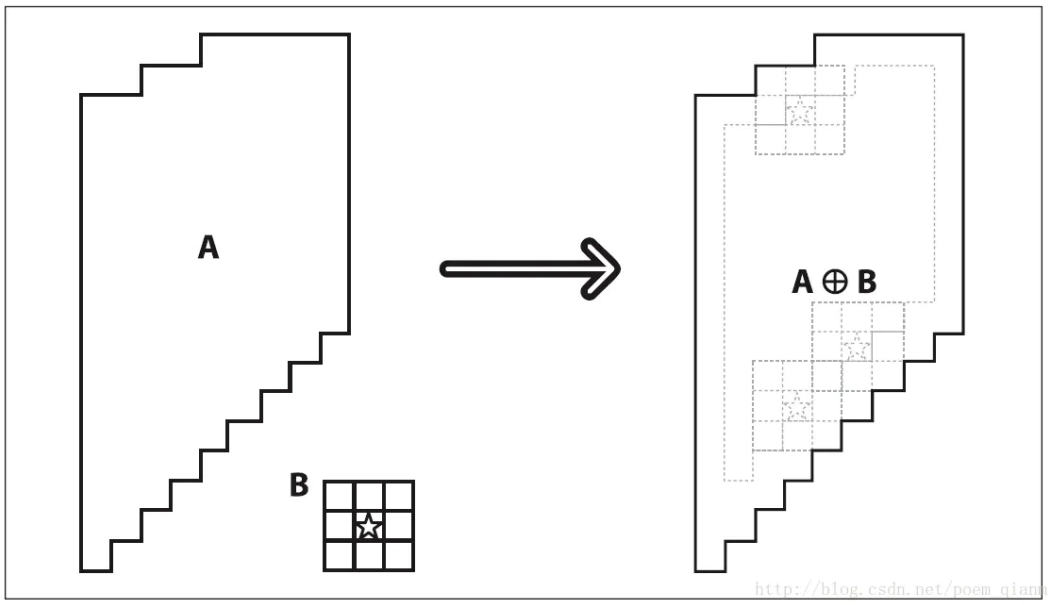
3. Opening

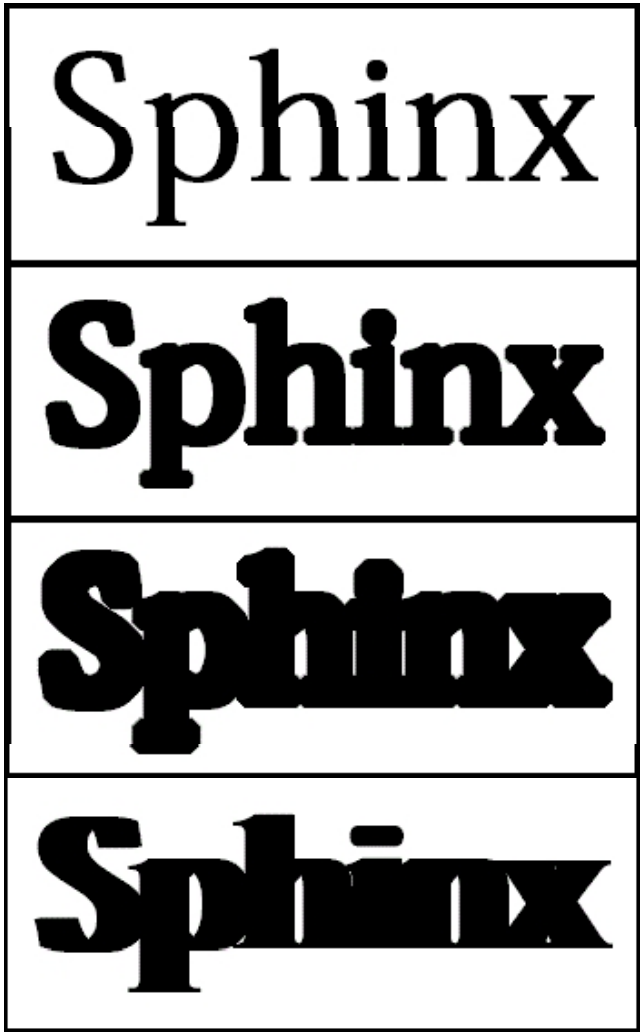
4. Closing

Morphological operator: Dilation

- Expands the objects.

$$A \oplus B = \{a + b \mid a \in A, b \in B\}$$





Orig.

SE:

Circ-5

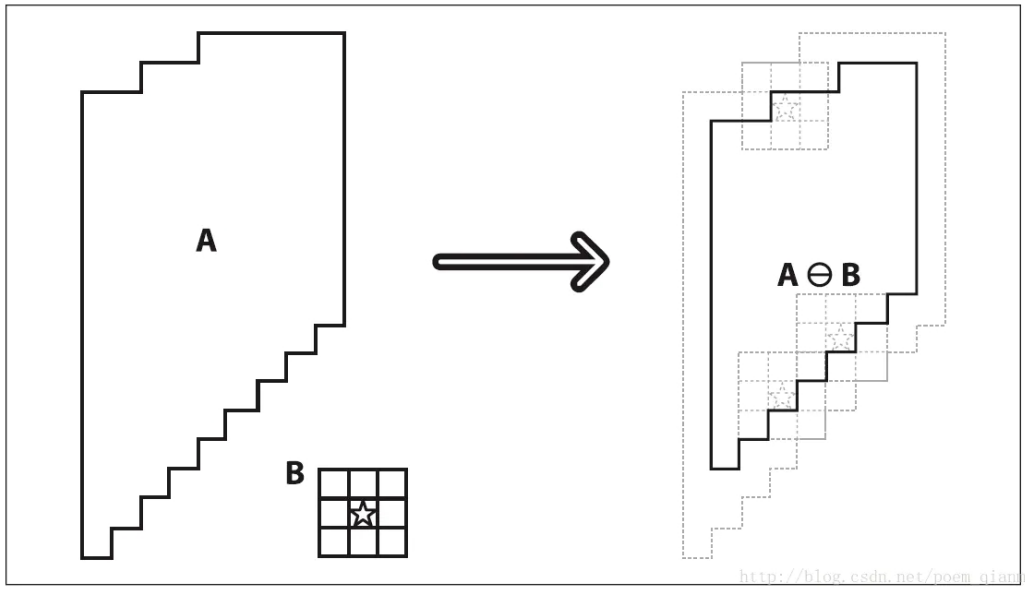
Circ-9

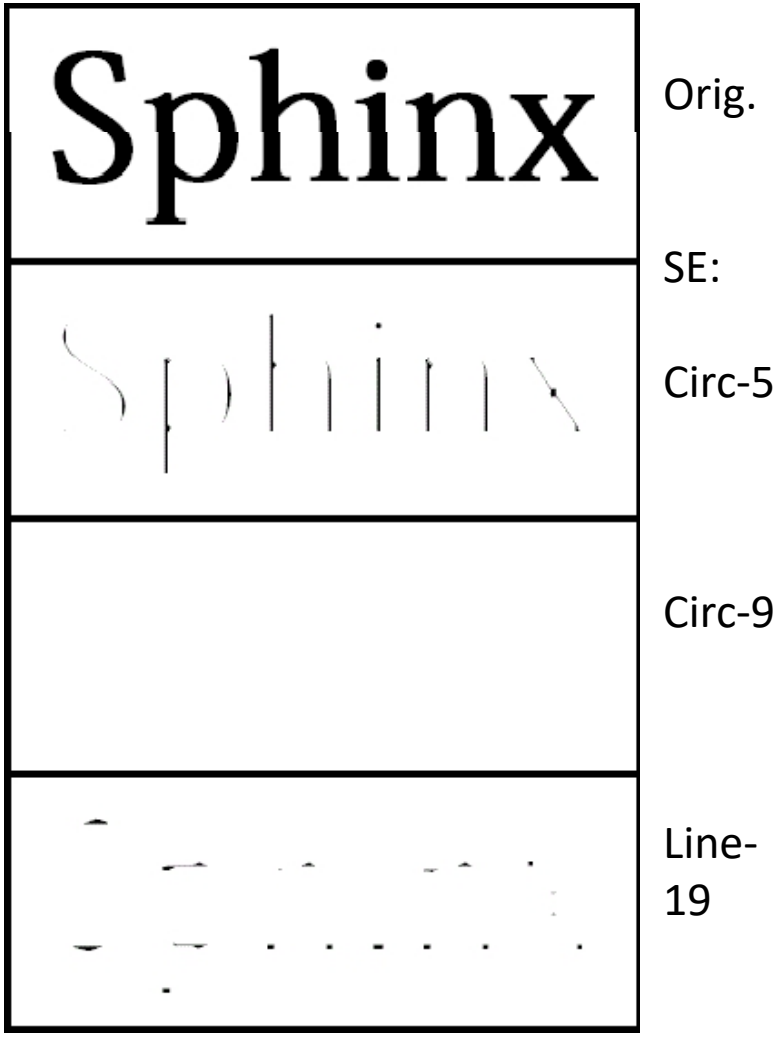
Line-19

Morphological operator: Erosion

- Shrinks the objects.

$$A \Theta B = \{p \mid B + p \in A\}$$





Morphological operator: Opening

- Removes objects or parts of it that cannot fit in SE.

$$A \circ B = (A \Theta B) \oplus B$$

where A is an object and

B is SE.

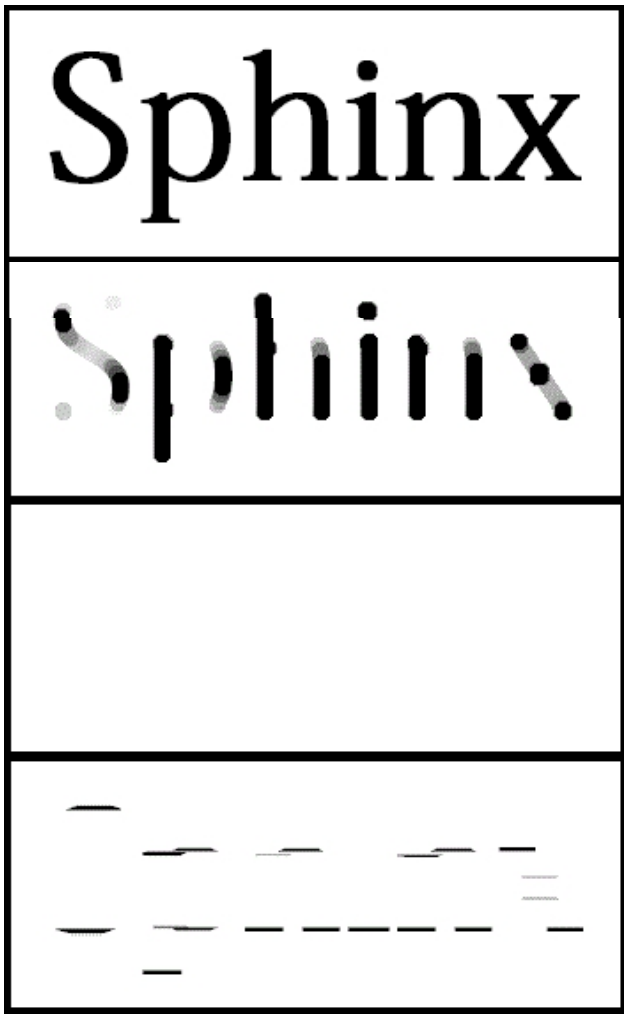
- **Properties:**

  Increasing,

  idempotent,

  anti-extensive.

- *It is a filter.*



Morphological operator: Closing

- Appends to objects parts of background if SE does not fit.

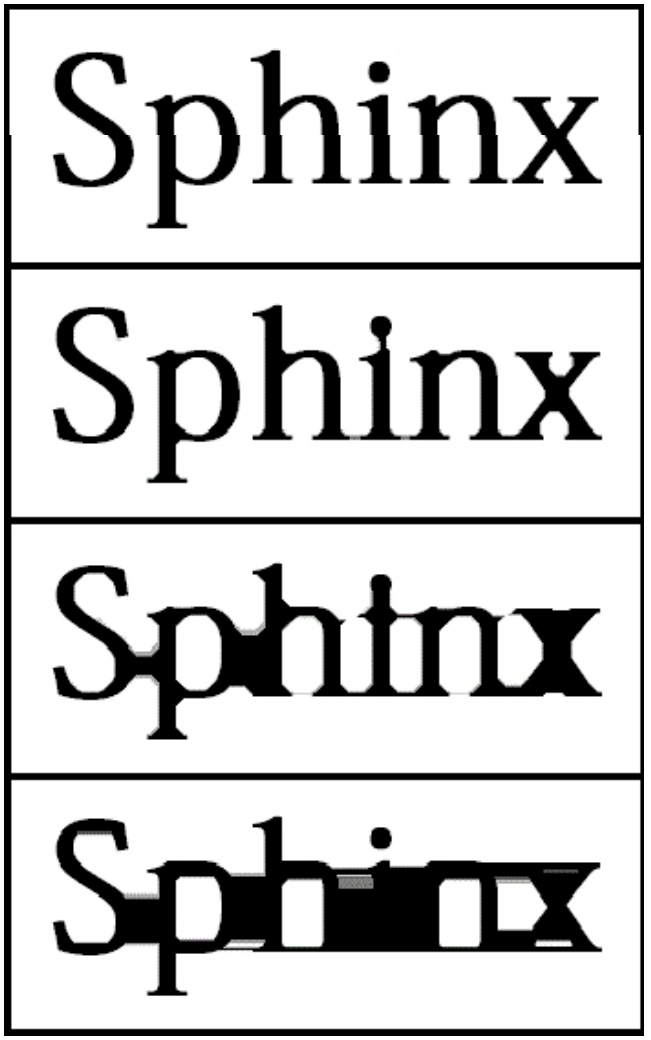
$$A \bullet B = (A \oplus B) \Theta B$$

where A is an object and

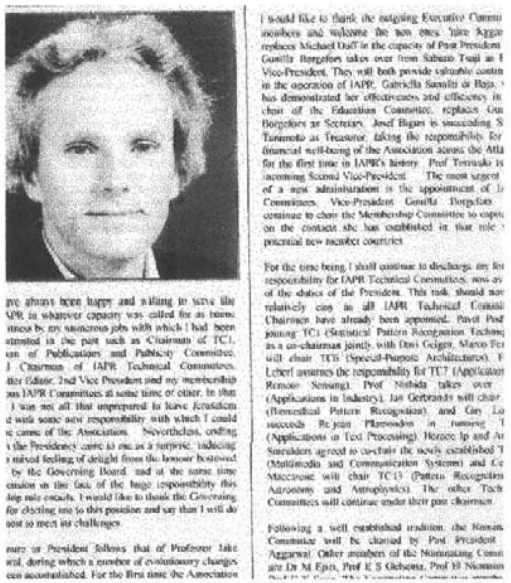
B is SE.

