

Document and Content Analysis

Summer 2009

Lecture 6
Document Image Analysis

Thomas Breuel
Faisal Shafait

- **Documents come to us in**
 - Paper format (letters, books, newspapers, magazine, ...)
 - Electronic format (E-Mail, PDF, Word, Web page, ...)
- **For better handling of documents around us, both formats should be seamlessly interchangeable**

- **Documents come to us in**
 - Paper format (letters, books, newspapers, magazine, ...)
 - Electronic format (E-Mail, PDF, Word, Web page, ...)
- **For better handling of documents around us, both formats should be seamlessly interchangeable**
- **How do we convert an electronic document to paper format?**

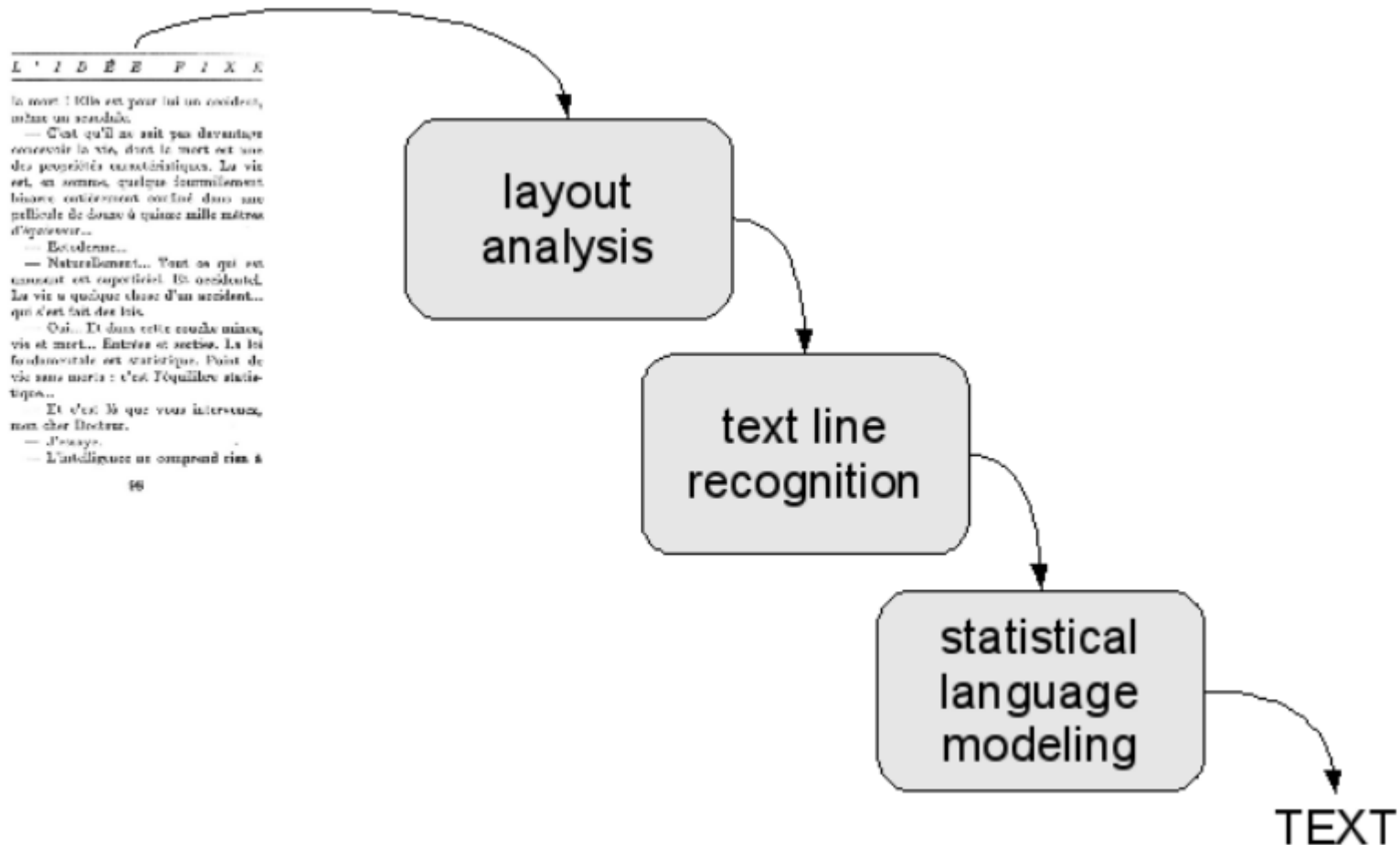
- **Documents come to us in**
 - Paper format (letters, books, newspapers, magazine, ...)
 - Electronic format (E-Mail, PDF, Word, Web page, ...)
- **For better handling of documents around us, both formats should be seamlessly interchangeable**
- **How do we convert an electronic document to paper format?**
- **How do we convert a paper document into electronic format?**

Optical Character Recognition (OCR)

Optical Character Recognition (OCR)

We will talk about OCR in the next FOUR lectures

Flow chart for OCR



Outline

Optical Character Recognition

Capturing document images

Binarization

Skew Correction

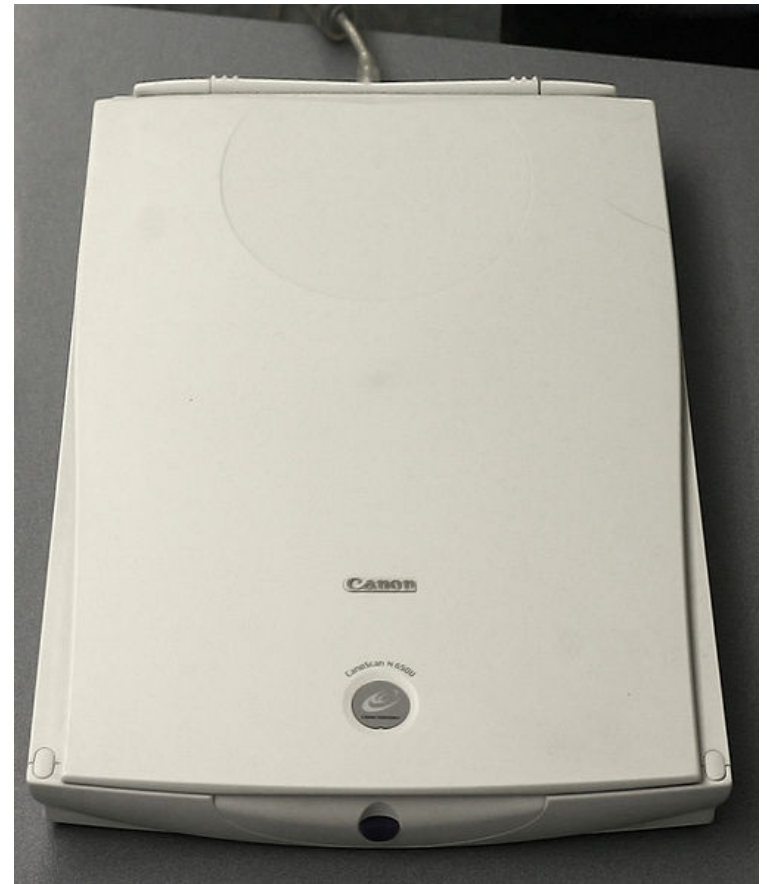
Page Segmentation

Layout Analysis

Capturing document images

- **Flat-bed scanners**

- A glass plate with a light source underneath
- Page is placed upside down
- An array of CCD sensors moves over the page and collected the light reflected by the page
- Suitable for processing a few pages occasionally



Automatic Feed Scanners

- **High scan throughput (more than 100 page per minute)**
- **Can not scan bound material (e.g. books)**