- **Properties:**

  Increasing, idempotent, and extensive.
- It is a filter.
- *Opening & closing are dual.*

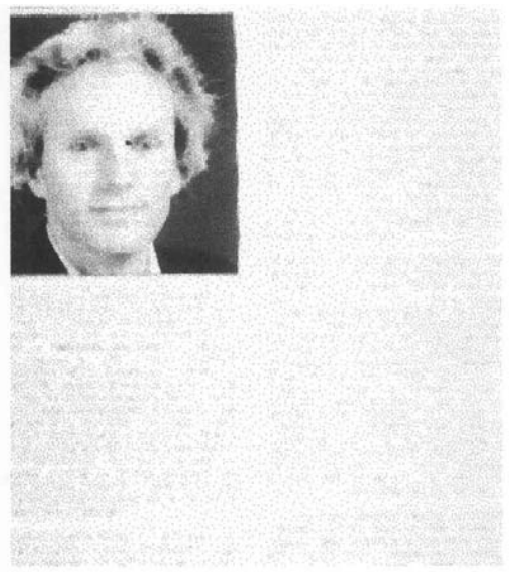


Morphological approach: An example



(a) Original image

(b) Closed image



(c) Opened image

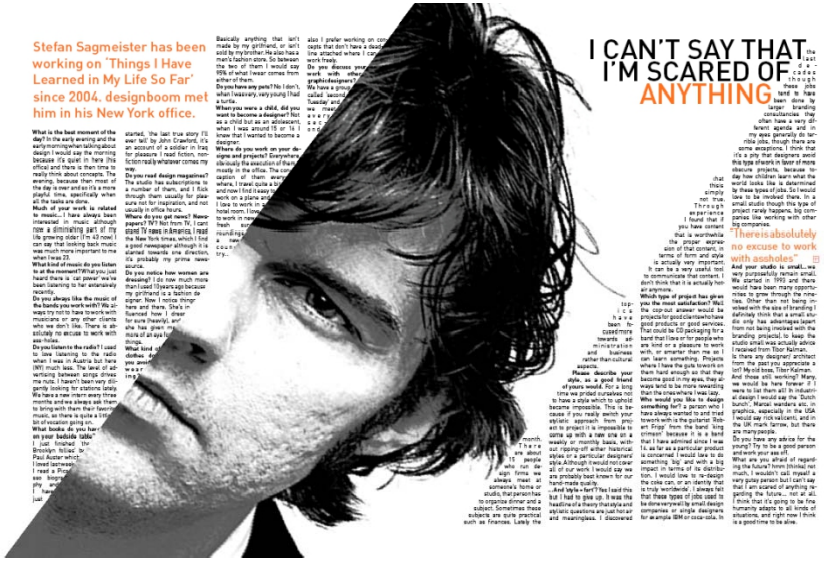
Issues in document page scanning

- Resolution
- Back page impression[封底压痕]
- Granular noise
- Blotted text *(specially in old documents)*
- Bending of pages at the binding
- Skew

  (due to placement of the page in the scanner)

Results

Input test image



Resultant (labeled) image

Results

Input test image

Resultant (labeled) image

Results

Input test image

Resultant (labeled) image

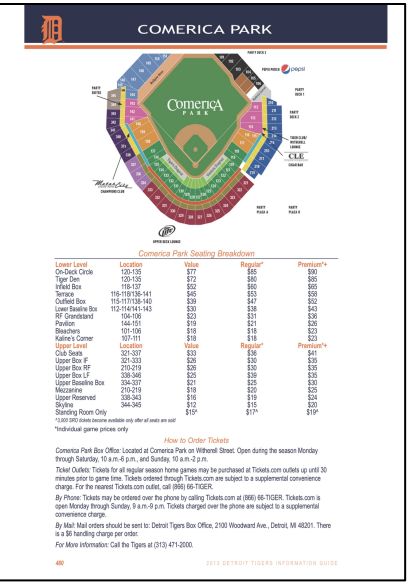
Results

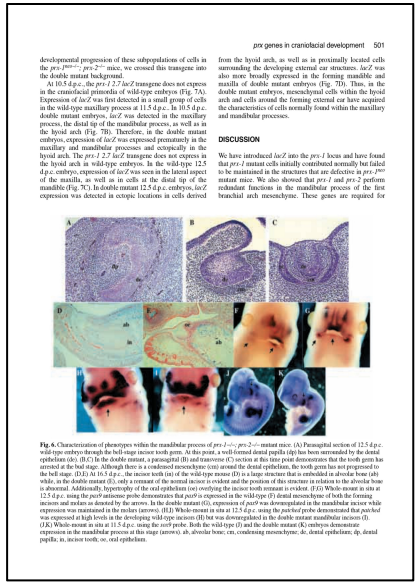
Input test image

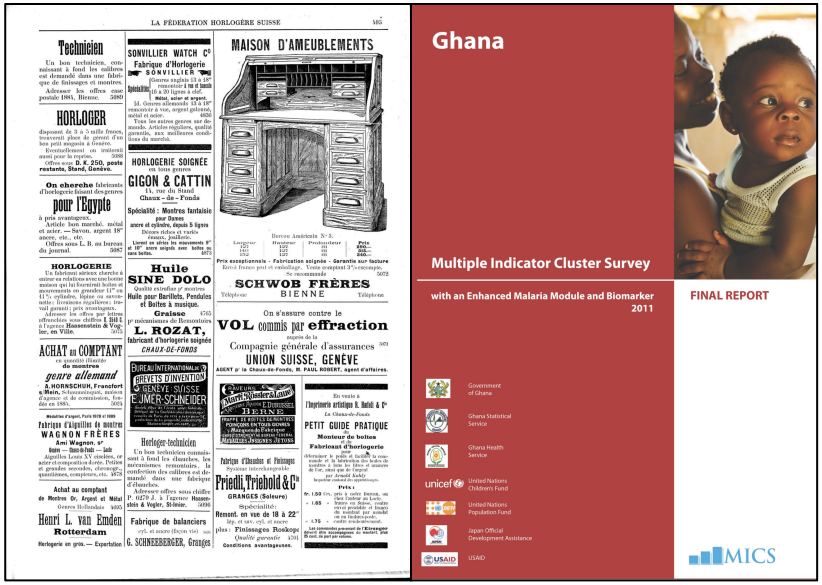
Resultant (labeled) image

Benchmark database

- DSSE-200 [1]







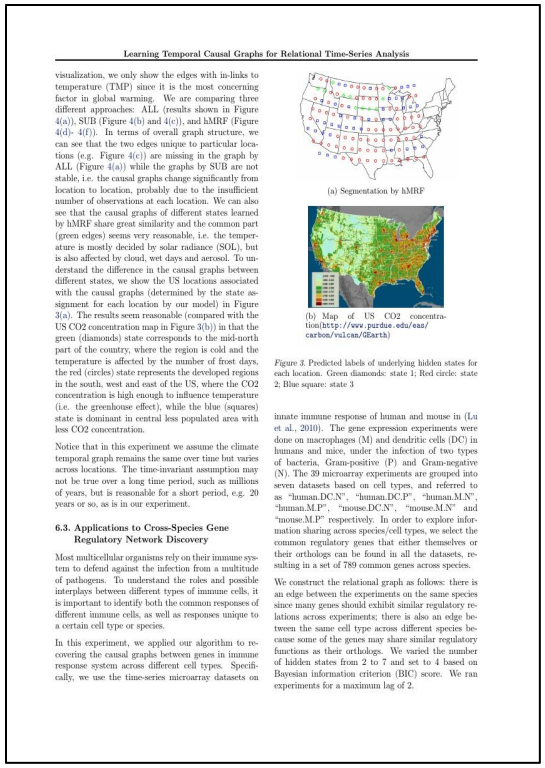
Summary

- High-speed with level, smooth ride
- Shallow channel & port accessibility
- Low wake-wash & wave creation
- High reliability, safety, survivability
- Favorable economics
- Can be built now and perform well
- Planned program for further improvement

[1]X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. Lee Giles. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks.In Conference on Computer Vision and Pattern Recognition (CVPR). 5315–5324.

Benchmark database

- CS-150 [1]



Learning Temporal Causal Graphs for Relational Time-Series Analysis

visualization, we only show the edges with in-links to temperature (TMP) since it is the most concerning factor in global warming. We are comparing three different approaches: ALL (results shown in Figure 4(a)), SUB (Figure 4(b) and 4(c)), and hMRF (Figure 4(d)- 4(f)). In terms of overall graph structure, we can see that the two edges unique to particular locations (e.g. Figure 4(c)) are missing in the graph by ALL (Figure 4(a)) while the graphs by SUB are not stable, i.e. the causal graphs change significantly from location to location, probably due to the insufficient number of observations at each location. We can also see that the causal graphs of different states learned by hMRF share great similarity and the common part (green edges) seems very reasonable, i.e. the temperature is mostly decided by solar radiance (SOL), but is also affected by cloud, wet days and aerosol. To understand the difference in the causal graphs between different states, we show the US locations associated with the causal graphs (determined by the state assignment for each location by our model) in Figure 3(a). The results seem reasonable (compared with the US CO2 concentration map in Figure 3(b)) in that the green (diamonds) state corresponds to the mid-north part of the country, where the region is cold and the temperature is affected by the number of frost days, the red (circles) state represents the developed regions in the south, west and east of the US, where the CO2 concentration is high enough to influence temperature (i.e. the greenhouse effect), while the blue (squares) state is dominant in central less populated area with less CO2 concentration.

Notice that in this experiment we assume the climate temporal graph remains the same over time but varies across locations. The time-invariant assumption may not be true over a long time period, such as millions of years, but is reasonable for a short period, e.g. 20 years or so, as is in our experiment.

6.3. Applications to Cross-Species Gene Regulatory Network Discovery

Most multicellular organisms rely on their immune system to defend against the infection from a multitude of pathogens. To understand the roles and possible interplays between different types of immune cells, it is important to identify both the common responses of different immune cells, as well as responses unique to a certain cell type or species.

In this experiment, we applied our algorithm to recovering the causal graphs between genes in immune response system across different cell types. Specifically, we use the time-series microarray datasets on

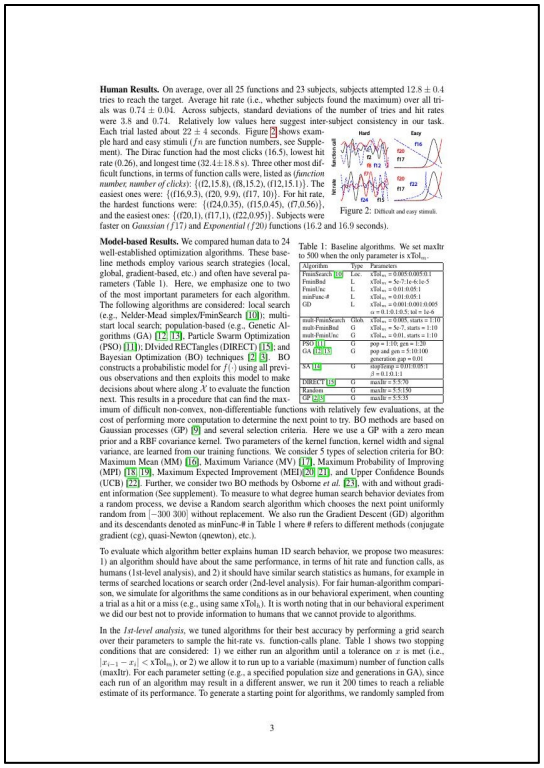
(a) Segmentation by hMRF

(b) Map of US CO2 concentration(http://www.purdue.edu/eas/carbon/vulcan/GEarth)

Figure 3. Predicted labels of underlying hidden states for each location. Green diamonds: state 1; Red circle: state 2; Blue square: state 3

innate immune response of human and mouse in (Lu et al., 2010). The gene expression experiments were done on macrophages (M) and dendritic cells (DC) in humans and mice, under the infection of two types of bacteria, Gram-positive (P) and Gram-negative (N). The 39 microarray experiments are grouped into seven datasets based on cell types, and referred to as "human.DC.N", "human.DC.P", "human.M.N", "human.M.P", "mouse.DC.N", "mouse.M.N" and "mouse.M.P" respectively. In order to explore information sharing across species/cell types, we select the common regulatory genes that either themselves or their orthologs can be found in all the datasets, resulting in a set of 789 common genes across species.

We construct the relational graph as follows: there is an edge between the experiments on the same species since many genes should exhibit similar regulatory relations across experiments; there is also an edge between the same cell type across different species because some of the genes may share similar regulatory functions as their orthologs. We varied the number of hidden states from 2 to 7 and set to 4 based on Bayesian information criterion (BIC) score. We ran experiments for a maximum lag of 2.



Human Results. On average, over all 25 functions and 23 subjects, subjects attempted 12.8 ± 0.4 tries to reach the target. Average hit rate (i.e., whether subjects found the maximum) over all trials was 0.74 ± 0.04. Across subjects, standard deviations of the number of tries and hit rates were 3.8 and 0.74. Relatively low values here suggest inter-subject consistency in our task. Each trial lasted about 22 ± 4 seconds. Figure 2 shows example hard and easy stimuli (f*n* are function numbers, see Supplement). The Dirac function had the most clicks (16.5), lowest hit rate (0.26), and longest time (32.4 ± 18.8 s). Three other most difficult functions, in terms of function calls were, listed as (*function number, number of clicks*): {(f2,15.8), (f8,15.2), (f12,15.1)}. The easiest ones were: {(f16,9.3), (f20, 9.9), (f17, 10)}. For hit rate, the hardest functions were: {(f24,0.35), (f15,0.45), (f7,0.56)}, and the easiest ones: {(f20,1), (f17,1), (f22,0.95)}. Subjects were faster on *Gaussian* (*f*17) and *Exponential* (*f*20) functions (16.2 and 16.9 seconds).

Figure 2: Difficult and easy stimuli.

Model-based Results. We compared human data to 24 well-established optimization algorithms. These baseline methods employ various search strategies (local, global, gradient-based, etc.) and often have several parameters (Table 1). Here, we emphasize one to two of the most important parameters for each algorithm. The following algorithms are considered: local search (e.g., Nelder-Mead simplex/FminSearch [10]); multi-start local search; population-based (e.g., Genetic Algorithms (GA) [12, 13], Particle Swarm Optimization (PSO) [11]); DIvided RECTangles (DIRECT) [15]; and Bayesian Optimization (BO) techniques [2, 3]. BO constructs a probabilistic model for $f(\cdot)$ using all previous observations and then exploits this model to make decisions about where along $\mathcal{X}$ to evaluate the function next. This results in a procedure that can find the maximum of difficult non-convex, non-differentiable functions with relatively few evaluations, at the cost of performing more computation to determine the next point to try. BO methods are based on Gaussian processes (GP) [9] and several selection criteria. Here we use a GP with a zero mean prior and a RBF covariance kernel. Two parameters of the kernel function, kernel width and signal variance, are learned from our training functions. We consider 5 types of selection criteria for BO: Maximum Mean (MM) [16], Maximum Variance (MV) [17], Maximum Probability of Improving (MPI) [18, 19], Maximum Expected Improvement (MEI)[20, 21], and Upper Confidence Bounds (UCB) [22]. Further, we consider two BO methods by Osborne *et al.* [23], with and without gradient information (See supplement). To measure to what degree human search behavior deviates from a random process, we devise a Random search algorithm which chooses the next point uniformly random from $[-300\ 300]$ without replacement. We also run the Gradient Descent (GD) algorithm and its descendants denoted as minFunc-# in Table 1 where # refers to different methods (conjugate gradient (cg), quasi-Newton (qnewton), etc.).

Table 1: Baseline algorithms. We set maxItr to 500 when the only parameter is $xTol_{rs}$.

Algorithm	Type	Parameters
FminSearch [10]	Loc.	$xTol_{rs}$ = 0.005 0.005 0.1
FminBnd	L.	$xTol_{rs}$ = 5e-7 1e-6 1e-5
FminUnc	L.	$xTol_{rs}$ = 0.01 0.05 1
minFunc-#	L.	$xTol_{rs}$ = 0.01 0.05 1
GD	L.	$xTol_{rs}$ = 0.001 0.001 0.005 α = 0.1 0.1 0.5 $\eta\delta$ = 1e-6
multi-FminSearch	Glob.	$xTol_{rs}$ = 0.005, starts = 1:10
multi-FminBnd	G	$xTol_{rs}$ = 5e-7, starts = 1:10
multi-FminUnc	G	$xTol_{rs}$ = 0.01, starts = 1:10
PSO [11]	G	pop = 1:10; gen = 1:20
GA [12,13]	G	pop and gen = 5:10:100; generation gap = 0.01
SA [14]	G	startpt = 0.01 0.05 1; β = 0.1:0.1:1
DIRECT [15]	G	maxIlr = 3:5:30
Random	G	maxIlr = 5:5:150
CGP [2,3]	G	maxIlr = 5:5:35

To evaluate which algorithm better explains human 1D search behavior, we propose two measures: 1) an algorithm should have about the same performance, in terms of hit rate and function calls, as humans (1st-level analysis), and 2) it should have similar search statistics as humans, for example in terms of searched locations or search order (2nd-level analysis). For fair human-algorithm comparison, we simulate for algorithms the same conditions as in our behavioral experiment, when counting a trial as a hit or a miss (e.g., using same $xTol_{rs}$). It is worth noting that in our behavioral experiment we did our best not to provide information to humans that we cannot provide to algorithms.

In the *1st-level analysis*, we tuned algorithms for their best accuracy by performing a grid search over their parameters to sample the hit-rate vs. function-calls plane. Table 1 shows two stopping conditions that are considered: 1) we either run an algorithm until a tolerance on x is met (i.e., $|x_{i-1} - x_i| < xTol_{rs}$) or 2) we allow it to run up to a variable (maximum) number of function calls (maxItr). For each parameter setting (e.g., a specified population size and generations in GA), since each run of an algorithm may result in a different answer, we run it 200 times to reach a reliable estimate of its performance. To generate a starting point for algorithms, we randomly sampled from

3

C. A. Clark and S. Divvala. 2015. Looking beyond text: Extracting figures, tables and captions from computer science papers. In The AAAI Conference on Artificial Intelligence Workshops.

Benchmark database

- ICDAR2015

Four Tech Waves To Watch

1 UTILITY COMPUTING

JUST TURN ON THE DATA

The idea is to make computing power into another pay-as-you-go service—like water or electricity. But beware of the hype

By Steve Hamm

ON SCIENCE
Sharon Begley

Our Imaginary, Hotter Selves

ANYONE WHO HAS EVER HAD A BAD HAIR DAY, WHEN LOOKING LIKE A latter-day Medusa makes you feel cranky and antisocial and plodding, can sympathize with the Oakland Raiders—and not because the players get helmet hair. The Raiders alternated between mostly black and mostly white uniforms, depending on whether they were playing at home or away. Knowing that appearance affects people's mood and outlook, psychologists wondered whether uniform color influenced the Raiders' aggressiveness. Using data from the 1970s and 1980s, they found that the team racked up way more penalty yards—a measure of aggression—when they wore black than when they wore white, for infractions both minor (encroachment) and major (roughing the kicker). The pattern held even when the scientists took into account different conditions and styles of play at home and away. But while the 1988 finding has become a classic in psychology, the explanation remains controversial. Do referees, because of black's cultural baggage, see black-clad players as meaner and badder than those in, say, baby blue? Or does wearing black make players see themselves as tougher and meaner—and therefore cause them to play that way?

Jeremy Bailenson and Nick Yee of Stanford University had this and other classic studies in mind when they started wondering about the effect of being able to alter one's appearance. They weren't going to study wardrobe choices, however. Their quarry is avatars, digital representations of players in such games as Second Life. "Your physical appearance changes how people treat you," says Bailenson. "But independent of that, when you perceive yourself in a certain way, you act differently." He and Yee call it "the Proteus effect," after the shape-changing Greek god. The effect of appearance on behavior, they find, carries over from the virtual world to the real one, with intriguing consequences.

In one Stanford study, volunteers were assigned avatars who ranged from attractive to plain. It is one of life's inequities that the world sees attractive people as possessing a long list of desirable traits, including honesty, generosity and kindness. Perhaps as a result, people judged attractive are more self-confident than ugly ducklings, and so tend to be extroverted. Using a virtual-reality headset, the volunteers—actually, their avatars—walked across a room to interact with another avatar. Those with attractive avatars got within three feet of the stranger; those with homely ones kept almost six feet away. How much "personal space" one needs is inversely proportional to self-confidence, which having an attractive avatar increases. When the stranger asked the players to "tell me a little about yourself," good-looking avatars revealed more: feeling attractive increases self-esteem and therefore friendliness.

The Proteus effect spilled into the real world. After their virtual-reality session, players were shown photos from an online dating site and asked to pick those who "would be interested in you." Players who had been assigned attractive avatars picked more-attractive candidates than did players (of equal pulchritude in real life) who had been represented as homely avatars. Male players were also asked to enter personal information for an online dating site. In this situation men routinely inflate their height by an average of one inch. But those who had had an attractive avatar told the truth.

Western society sees taller people as more competent and having greater leadership potential than shorter people. The effect of virtual height, too, bleeds into real life, the Stanford scientists find. After their avatar roamed through a virtual world, players took seats in the real world to play a split-the-loot game. Player One proposed ways to divide $100 which Player Two could accept or reject, in which case neither player got anything. People tend to reject

Avatars might serve therapeutic purposes, helping those with social phobia become more confident.

unfair offers, even though accepting a lopsided $99/$1 split leaves them objectively better off than walking away in a righteous huff. Players fresh from being a tall avatar showed their cockiness, proposing, on average, $61/$39 splits. People with short avatars averaged offers of $52/$48. Yee and Bailenson reported last year. When it was their turn to weigh an insulting $75/$25 offer, players with tall avatars rejected it 62 percent of the time; those with short avatars stood up for themselves a wimpish 28 percent of the time.

The goal of the research isn't to tell Second Lifers they'll have more virtual friends, money, power and other goodies if they create a hot, young, powerful avatar. Most do so anyway. Instead, avatars might serve therapeutic purposes, helping those with social phobia, say, become more confident and friendly in real life. The work also underlines the power of new media to affect our behavior: players who roamed a virtual world as a KKK-clad avatar felt more aggressive than they did before playing the game, while those whose avatar wore a doctor's coat scored higher on a test of friendliness. It's not clear how long the spillover to the real world lasts. But even if it's only a few hours the potential is impressive: online players spend, on average, 20 hours a week as their avatar.

Briefing

Verbatim

'Nobody is going to be allowed to do anything here.'

WAHEED ARSHAD, Pakistani Major General, on reports that the U.S. may expand military and CIA authority to pursue Islamic militants in Pakistan

'If you are a believer in miracles, this would be one.'

PHILIP S. BARIE, chief of the division of critical care at New York–Presbyterian Hospital, on the recuperation of Alcides Moreno, a 37-year-old window washer who fell 47 stories last month

'We should not resort to violence even if we have differences.'

MAUMOON GAYOOM, President of the Maldives, after an attempt to assassinate him was foiled when a boy scout waiting to greet the island nation's leader grabbed the attacker's knife as he lunged out of a crowd

'This is who I am. This is my life.'

SERGEANT DARREN MANZELLA, U.S. Army medic and a decorated veteran of Iraq, discussing his homosexuality in a Jan. 8 press conference. Manzella has been openly gay since August 2006, despite the military's "Don't ask, don't tell" policy; its critics have argued that his case proves the policy is arbitrarily enforced

'They should either kill me or organize a second round of elections.'

LEVAN GACHECHILADZE, Georgian opposition leader, after his resounding defeat by pro-Western incumbent Mikhail Saakashvili in the country's Jan. 5 Presidential elections. Gachechiladze claimed the voting was rigged

'This is the first red carpet that I've really walked down where I didn't have to think about holding in my stomach.'

HALLE BERRY, on being pregnant, while accepting the Desert Palm Achievement Award at the Palm Springs International Film Festival on Jan. 5

MORE AT TIME.COM
For more daily sound bites, visit time.com/quotes

NUMBERS

CENSORSHIP
16,600%
Percentage by which Burmese authorities have increased the license fee to receive satellite TV, the country's only broadcast-news source not controlled by the state

300%
Percentage by which the new fee—about $800—surpasses the annual salary of a Burmese schoolteacher. The increase is seen as a move to block access to news critical of the ruling military junta

HEALTH
65%
Percentage of U.S. adults who are exercising, dieting or both during any nonholiday period

29
Number of pounds (13 kg) that at least 25% of American adults currently on a diet would like to lose. Regular exercisers said they want to drop an average of 14 lbs. (6.3 kg)

GOLD
$884
Record price per ounce of gold set on Jan. 8, as investors reacted to market turbulence and high oil prices, beating the previous all-time record of $875 set in 1980

$2,200
1980's record price for gold, when adjusted for inflation

WEALTH
$46,380
The U.K.'s per capita gross domestic product this year, $500 more than U.S. per capita GDP. For the first time since the 19th century, the average Briton will earn more than the average American

1993
The year following Britain's last major recession, when the country's per capita GDP was 34% less than that of the U.S.

Sources: Reuters, A.P., Pew Research Center (2), International Herald Tribune, A.P., BBC (2)

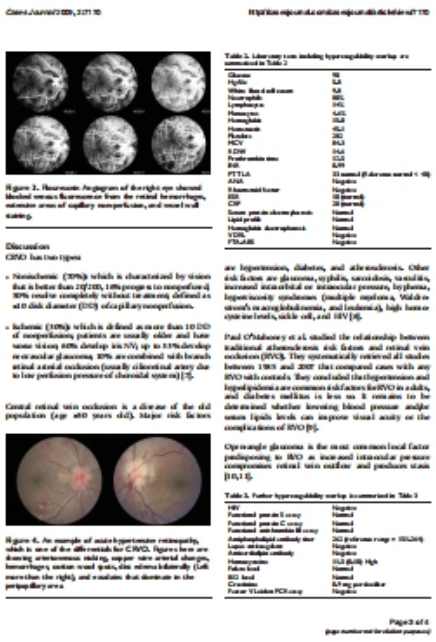
12

TIME January 21, 2008

A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, ICDAR2015 485 competition on recognition of documents with complex layouts-rdcl2015, in: I- 486 APR International Conference on Document Analysis and Recognition (ICDAR), 487 2015, pp. 1151–1155.

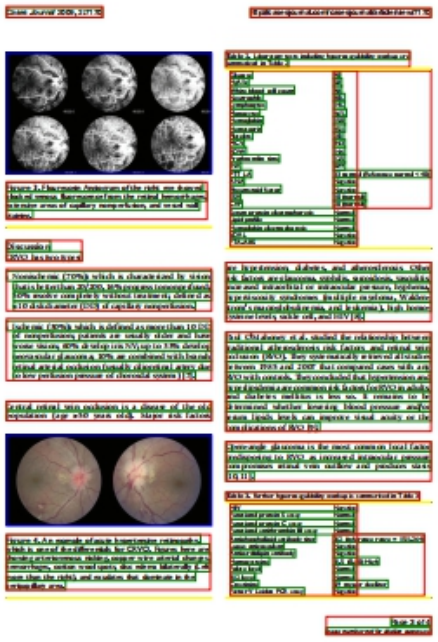
Benchmark database

- PubLayNet

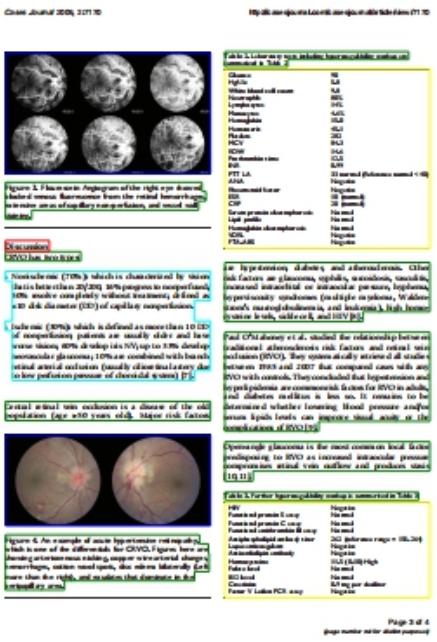


(a) PDF representation.

(b) XML representation.



(c) PDFMiner output on (a).



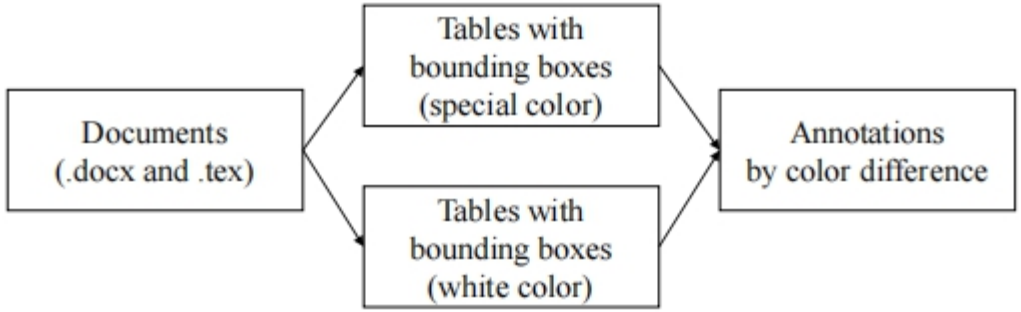
(d) Annotations generated by matching (b) and (c).