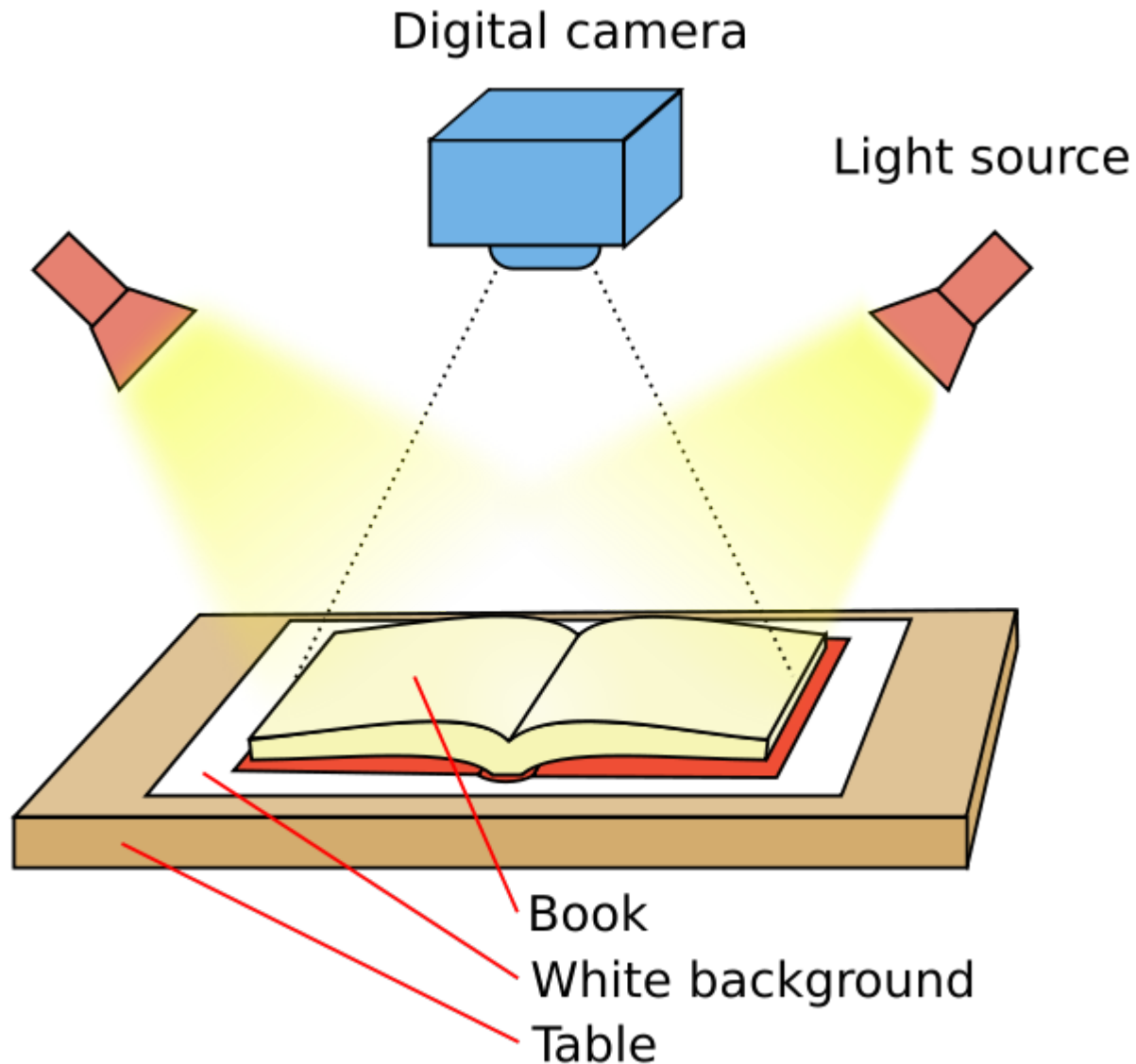


Book Scanners

- **Scan bound volumes using overhead camera**
- **Some book scanners have automatic page turning (not very reliable)**



Book Scanners Principle

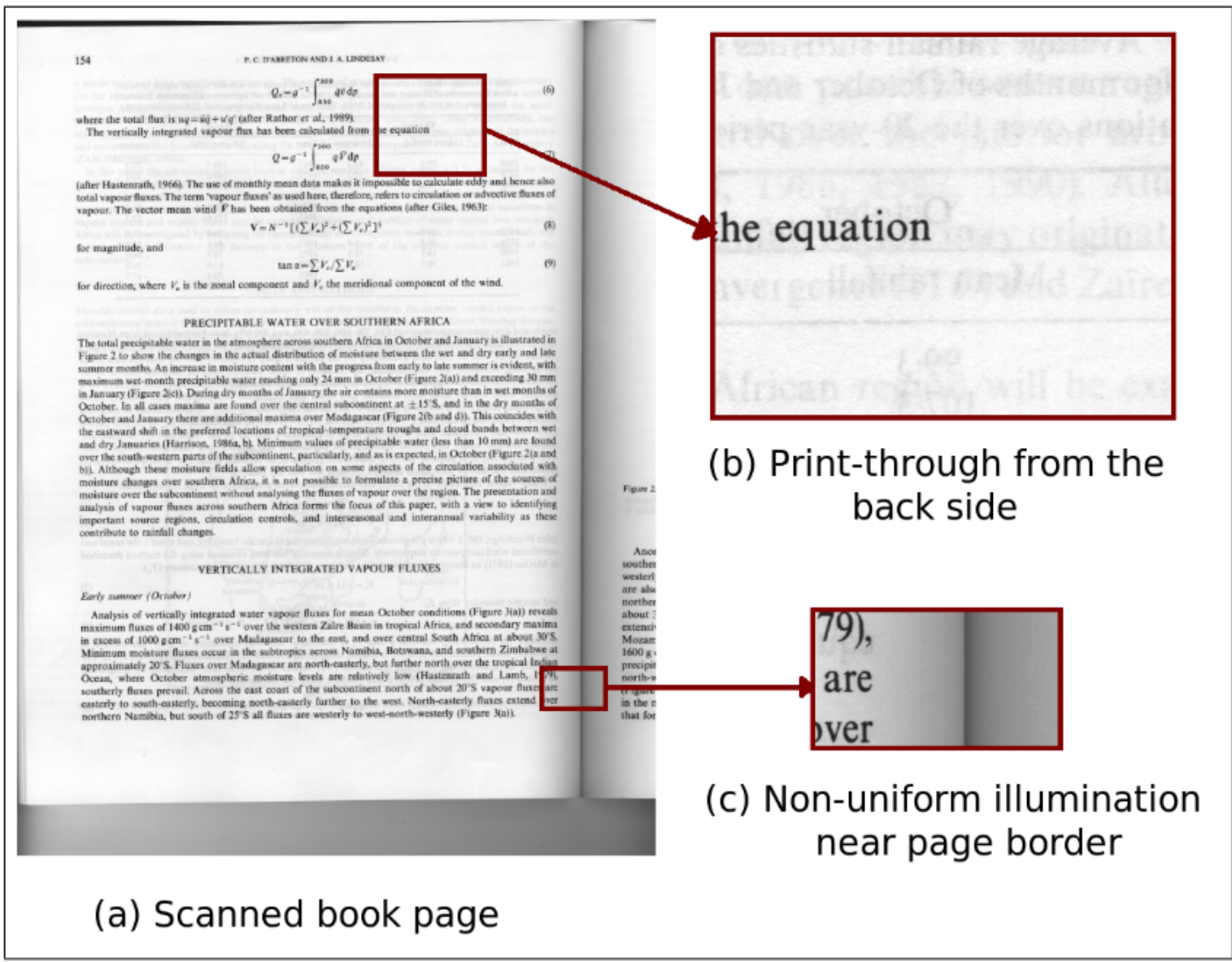


Micro-film Scanners

- **Scan material that is out-of-print**
- **Used for digitizing historical documents (newspapers, books)**



A typical scanned book page



Binarization

- **Scanners capture a greyscale/color document**
- **Most of the OCR systems work on binary images**
- **Binarization is an important first step in most of the document analysis systems**

Effect of binarization on OCR

north over the tropical Indian

north over the tropical Indian

OCR →

north over thc tropical Indian

north over the tropical Indian

OCR →

north over thc tropical

north over the tropical Indian

OCR →

north mer thc tmpieul lmdhm

Binarization algorithms

- **The goal of binarization algorithm is to define a threshold.**
- **Two main classes:**
 - Global binarization

$$o(x, y) = \begin{cases} 0 & \text{if } g(x, y) \leq T \\ 255 & \text{otherwise} \end{cases}$$

- Local binarization

$$o(x, y) = \begin{cases} 0 & \text{if } g(x, y) \leq t(x, y) \\ 255 & \text{otherwise} \end{cases}$$

Global Binarization

- **Just set** $T = 128$

Global Binarization

- **set**
$$T = \frac{\min_{(x,y)} g(x,y) + \max_{(x,y)} g(x,y)}{2}$$

Otsu Global Thresholding

Let h be the normalized histogram of the image

$$p_1 = \sum_{g=0}^T h_g$$

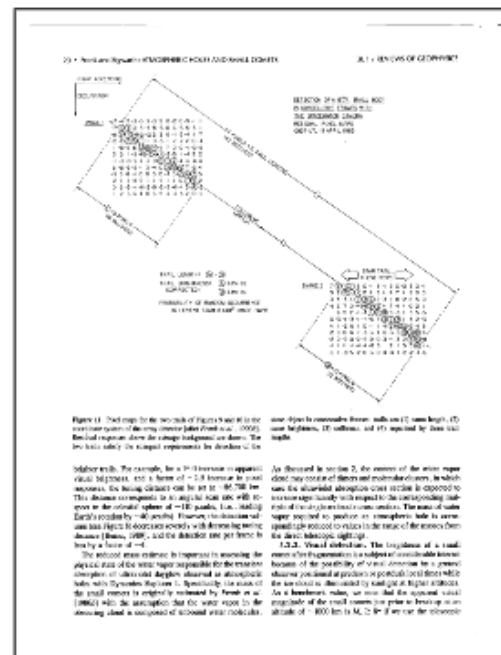
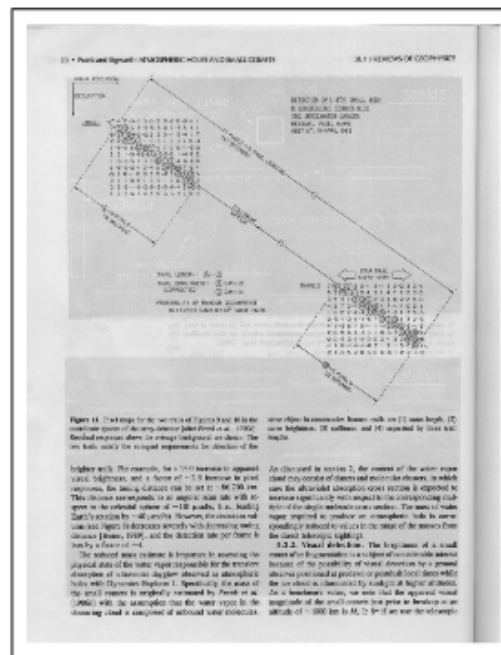
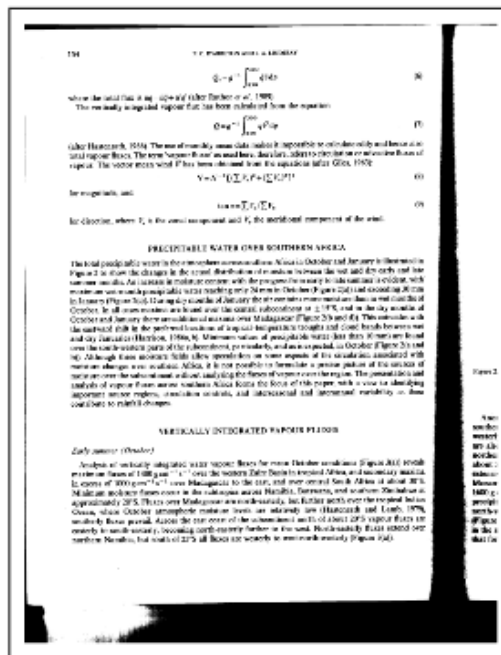
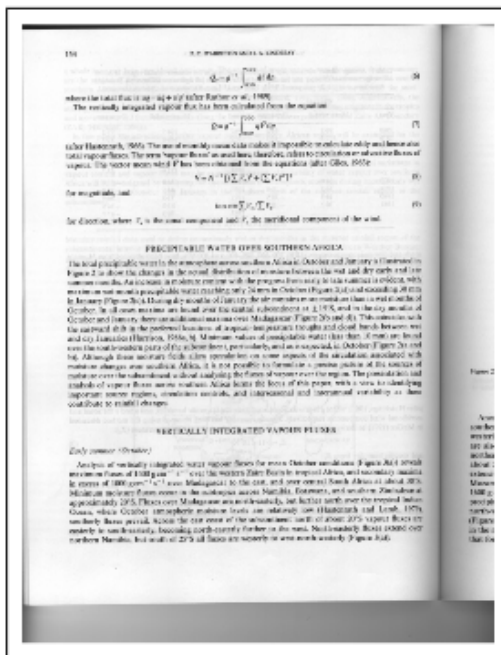
$$\mu_1 = \frac{1}{p_1} \sum_{g=0}^T g h_g$$

$$p_2 = \sum_{g=T+1}^{L-1} h_g = 1 - p_1$$

$$\mu_2 = \frac{1}{p_2} \sum_{g=T+1}^{L-1} g h_g$$

$$\hat{T} = \arg \max_T p_1 p_2 (\mu_1 - \mu_2)^2$$

Otsu Global Thresholding



(a) Input image

(b) Otsu's result

(c) Input image

(d) Otsu's result

Local Adaptive Thresholding

- **Adapt to local variations in intensity by taking a $w \times w$ window around each pixel**

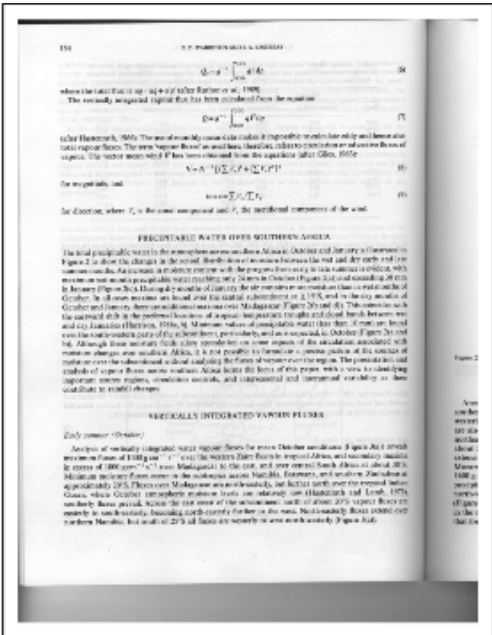
$$o(x, y) = \begin{cases} 0 & \text{if } g(x, y) \leq t(x, y) \\ 255 & \text{otherwise} \end{cases}$$

White (1983): $t(x, y) = km(x, y)$

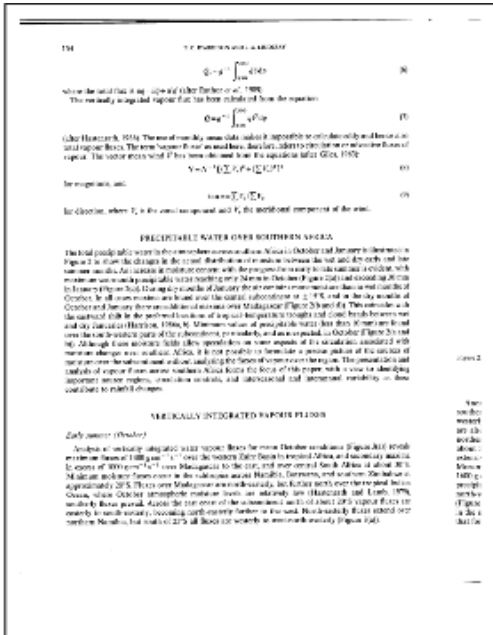
Niblack (1986): $t(x, y) = m(x, y) + ks(x, y)$

Sauvola (2000): $t(x, y) = m(x, y) \left[1 + k \left(\frac{s(x, y)}{R} - 1 \right) \right]$

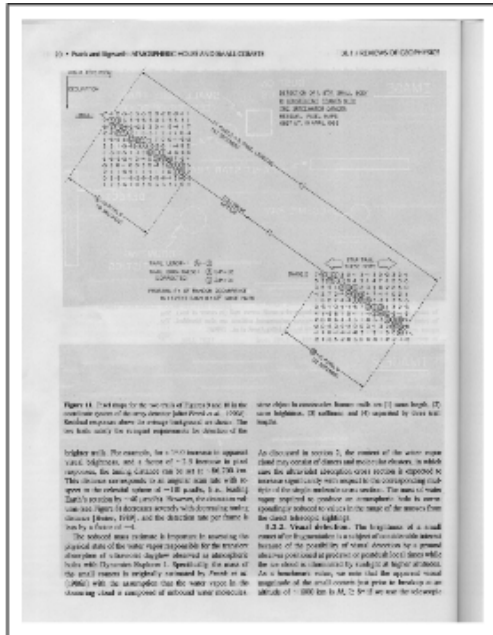
Sauvola Local Thresholding



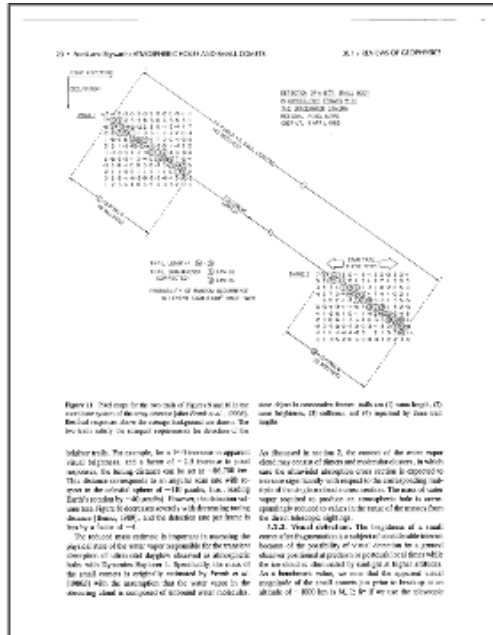
(a) Input image



(b) Sauvola's result



(c) Input image



(d) Sauvola's result

Local Vs Global Thresholding

- **Global Thresholding methods are:**

- Fast
- Give good results when illumination over a page is uniform
- Fail when there are local changes in illumination