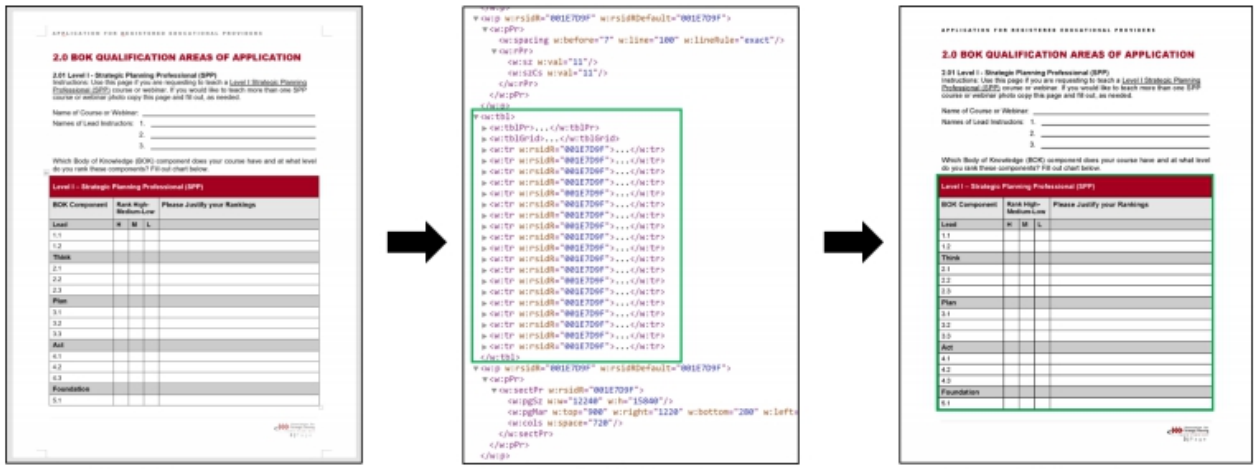
PDFMiner

Zhong X, Tang J, Yepes A J. Publaynet: largest dataset ever for document layout analysis[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 1015-1022.

Benchmark database

- TableBank

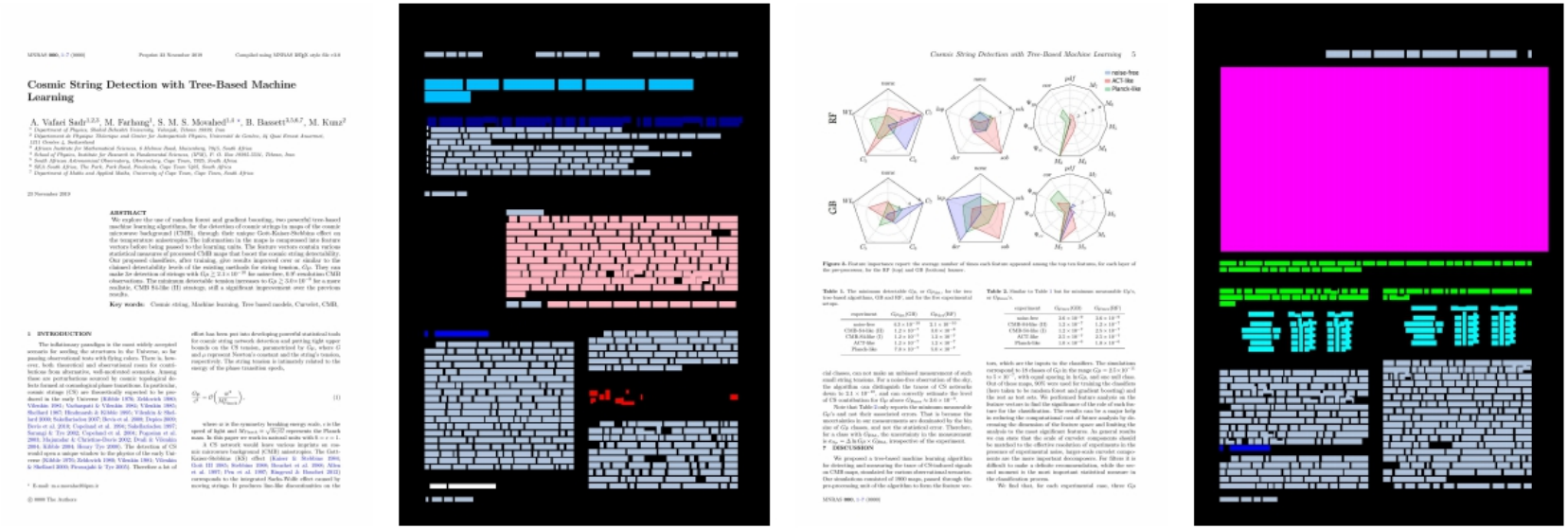




Li M, Cui L, Huang S, et al. Tablebank: Table benchmark for image-based table detection and recognition[C]//Proceedings of The 12th Language Resources and Evaluation Conference. 2020: 1918-1925.

Benchmark database

- DocBank

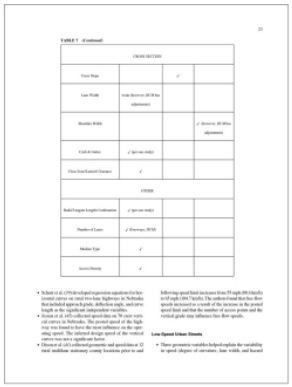


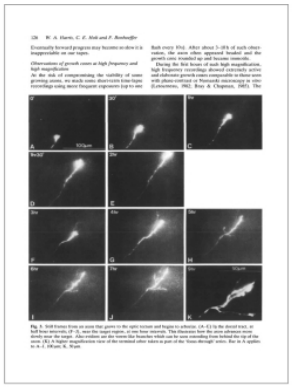
(a) (b) (c) (d)

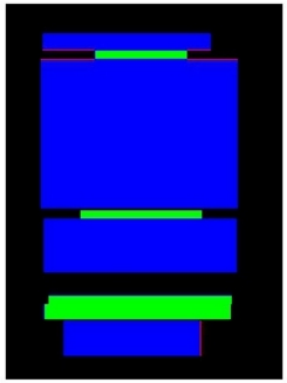
Figure 1: Example annotations of the DocBank. The colors of semantic structure labels are: Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, Title

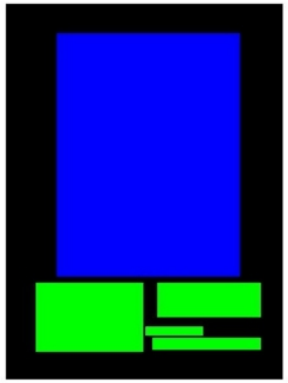
Li M, Xu Y, Cui L, et al. Docbank: A benchmark dataset for document layout analysis[J]. arXiv preprint arXiv:2006.01038, 2020.

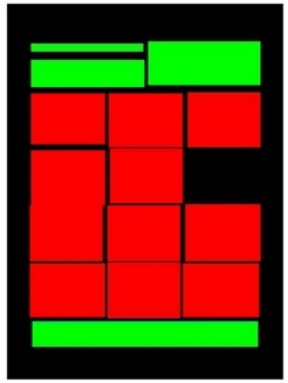
Benchmark database

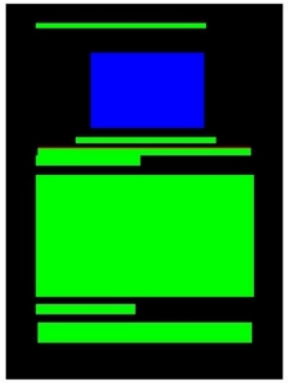


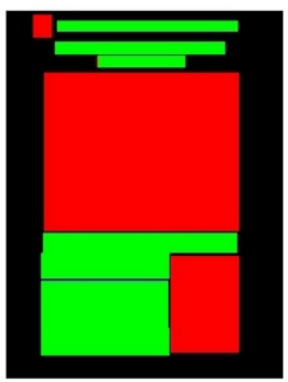


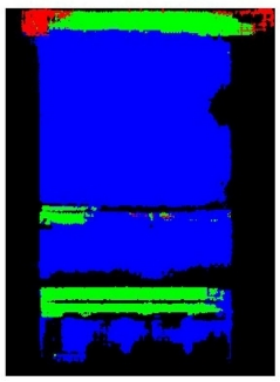


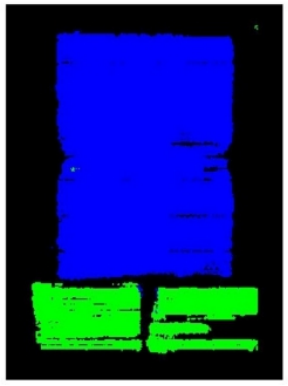


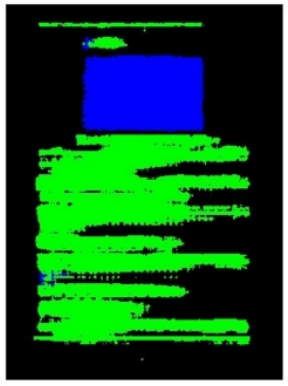


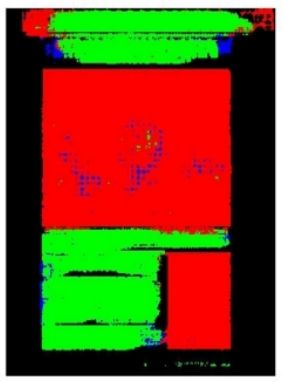


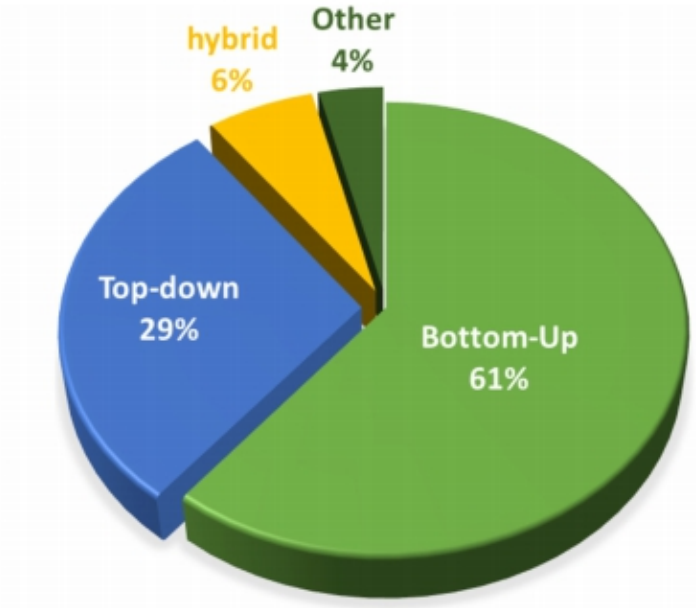




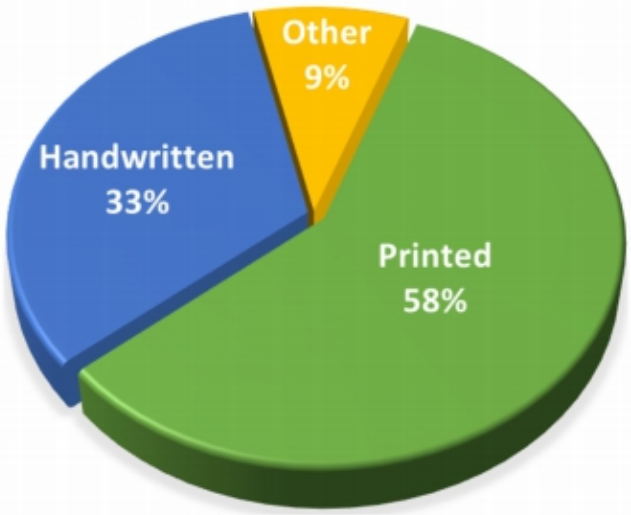




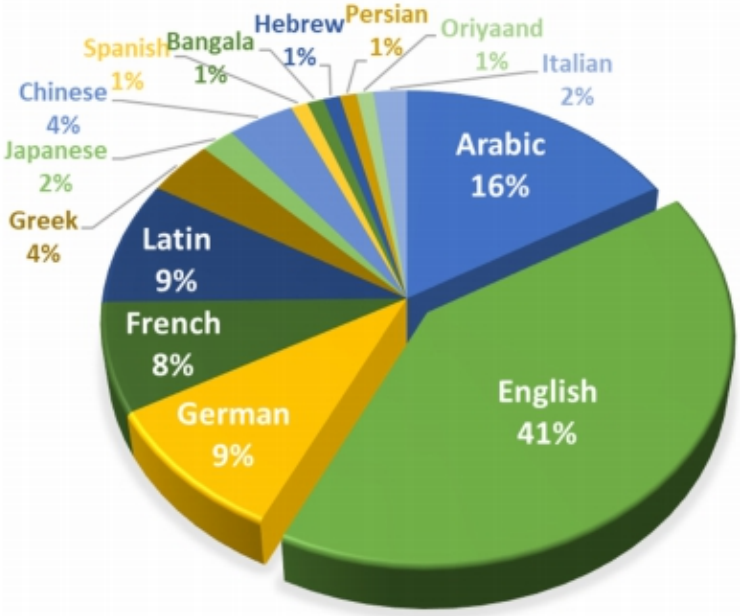




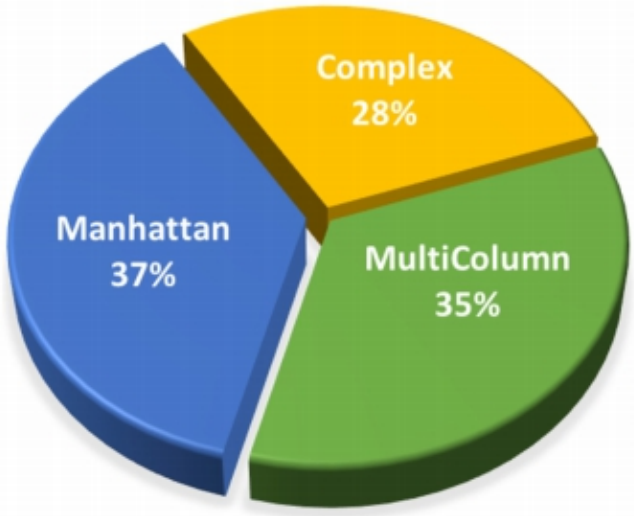
(a) Strategy based distribution



(b) Document Type Distribution



(c) Language based Distribution



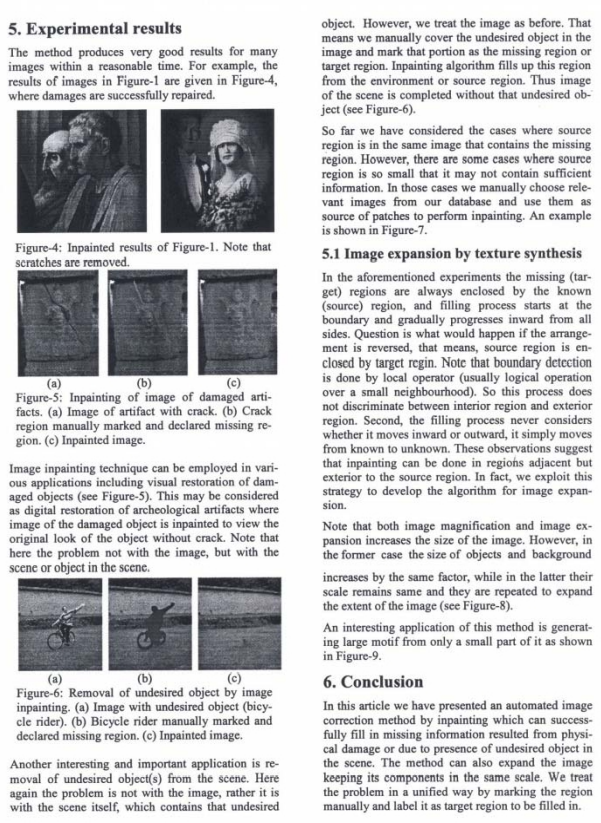
(d) Document Layout Distribution

Deep learning

- Popular technique for unsupervised feature extraction for supervised applications
  - Ex. object recognition.
- Utilizes HUGE number of instances to train relatively simpler system to perform more complicated task.
- Training samples may be outcome of controlled or uncontrolled data acquisition.
- Requires very high computational resources for implementing a reasonably meaningful system.