- **Local Thresholding methods are:**

- Slow
- Adapt to local changes in illumination
- Perform well for both uniform and non-uniform illumination

Shafait Binarization (2008)

- **Use integral images for computing local thresholds**

$$I(x, y) = \sum_{i=0}^x \sum_{j=0}^y g(i, j)$$

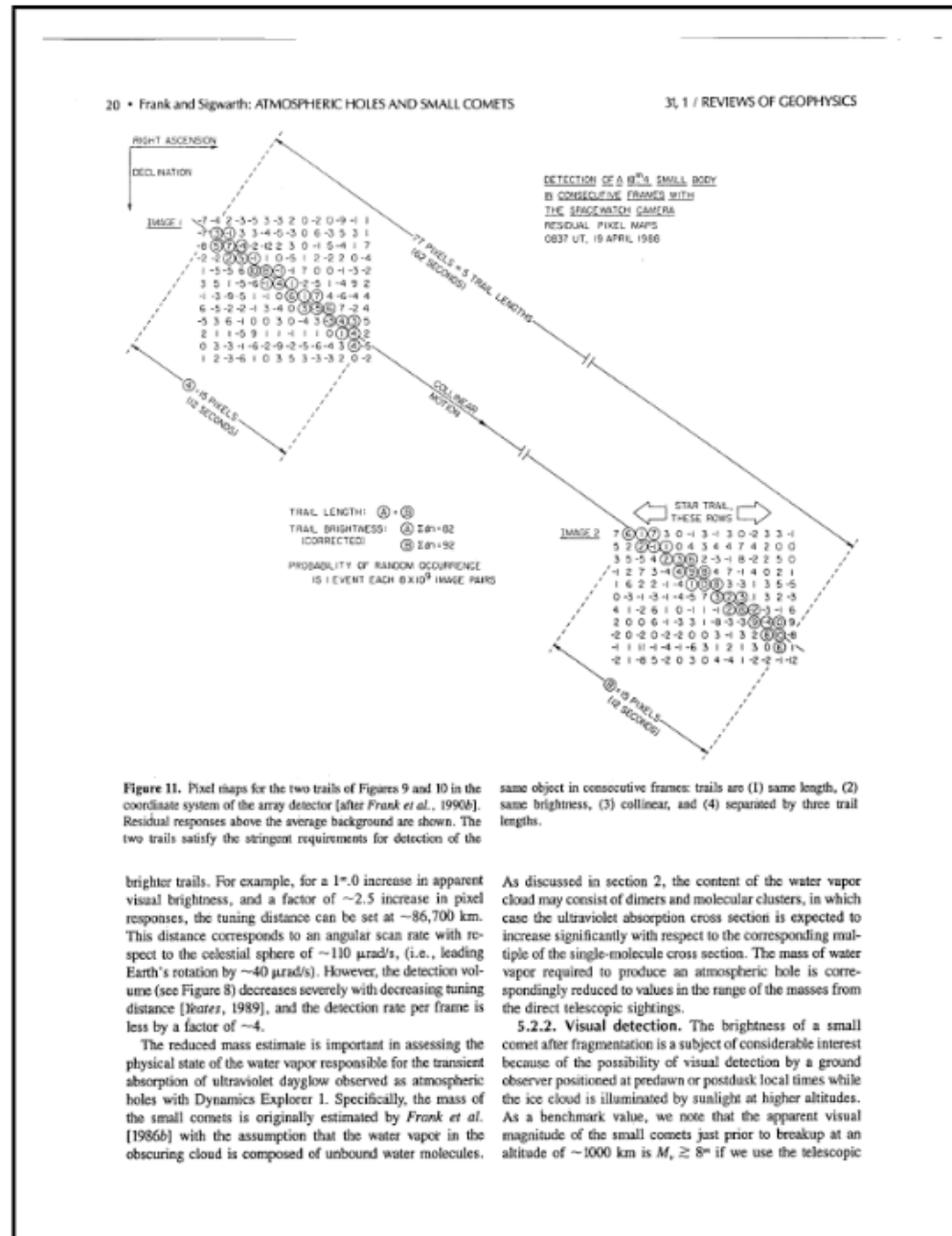
- **Local mean and variance can be computed in linear time**

$$m(x, y) = (I(x + w/2, y + w/2) + I(x - w/2, y - w/2) - I(x + w/2, y - w/2) - I(x - w/2, y + w/2)) / w^2$$

$$s^2(x, y) = \frac{1}{w^2} \sum_{i=x-w/2}^{x+w/2} \sum_{j=y-w/2}^{y+w/2} g^2(i, j) - m^2(x, y)$$

- **Same performance as local thresholding in time close to global thresholding**

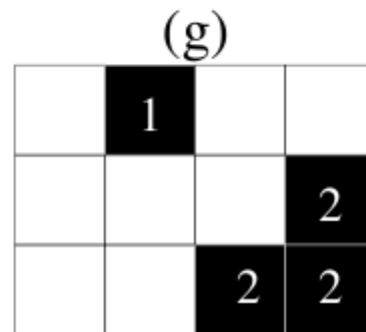
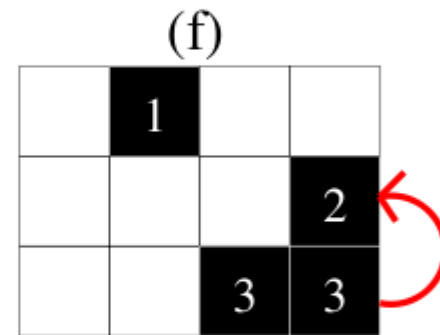
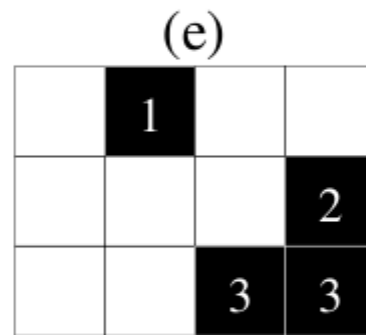
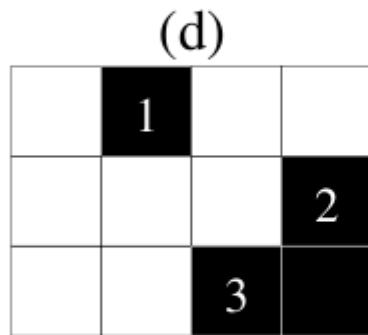
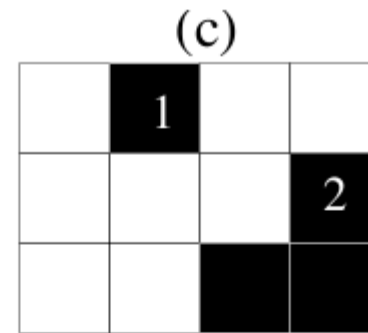
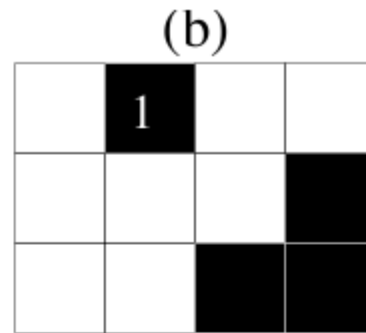
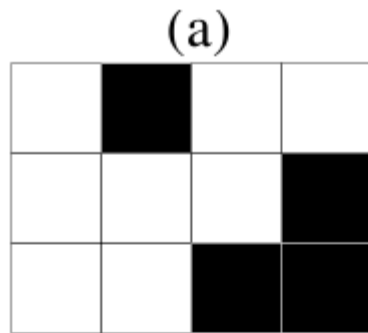
Connected Component Analysis



Connected Component Analysis

- **Scan the image row by row**
- **When a black pixel is encountered, assign it a label:**
 - If left neighbor pixel is white, a new label is assigned to the current black pixel
 - If left neighbor is black, its label is copied to the current pixel
- **If the upper neighbor pixel is black, merge the label of the current pixel and that of upper neighbor**

Connected Component Analysis Example



Other Pre-processing Tasks

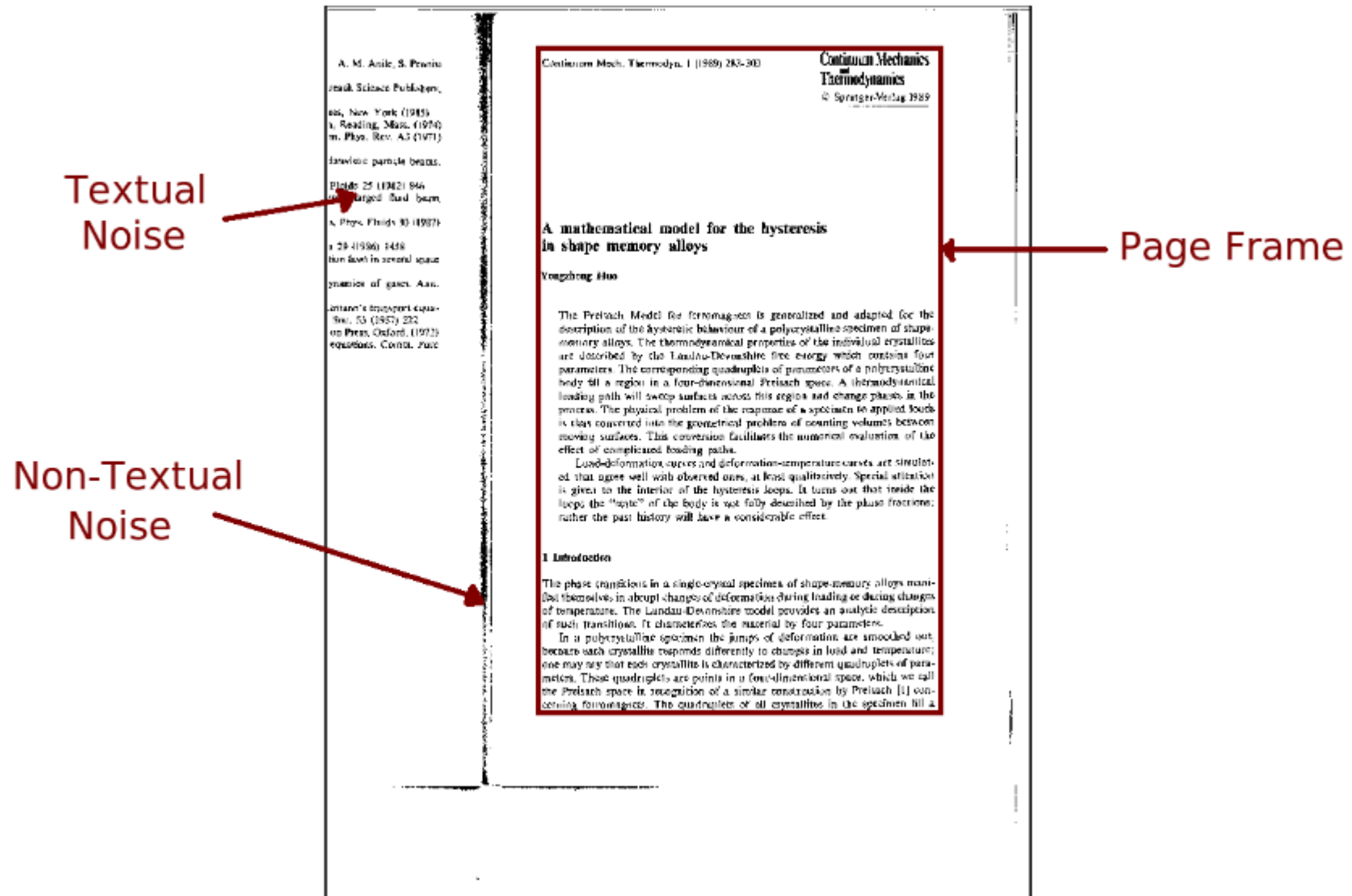
- **Orientation detection**
- **Marginal noise removal**
- **Skew correction**

Orientation Detection

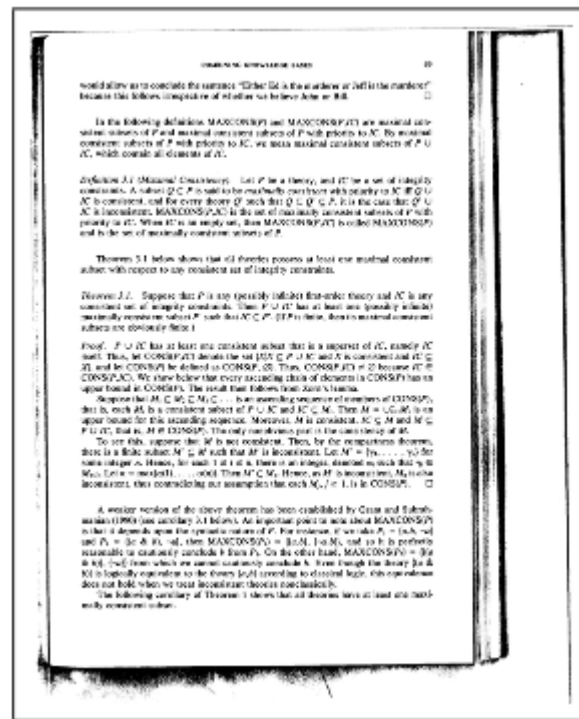
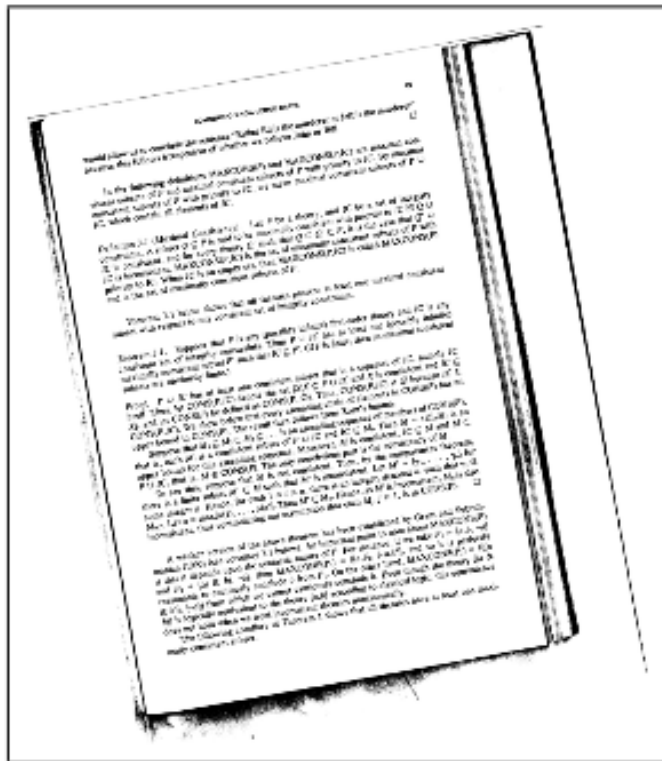
[illegible]

Row	Column	Value
1	1	000
2	1	000
3	1	000
4	1	000
5	1	000
6	1	000
7	1	000
8	1	000
9	1	000
10	1	000
11	1	000
12	1	000
13	1	000
14	1	000
15	1	000
16	1	000
17	1	000
18	1	000
19	1	000
20	1	000
21	1	000
22	1	000
23	1	000
24	1	000
25	1	000
26	1	000
27	1	000
28	1	000
29	1	000
30	1	000
31	1	000
32	1	000
33	1	000
34	1	000
35	1	000
36	1	000
37	1	000
38	1	000
39	1	000
40	1	000
41	1	000
42	1	000
43	1	000
44	1	000
45	1	000
46	1	000
47	1	000
48	1	000
49	1	000
50	1	000
51	1	000
52	1	000
53	1	000
54	1	000
55	1	000
56	1	000
57	1	000
58	1	000
59	1	000
60	1	000
61	1	000
62	1	000
63	1	000
64	1	000
65	1	000
66	1	000
67	1	000
68	1	000
69	1	000
70	1	000
71	1	000
72	1	000
73	1	000
74	1	000
75	1	000
76	1	000
77	1	000
78	1	000
79	1	000
80	1	000
81	1	000
82	1	000
83	1	000
84	1	000
85	1	000
86	1	000
87	1	000
88	1	000
89	1	000
90	1	000
91	1	000
92	1	000
93	1	000
94	1	000
95	1	000
96	1	000
97	1	000
98	1	000
99	1	000
100	1	000

Marginal Noise Removal



Skew Correction



Page Segmentation



Figure 4. View of the canoe raised above, illustrating the board and endboard fitting remaining within the walls of the raised deck.