Detect text area using CNN

Input: *A document image*



5. Experimental results

The method produces very good results for many images within a reasonable time. For example, the results of images in Figure-1 are given in Figure-4, where damages are successfully repaired.

Figure-4: Inpainted results of Figure-1. Note that scratches are removed.

(a) (b) (c)

Figure-5: Inpainting of image of damaged artifacts. (a) Image of artifact with crack. (b) Crack region manually marked and declared missing region. (c) Inpainted image.

Image inpainting technique can be employed in various applications including visual restoration of damaged objects (see Figure-5). This may be considered as digital restoration of archeological artifacts where image of the damaged object is inpainted to view the original look of the object without crack. Note that here the problem not with the image, but with the scene or object in the scene.

(a) (b) (c)

Figure-6: Removal of undesired object by image inpainting. (a) Image with undesired object (bicycle rider). (b) Bicycle rider manually marked and declared missing region. (c) Inpainted image.

Another interesting and important application is removal of undesired object(s) from the scene. Here again the problem is not with the image, rather it is with the scene itself, which contains that undesired object. However, we treat the image as before. That means we manually cover the undesired object in the image and mark that portion as the missing region or target region. Inpainting algorithm fills up this region from the environment or source region. Thus image of the scene is completed without that undesired object (see Figure-6).

So far we have considered the cases where source region is in the same image that contains the missing region. However, there are some cases where source region is so small that it may not contain sufficient information. In those cases we manually choose relevant images from our database and use them as source of patches to perform inpainting. An example is shown in Figure-7.

5.1 Image expansion by texture synthesis

In the aforementioned experiments the missing (target) regions are always enclosed by the known (source) region, and filling process starts at the boundary and gradually progresses inward from all sides. Question is what would happen if the arrangement is reversed, that means, source region is enclosed by target regin. Note that boundary detection is done by local operator (usually logical operation over a small neighbourhood). So this process does not discriminate between interior region and exterior region. Second, the filling process never considers whether it moves inward or outward, it simply moves from known to unknown. These observations suggest that inpainting can be done in regions adjacent but exterior to the source region. In fact, we exploit this strategy to develop the algorithm for image expansion.

Note that both image magnification and image expansion increases the size of the image. However, in the former case the size of objects and background increases by the same factor, while in the latter their scale remains same and they are repeated to expand the extent of the image (see Figure-8).

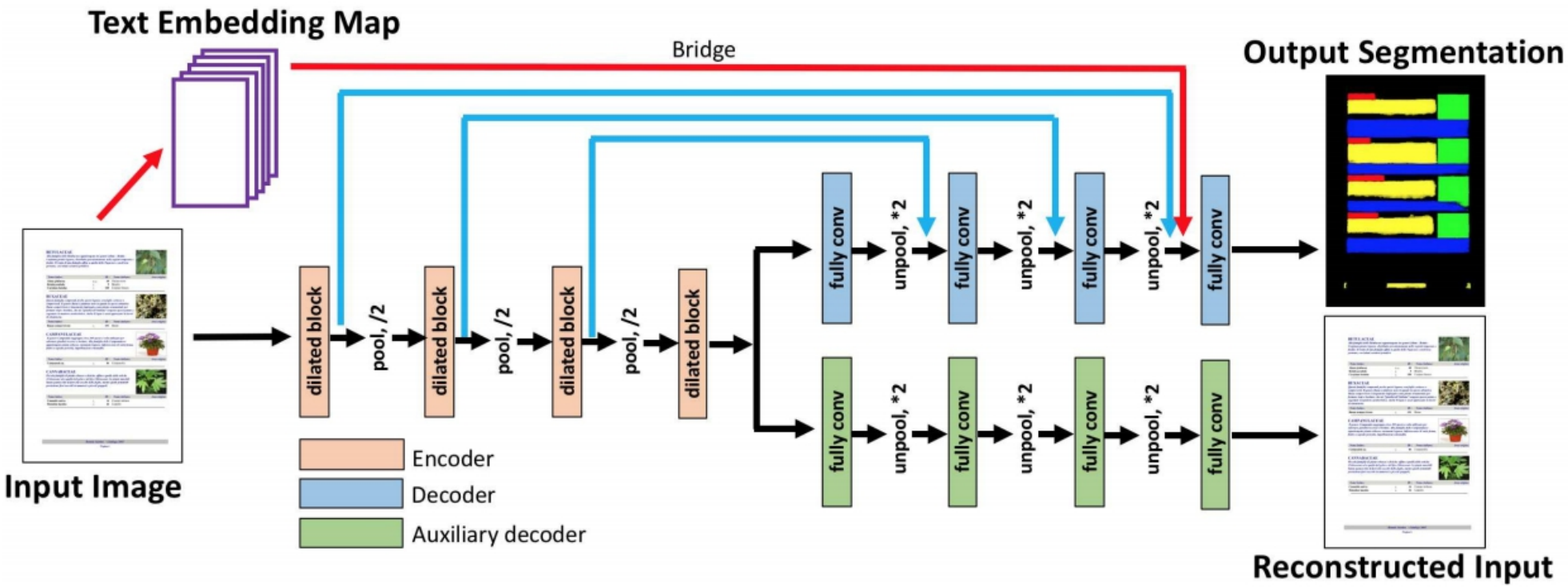
An interesting application of this method is generating large motif from only a small part of it as shown in Figure-9.

6. Conclusion

In this article we have presented an automated image correction method by inpainting which can successfully fill in missing information resulted from physical damage or due to presence of undesired object in the scene. The method can also expand the image keeping its components in the same scale. We treat the problem in a unified way by marking the region manually and label it as target region to be filled in.

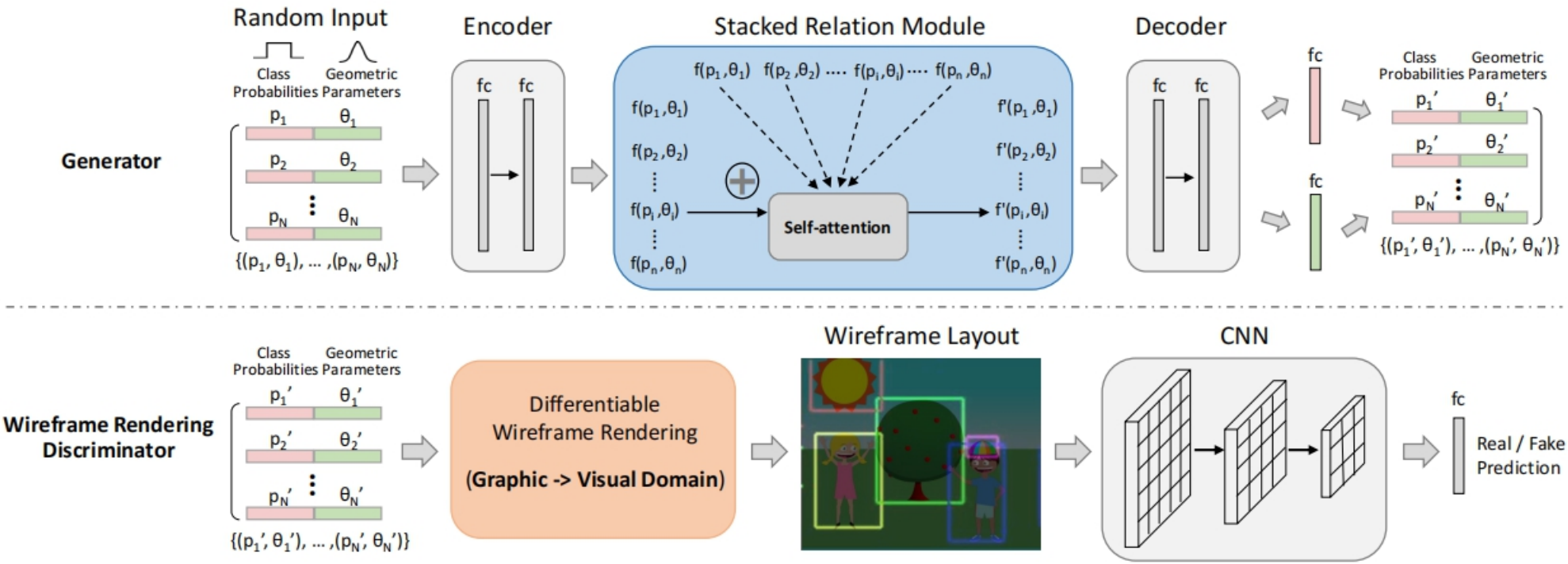
Output:*Text / Non-text area*

Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks



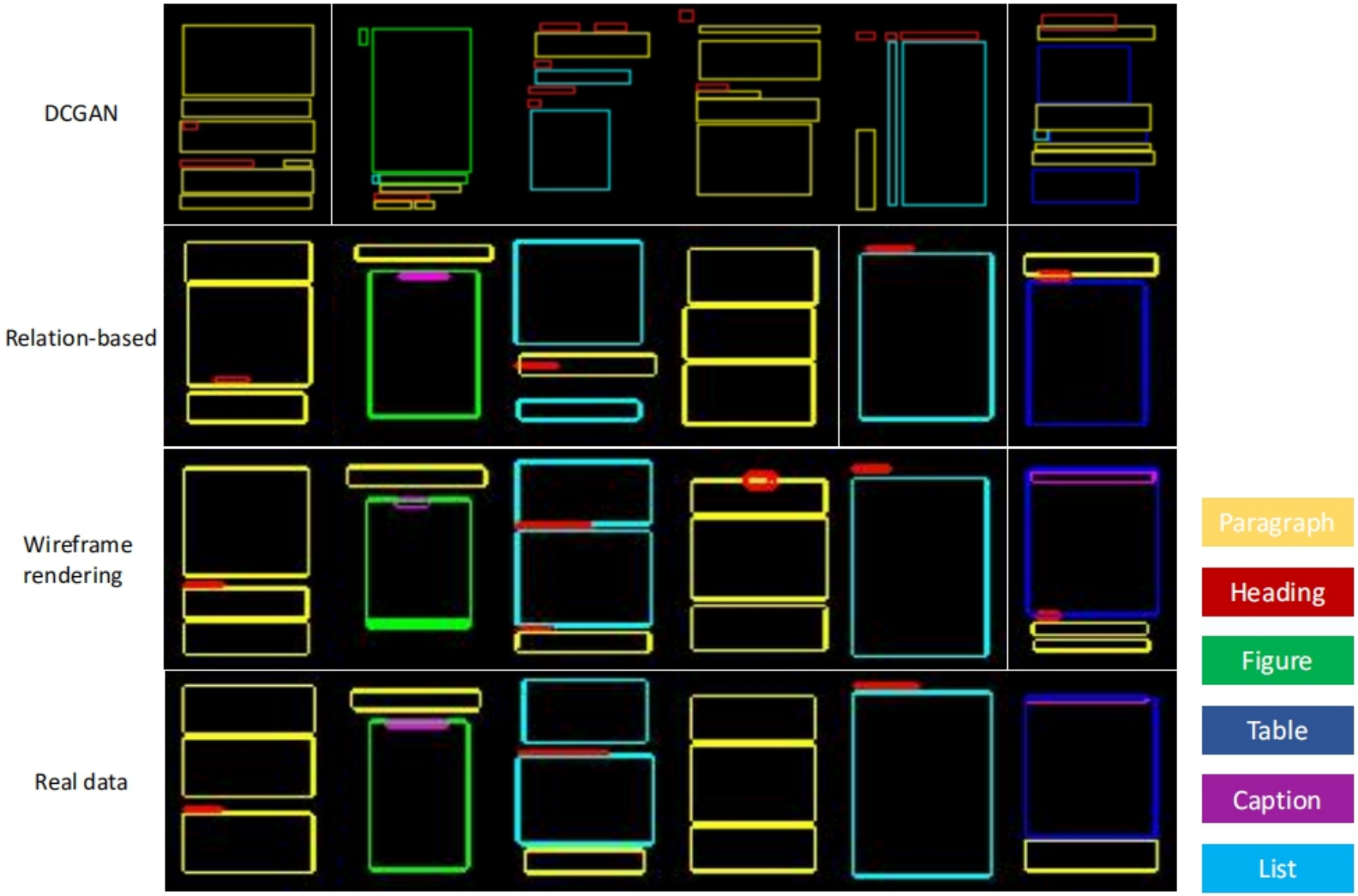
Yang X, Yumer E, Asente P, et al. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5315-5324.

LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks



Li J, Yang J, Hertzmann A, et al. LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks



Li J, Yang J, Hertzmann A, et al. LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks

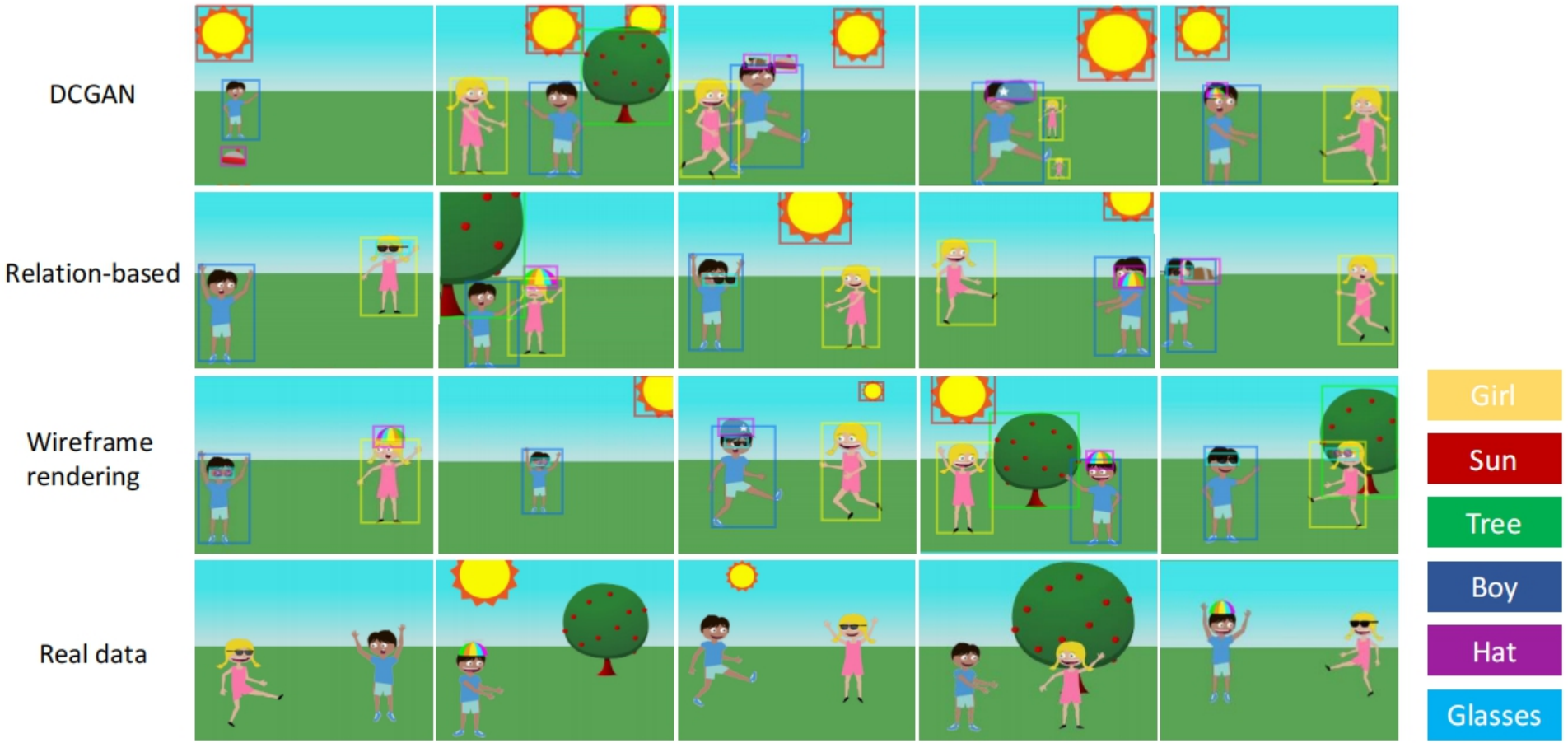


Fig. 8: Comparisons of Clipart abstract scenes from DCGAN, LayoutGAN with different discriminators and the real data.

Li J, Yang J, Hertzmann A, et al. LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

Multi-domain Document Layout Understanding Using Few-Shot Object Detection

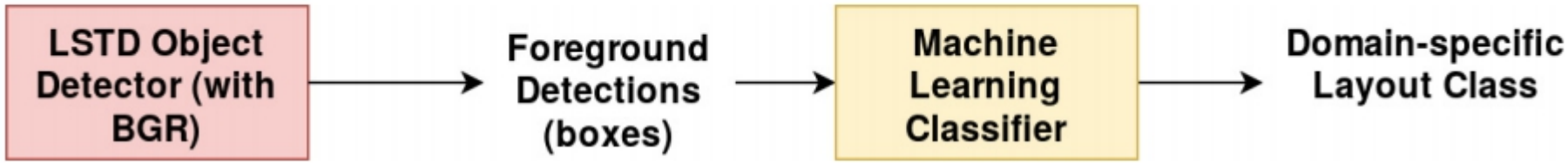


Fig. 1. Overview of the proposed method

The task can be described as few shot document layout understanding. Our methodology consists of the following parts

1. Creating the artificial (Source) dataset.
2. Pretraining the model on the Source dataset.
3. Finetuning the model on the domain-specific (Target) dataset.
4. Training the ML classifier on the Target dataset (is combined with Step 3).

Singh P, Varadarajan S, Singh A N, et al. Multi-domain Document Layout Understanding Using Few-Shot Object Detection[C]//International Conference on Image Analysis and Recognition. Springer, Cham, 2020: 89-99.

Layoutlm: Pre-training of text and layout for document image understanding

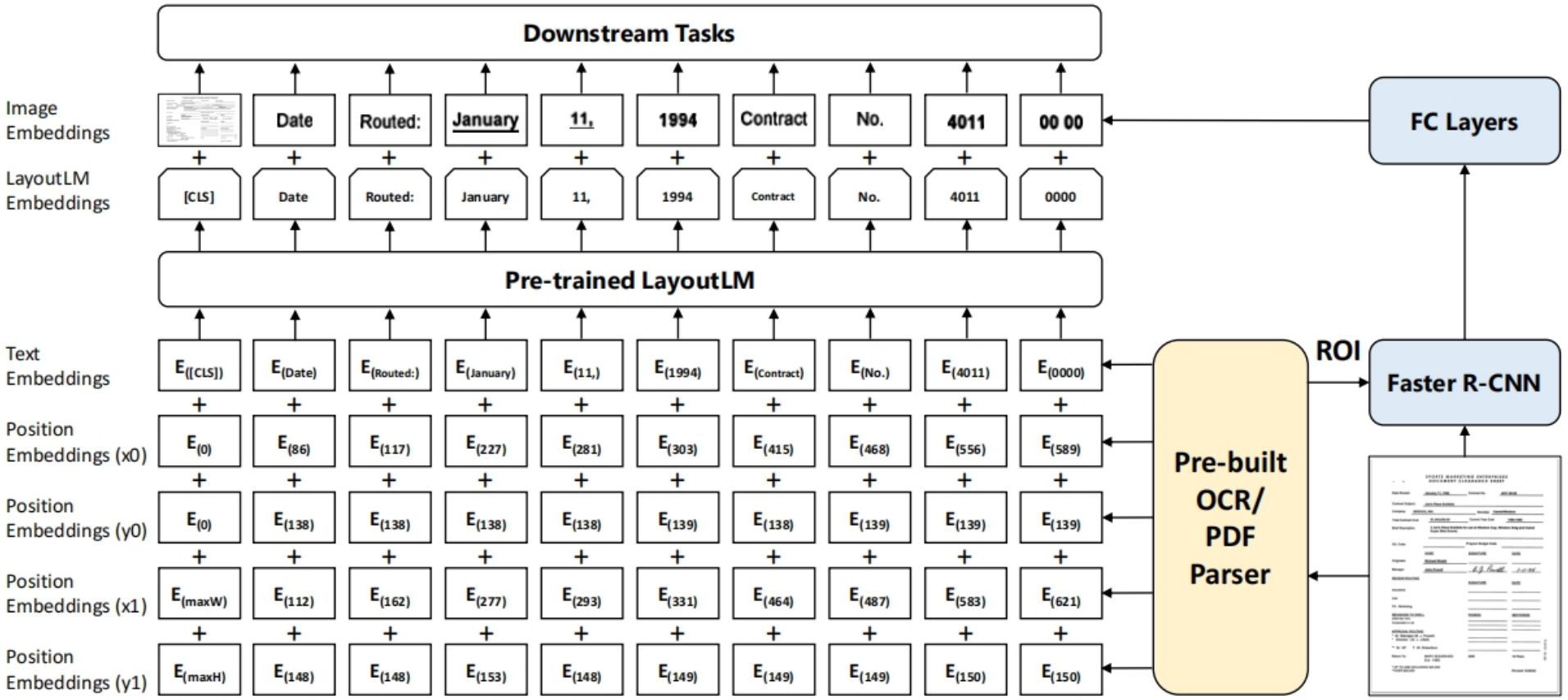
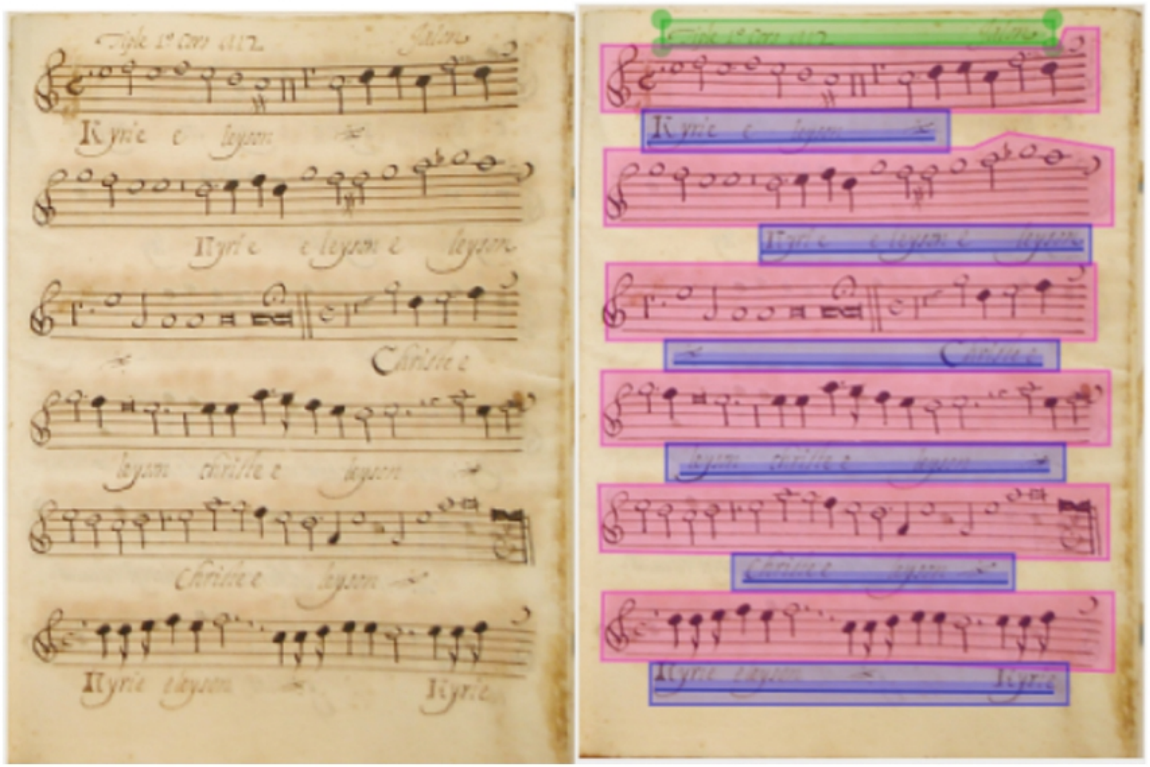


Figure 2: An example of LayoutLM, where 2-D layout and image embeddings are integrated into the original BERT architecture. The LayoutLM embeddings and image embeddings from Faster R-CNN work together for downstream tasks.

Xu Y, Li M, Cui L, et al. Layoutlm: Pre-training of text and layout for document image understanding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1192-1200.

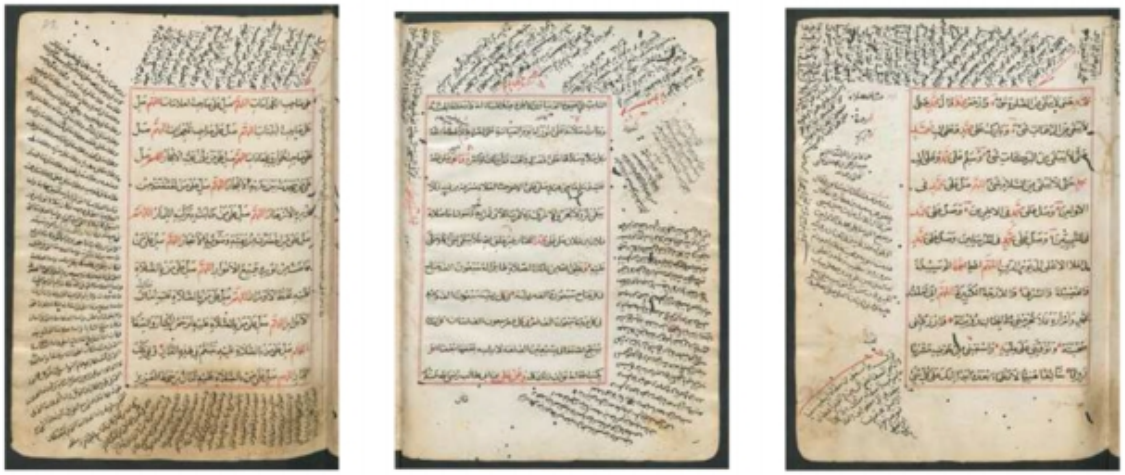
Multi-task Layout Analysis of Handwritten Musical Scores



- *A-net*: `C64:C128B:C256B:C512B:C1:Sigmoid`. Activation LeakyReLU.
- *M-net*:
  - Encoder: `C64:C128B:C256B:C512B:C512B:C512B:C512B:C512`. Activation LeakyReLU.
  - Decoder: `C512BD:C512BD:C512BD:C515B:C256B:C128B: C64B:ReLU:C` $K^{t=1} + K^{t=2}$: SoftMax. Activation ReLU.

Quirós L, Toselli A H, Vidal E. Multi-task layout analysis of handwritten musical scores[C]//Iberian Conference on Pattern Recognition and Image Analysis. Springer, Cham, 2019: 123-134.