[illegible]

future research, focusing on the fact that the main reason for having a low level of awareness of the health system are individuals for example the socially excluded (number 10). While rudimentary awareness may only indicate developed levels of cognitive skills, capacity and cultural factors related to the health system, some basic understanding of health care is available to the majority of the population. However, much of this understanding is generally a tradition, not a practice, and it is well separated from places where the individuals involved may be engaged in essential health care. Where the gap is so significant and the individuals have no formal training, the health system has to be reformed (number 10).

Several studies have found that exposure to the highly spinous, needle-like RFLP causes a prolonged length of time to complete the removal of the sharp, uncoated needle and increases the risk of a serious injury that has occurred between the participant and the case during pregnancy [2]. Although the use of a needle is not a direct cause of exposure, the handling of the needle is a likely cause. It is important to note that the use of a needle is not a possible cause of exposure.

KEY TOPICS TO REMEMBER

1. In the early 1990s, when the stock market was in a record decline, the Federal Reserve lowered the discount rate to 5 percent, the lowest level in the history of the United States. This move was intended to stimulate the economy by making borrowing cheaper. However, the Fed's actions were criticized for being too aggressive, leading to a sharp increase in the price of Treasury bonds and a decline in the price of stocks. The Fed's actions were also criticized for being too late, as the economy had already begun to recover by the time the rate was lowered.

Product category: The 1990s were a time when the need to smooth the supply of capital to the manufacturing and small plant sectors of the economy became acute. The need for a new class of financial instruments, able to provide high quality, high yield financing was strong. Many thousands of investment grade, medium to large size companies were being sold to private and public investors and many new ones were being established from scratch. The need for a new class of



Figure 6. 24. Plot of $\ln(\text{rate})$ versus $\ln(\text{time})$ for the reaction of 2,4-dinitrophenol with 2,4-dinitrophenol.

(a) Segmentation A



Figure 4. View of finished, soaked and bleached showing the complete and good ring remaining within the rope on the upper 10 cm of the rope (from the head of the fish).

and the other of internal sources (no external forces). Assuming that the number of waves is large and that the wave length is small, the perturbation is reduced to a set of linearized equations. The linearized equations are then solved numerically, and the results are compared with the exact solution. The results show that the linearized equations are in good agreement with the exact solution, and that the perturbation is reduced to a set of linearized equations.

where the authors, including the author of this review, are in the conflict of interest. However, the findings reported in this review looking at the relationship between the presence of the marital homicide suspect and the likelihood of the subsequently substantiated sexual abuse are well developed. There is supplementary support for, rather than, a role for the husband's alleged sexual abuse in the likelihood of the rape. However, it seems that the sexual abuse is not normally a positive, but, a positive, indicator of subsequent child sexual abuse. It is well established that prior sexual abuse is not a necessary condition for subsequent sexual abuse. The authors also note that there is a need for further research in this area.

from a population of 1000. Each cell is cultured, and a selection based on the selection of cells by RT-PCR is made. In general, length polymorphism analysis is not used in the direct sequencing method because the DNA generated through the PCR may not be used to create the parent plasmid and thus is not suitable for sequencing. Although the method is not suitable for cloning, it is useful in cases where the sequencing is not required and the information on the sequence is sufficient for the purpose of the study.

Info: **Global warming**

In the past decade, there has been a lot of discussion about global warming. Global warming is the gradual increase in the average temperature of the earth's atmosphere predicted by the expansion that will occur over the next few decades as a result of increased concentrations of greenhouse gases in the atmosphere. The most common greenhouse gas is carbon dioxide (CO_2). Other greenhouse gases include methane (CH_4) and nitrous oxide (N_2O). These gases are released into the atmosphere by burning fossil fuels, deforestation, and other human activities. The increase in the concentration of these gases in the atmosphere is causing the earth's temperature to rise, which is leading to a variety of problems, including melting glaciers, rising sea levels, and more frequent extreme weather events.

Abstract pages
 In the 1970s, when there were almost no journals devoted to psychology, the American Psychological Association published a journal devoted to the study of human factors. The journal was called *Human Factors* and it was the only journal in the field. The journal was published by the American Psychological Association and it was the only journal in the field. The journal was published by the American Psychological Association and it was the only journal in the field.

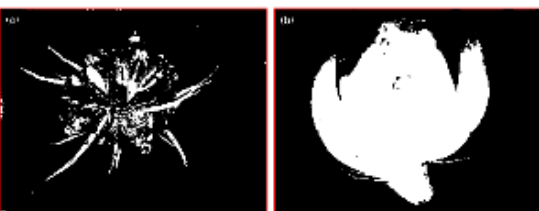


Figure 1: A 3D view of the sensor is overlaid on a 3D model of the section of a marine fin.

(b) Segmentation B



Figure 4. View of fish taken raised at 0°C for 4 months; the bones due endoskeleton remaining in the muscle as the fish is raised and with water salt extract.

[illegible]

relative abundance, resulting in a more stable biomass in the middle class than in the extreme class. According to the above reasoning, the abundance of the intermediate class should be the maximum among the three classes. [4]. While the intermediate abundance may vary with and developed along a succession, the number of species and the number of individuals in the intermediate class should be the maximum. In the present study, the number of species and the number of individuals in the intermediate class were the maximum, and the number of individuals in the extreme class was the minimum. The results of the present study are in good agreement with the above reasoning.

and expression of myelin oligodendrocyte glycoprotein (MOG) by fully myelinated but RFLP-resistant and myelin oligodendrocyte glycoprotein-negative strains of mice. In short-term myelination experiments in 20% acetone, oligodendrocytes may also have a role between the paranodal and the axolemma, perhaps by affecting the axolemma, as well as being directly involved in the myelination process. It is also possible that oligodendrocytes may have a role in the paranodal region.

[illegible]

Global oil supply
In the 1940s, great efforts were made to develop the supply of fuel by the American Government. The American supply of oil was not sufficient to meet the needs of the American military. High quality fuel was produced in Germany. Many thousands of tons of fuel were shipped to the American military. The American military was able to use this fuel to power its ships and aircraft. The American military was able to use this fuel to power its ships and aircraft. The American military was able to use this fuel to power its ships and aircraft.



TABLE 3. The Effect of the Optimization Strategy Used to Find the Best Solution of a Problem

(c) Segmentation C

Incorrect Page Segmentation



Figure 4 View of fruit shed naked (a) and (b) showing the complete androecial ring remaining within the tepals on the spike. (c) Fruit shed with all tepals attached.

(a) Input page segment

```

tv \_ - 14
la) \_ )âœ‰ , ` (C)
Â» ` fr /.1 Â·r , ` WM. âœ‰
7 if Â¢âœ‰ Â¥, ` { Ã©>Ã©_j l;I XII
K;~Â· - _V . Â· ` _ ~ M" JF Q} { fÂ·'; Ai ' V _ I \ g,
âœ‰ _ it ~ _y, i * t
f jgj. _ it ` l âœ‰ f
1 < .Â· âœ‰ Â· V Â·- âœ‰Â¢ âœ‰ âœ‰ âœ‰ F {U
Â· ~ . 7 ~ Ã© . i / I / _ âœ‰ * âœ‰ K Â»

```

Figure 4 View of fruit shed naked (a) and (b) showing the complete androecial ring remaining within the tepals on the spike. (c) Fruit shed with all tepals attached.

(b) OCR result

Incorrect Page Segmentation

into six (usually) additional seedless lobes of female mesocarp (supplementary carpels) surrounding the central fruit (figure 5a, b). These parthenocarpic lobes synthesised carotene and lipid and ripened in concert with the kernel-containing fertile ovary. This additional lipid-rich mesocarp offered a potential for high yields, and certain seedlings and at least one genetic line of oil palm was found which routinely produced such fruit. The promise of high yields from these so-called 'mantled' fruit was not fulfilled, however, perhaps because, although the fruit ripened, it was not shed. In the absence of the usual signal of the first few ripe fruits that fall to the ground, bunches on the mantled palms were left unheeded and the fruit were quick to rot on their spikelets.

Clonal oil palms

In the 1980s, great efforts were made to upgrade the yields of lipid by the introduction of clonal plant material raised by tissue culture from root or shoot fragments taken from elite, high quality, high lipid-producing palms. Many thousands of these clonally propagated individuals are now bearing fruit in plantation trials around the world and improved yields have resulted from these plantings. Certain of the tissue

culture procedures, involving the use of plant hormones in the media have, however, also led to a proportion of the palms showing sexual abnormalities that resemble the naturally occurring mantled fruit [4]. While the rudimentary androecium may form very well-developed lobes of supplementary carpels that extend the whole circlet of the androecial ring, sometimes, only one or two small lobes may arise while the remainder of the ring may be normal. In such fruit, abscission occurs normally at position 1, but at positions 2 and 3, cell separation takes place only where the rudimentary androecial ring has remained as aborted staminal tissue. Where the ring has differentiated into mesocarp tissue, the fruit remains attached to the bases of the tepals (figure 6a, b).

Control of this second stage of fruit abscission, and hence of fruit shedding, can therefore be manipulated by altering the developmental programme of the cells of the rudimentary androecium early in differentiation and before anthesis. Evidence from clonal propagation biotechnology now indicates that the levels of hormones used in tissue culture can determine the degree of mantling expressed by a palm several years later when it starts to flower. Because the condition does not show con-

into six (usually) additional seedless lobes culture procedures, involving the use of female mesocarp (supplementary plant hormones in the media have, however, also led to a proportion of the palms showing sexual abnormalities that resemble the naturally occurring mantled fruit [4]. While the rudimentary androecium may form very well-developed lobes of supplementary carpels that extend the whole circlet of the androecial ring, sometimes, only one or two small lobes may arise while the remainder of the ring may be normal. In such fruit, abscission occurs normally at position 1, but at positions 2 and 3, cell separation takes place only where the rudimentary androecial ring has remained as aborted staminal tissue. Where the ring has differentiated into mesocarp tissue, the fruit remains attached to the bases of the tepals (figure 6a, b).

Control of this second stage of fruit ab-

Clonal oil palms scission, and hence of fruit shedding, can

In the 1980s, great efforts were made to upgrade the yields of lipid by the introduction of clonal plant material raised by tissue culture from root or shoot fragments taken from elite, high quality, high lipid-producing palms. Many thousands of these clonally propagated individuals are now bearing fruit in plantation trials around the world and improved yields have resulted from these plantings. Certain of the tissue

(a) Input page segment

(b) OCR result

Page Segmentation Algorithms

- **Run-length Smearing Algorithm (1982)**
- **Recursive X-Y Cuts (1984)**
- **Whitespace Analysis (1994)**
- **Docstrum (1993)**
- **Voronoi (1998)**
- **RAST (2002) – *by Thomas Breuel***

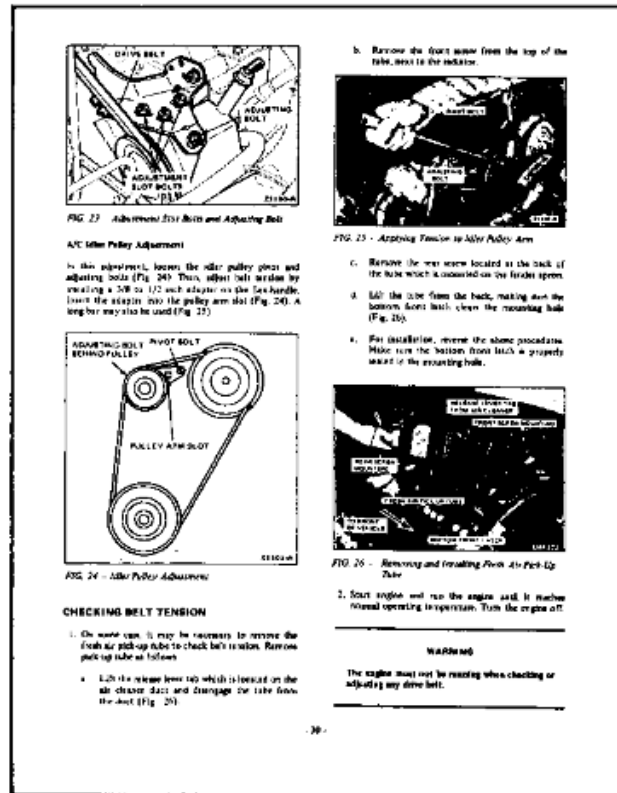
Run-Length Smearing Algorithm

- **Works on binary image**
- **White pixels represented by 0 and black by 1**
- **A binary sequence x is changed into y :**
 - 1's in x remain unchanged in y
 - 0's in x are changed to 1's in y if the number of adjacent 0's in x is less than or equal to a pre-defined threshold T .
- **This process is first repeated row-wise and then column-wise to get two distinct images**
- **The two images are combined using AND op.**

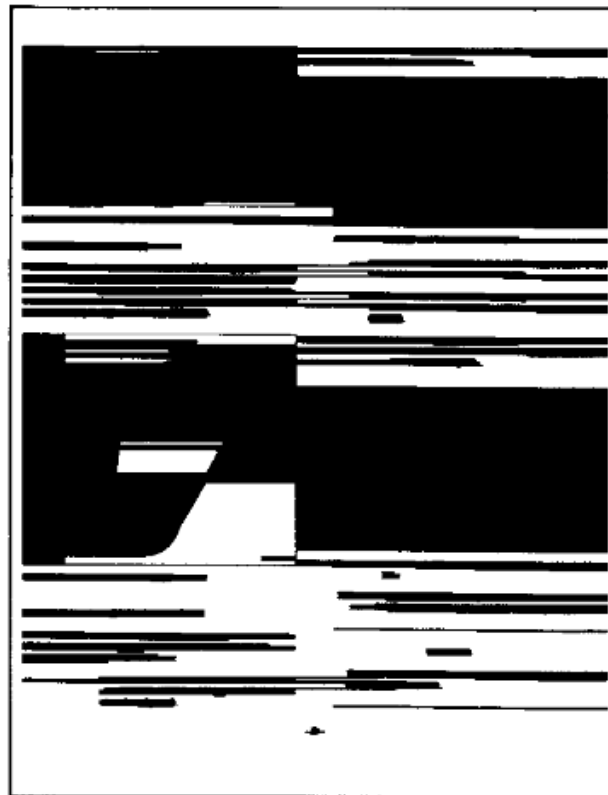
Run-Length Smearing Algorithm

- **A smooth final bitmap is obtained by again smearing in horizontal direction.**
- **Connected components in the final bitmap correspond to segments in the image.**