Historical document layout analysis using anisotropic diffusion and geometric features

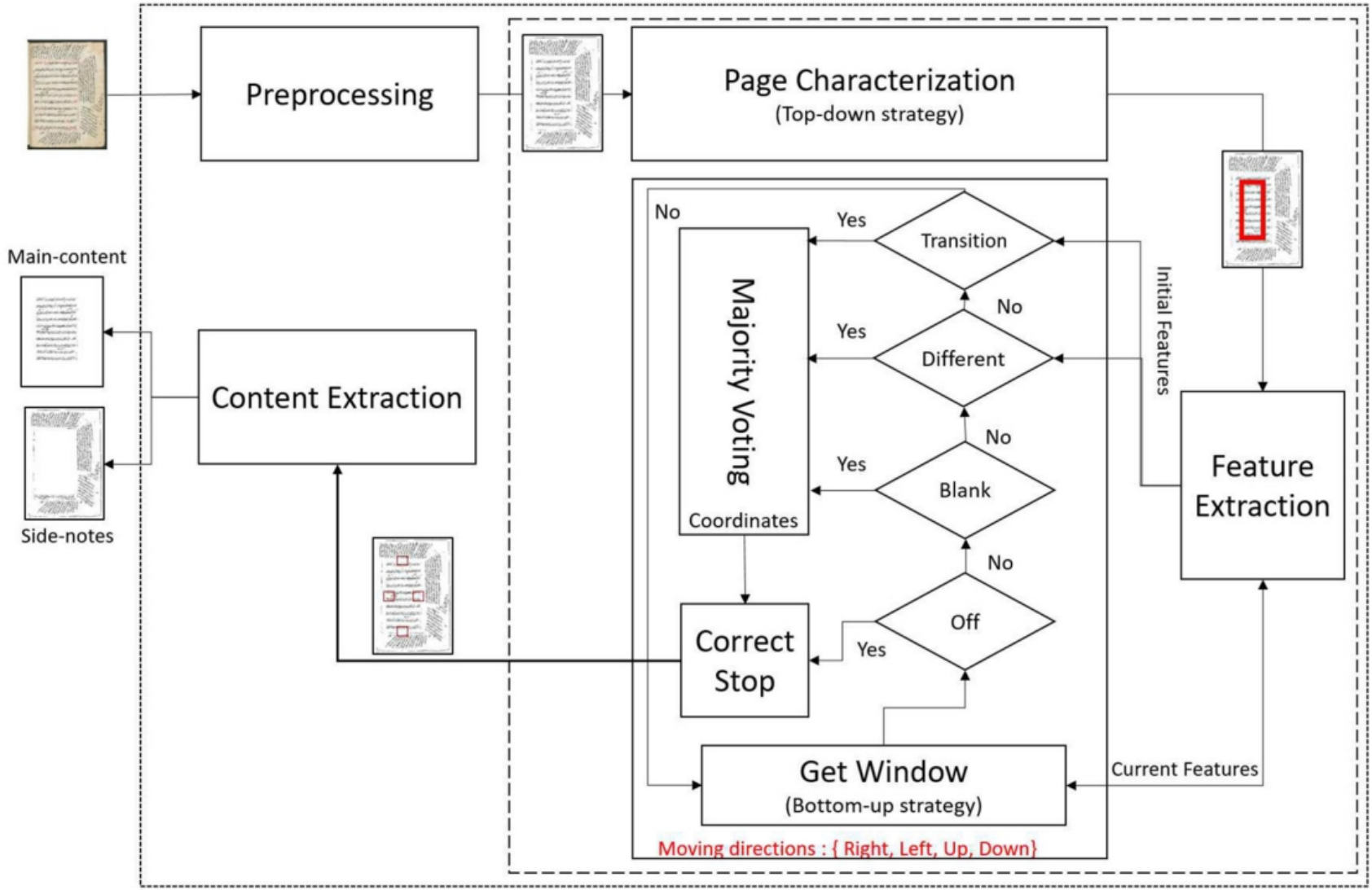


(a) Islamic Heritage Project (IHP), Harvard library

(b) Bukhari et al. [10] dataset

BinMakhashen G M, Mahmoud S A. Historical document layout analysis using anisotropic diffusion and geometric features[J]. International Journal on Digital Libraries, 2020: 1-14.

Historical document layout analysis using anisotropic diffusion and geometric features

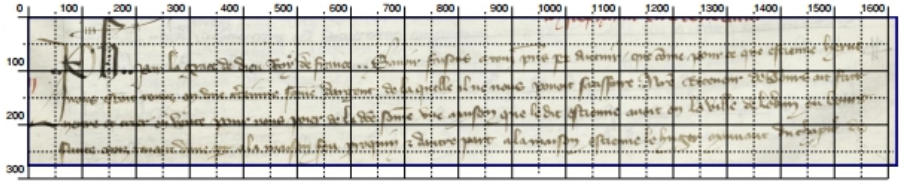


BinMakhashen G M, Mahmoud S A. Historical document layout analysis using anisotropic diffusion and geometric features[J]. International Journal on Digital Libraries, 2020: 1-14.

Text Content Based Layout Analysis

We use the word relevance probabilities provided by this map to calculate relevant text content based features at the pixel level.


```
...
<spot kw="TOUZ"          s=1.000  x=878  y=72   w=65  h=47  gt=0 />
<spot kw="UNE"           s=1.000  x=739  y=181  w=53  h=32  gt=1 />
<spot kw="VILLE"         s=1.000  x=1289 y=181  w=66  h=32  gt=1 />
<spot kw="FEU"           s=0.999  x=547  y=230  w=35  h=29  gt=1 />
<spot kw="MAISON"        s=0.999  x=820  y=181  w=87  h=32  gt=1 />
...
<spot kw="MAVANT"        s=0.001  x=1296 y=230  w=104 h=29  gt=0 />
<spot kw="PHILIPS"       s=0.001  x=105  y=72   w=117 h=47  gt=0 />
<spot kw="SAATISFAIRE"   s=0.001  x=1075 y=124  w=113 h=37  gt=0 />
```



Philippes par la grace de Dieu roys de France. Savoir faisons à tous ... maison feu Perrequin et d'autre part à la maison Estienne le Huger mouvant du chapitre de

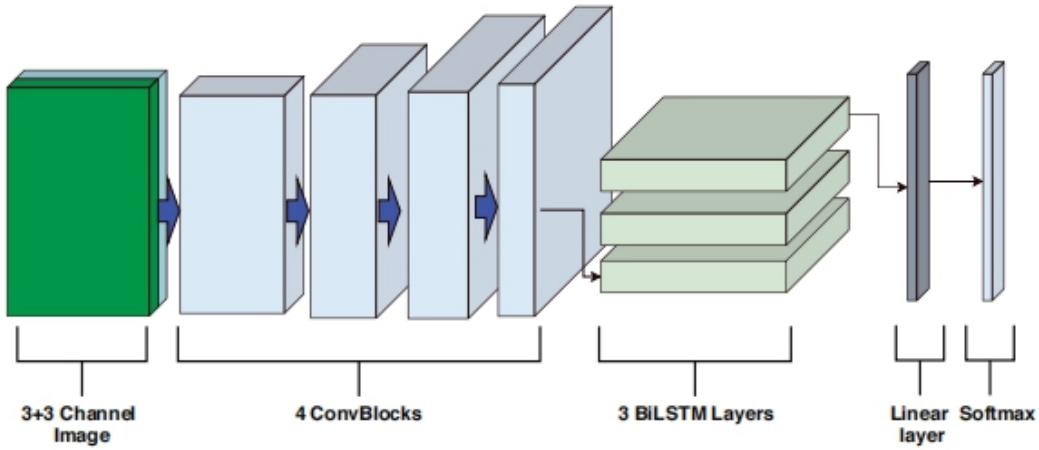


Fig. 3. Illustration of the network design. 6 channel image is created the concatenation of the input image in RGB channels with the 3 images in grey scale created with textual information. The attention variables that are multiplying the textual information are not shown in the figure.

Prieto J R, Bosch V, Vidal E, et al. Text Content Based Layout Analysis[C]//2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2020: 258-263.

Deep Arabic document layout analysis

Figure 2: Original document image (left) and the result after ARLSA (right).

Figure 3: Original document image (left) and the result after identifying connected components' regions (right).

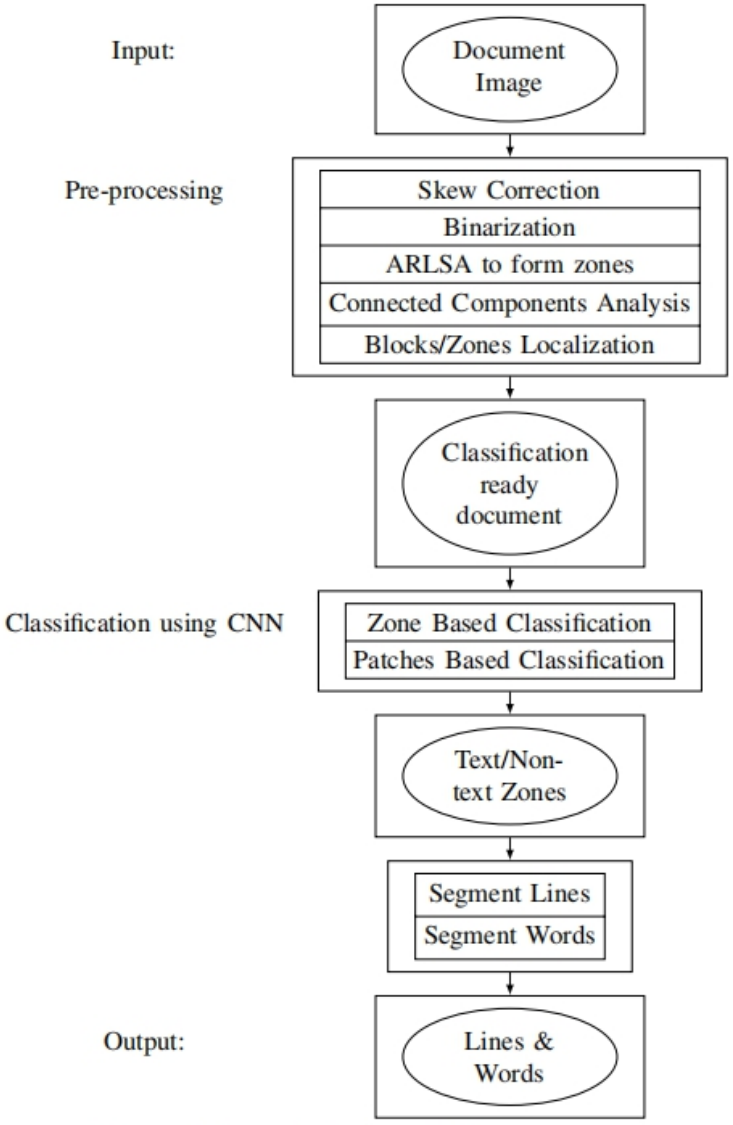


Figure 1: Proposed System Architecture

Amer I M, Hamdy S, Mostafa M G M. Deep Arabic document layout analysis[C]//2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS). IEEE, 2017: 224-231.

Document Analysis And Classification Based On Passing Window

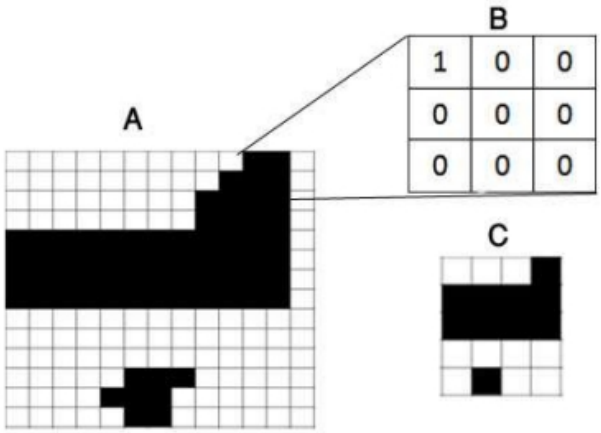


Fig.1: an example of rescaling the images with 3*3 window

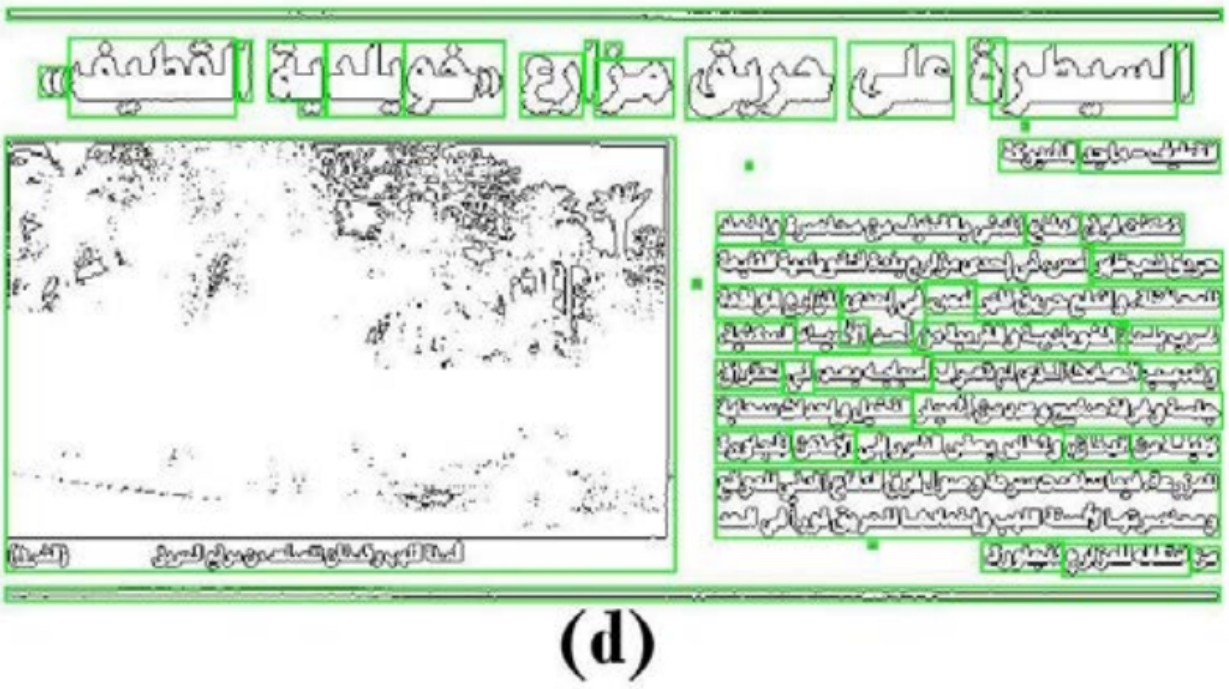


Fig.2:the proposed algorithm for segmentation; (a) a sample image; (b), and (c)are the result images after applying the first and second steps, and (d) shows the boundaries of the resulting regions

BAMASOOD Z A. Document Analysis And Classification Based On Passing Window[J]. Journal of Advances in Computer Engineering and Technology, 2020, 6(1): 39-46.

Fast CNN-based document layout analysis

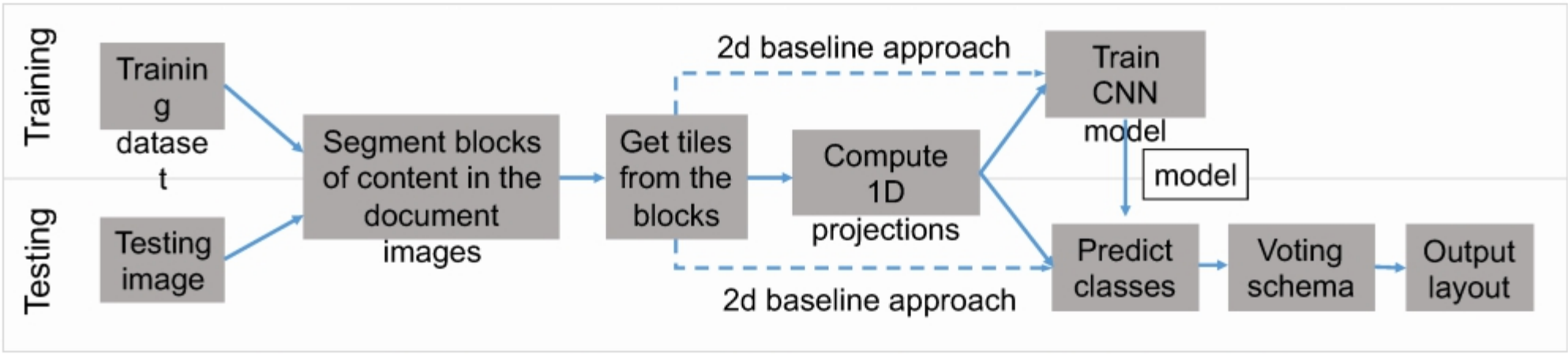


Figure 1. Block-diagram of out proposed methodology for document image layout analysis.

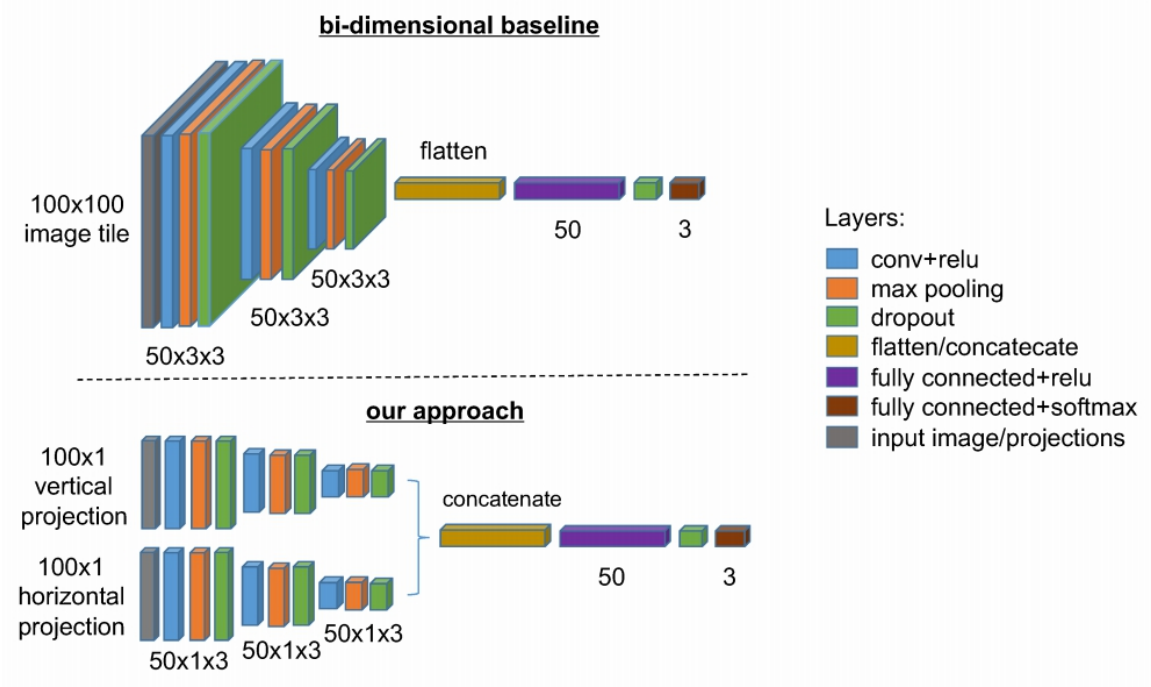


Figure 4. CNN models architectures: bi-dimensional baseline and the proposed one-dimensional approach.

Augusto Borges Oliveira D, Palhares Viana M. Fast CNN-based document layout analysis[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 1173-1180.

An Approach How to Automate Labeling Data for the Training ANN Models for Page Layout Analysis

– LaTeX source file converted to a DVI file, which could then be converted to PostScript with dvips. This, in turn, can be converted to a PDF file by ps2pdf [3] tool.

```
          latex              dvips              ps2pdf
text.tex -------> text.dvi -------> text.ps --------> text.pdf
```

– The step with conversion to PostScript can be skipped.

```
          latex            dvipdfm
text.tex -------> text.dvi -------> text.pdf
```

– Directly from the LaTeX source to PDF file by pdflatex program.

```
         pdflatex
text.tex --------> text.pdf
```

Mikhailov A. An Approach How to Automate Labeling Data for the Training ANN Models for Page Layout Analysis[J]. 2020.

An Approach How to Automate Labeling Data for the Training ANN Models for Page Layout Analysis

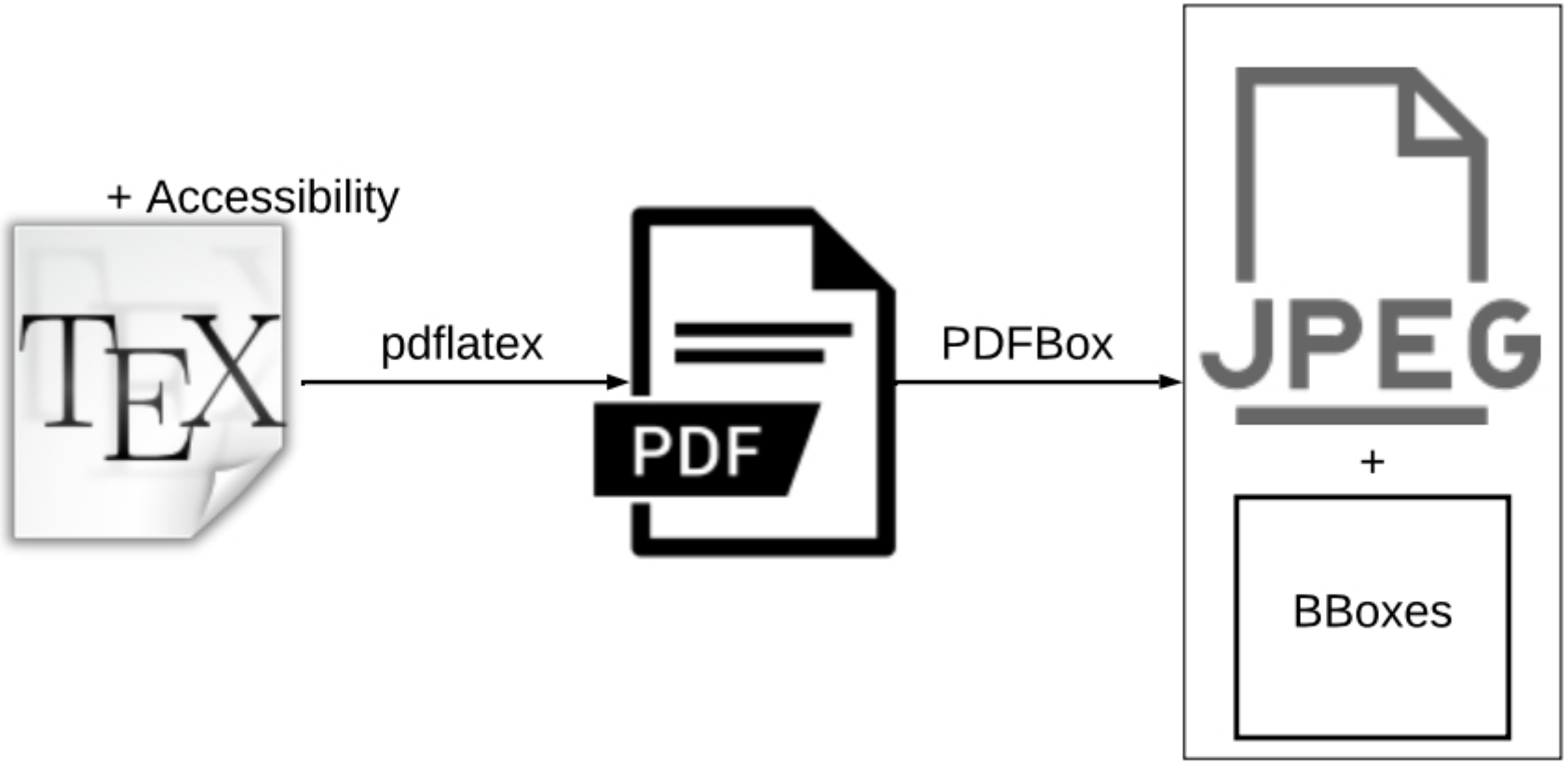


Fig. 1. Labeling PDFs process

Mikhailov A. An Approach How to Automate Labeling Data for the Training ANN Models for Page Layout Analysis[J]. 2020.

Pixel-Level Evaluation.

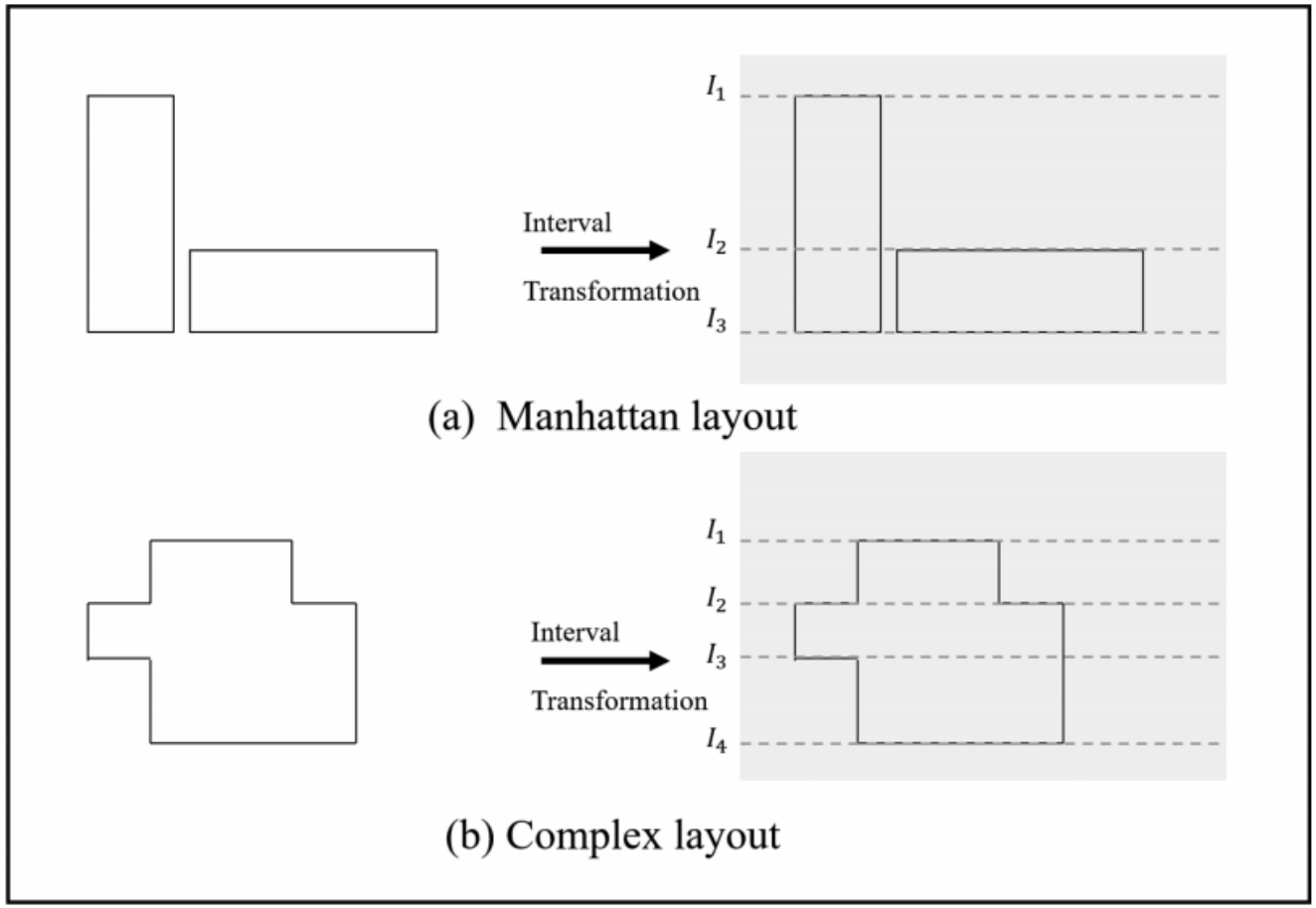
$$MS(i,j) = \alpha \frac{\left(T\left(G_j \cap R_j \cap I\right)\right)}{\left(T\left(\left(G_j \cup R_j\right) \cap I\right)\right)} \quad \text{where} \quad \alpha = \begin{cases} 1, & \text{if} \quad g_i = r_i \\ 0, & otherwise \end{cases}$$

$$DR = w_1 \frac{one2one}{N_i} + w_2 \frac{one2many}{N_i} + w_3 \frac{many2one}{N_i}$$

$$RR = w_4 \frac{one2one}{M_i} + w_5 \frac{one2many}{M_i} + w_6 \frac{many2one}{M_i}$$

$$F_{measure} = \frac{2 \times DR \times RR}{DR + RR}$$

Region-Level Evaluation.



$$IoU = \alpha_1 \frac{G_i \cap R_i}{t_i + (G_i \cup R_i) - (G_i \cap R_i)}$$

Performance evaluation

$$Acc = \frac{\sum_i M_{ii}}{\sum_{ij} M_{ij}}$$

$$P_i = \frac{M_{ii}}{\sum_j M_{ji}}$$

$$R_i = \frac{M_{ii}}{\sum_j M_{ij}}$$

$$F_1 = \frac{2 \times \sum_{i=1}^{n} P_i \times \sum_{i=1}^{n} R_i}{n(\sum_{i=1}^{n} P_i + \sum_{i=1}^{n} R_i)}$$

Thank you