Run-Length Smearing Algorithm

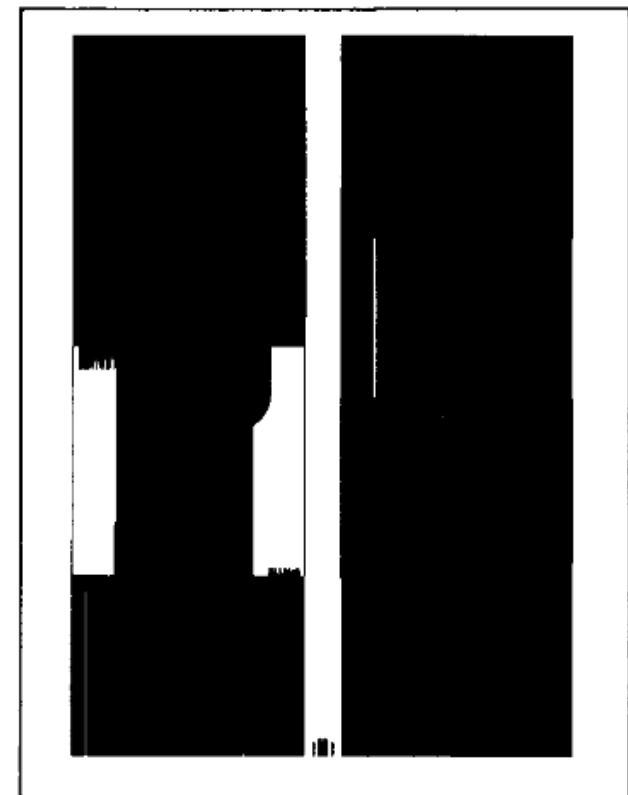


(a) Original Image



(b) Horizontally Smeared Image

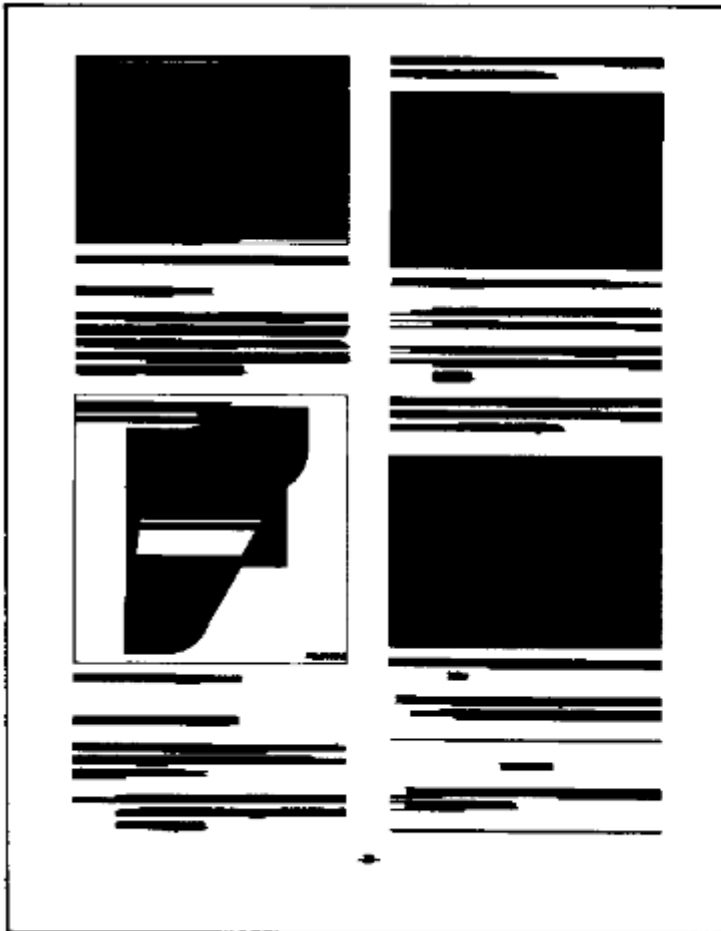
$$T_h = 300$$



(c) Vertically Smeared Image

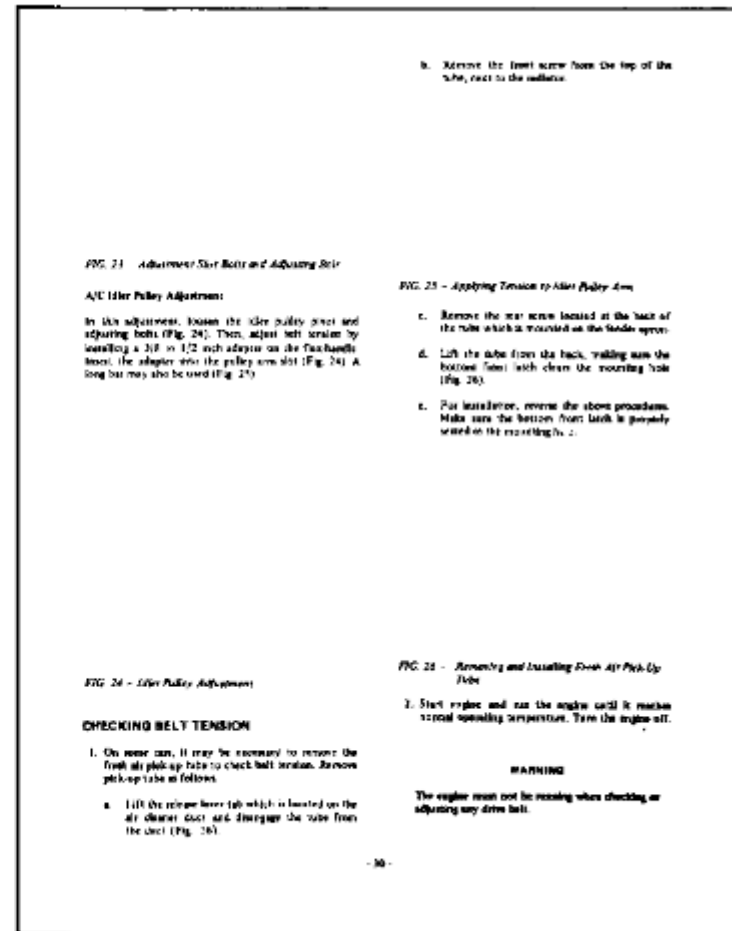
$$T_v = 500$$

Run-Length Smearing Algorithm



(d) Final Image after Smoothing

$$T_s = 30$$

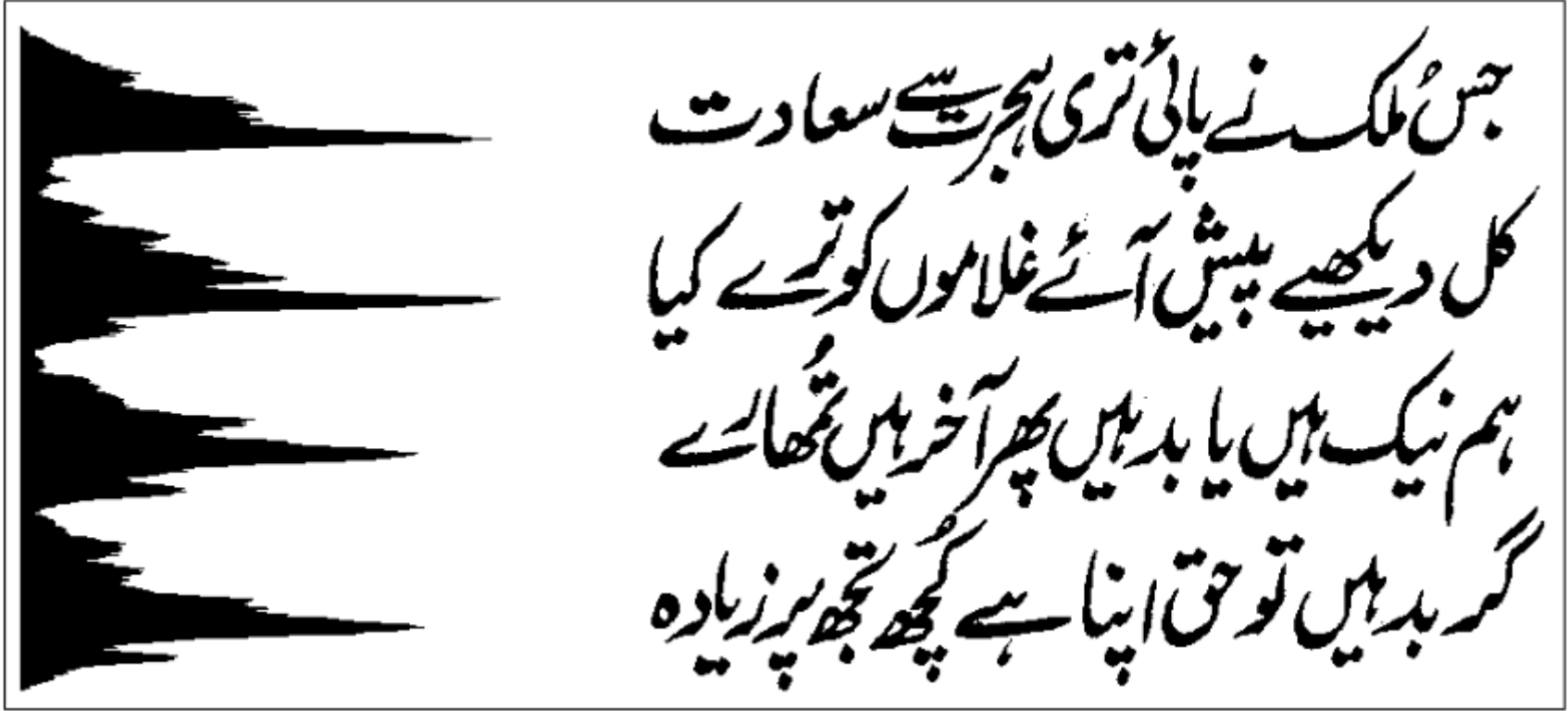


(e) Identified text regions

Recursive X-Y Cut Algorithm

- **Recursive analysis of projection profiles**
- **Projection profiles are obtained in two directions:**
 - **Horizontal:** Project the image on the **y-axis**.
 - The length of the projection is equal to the **height** of the image
 - The value at each index of projection is equal to the number of black pixels in that **row** of the image
 - **Vertical:** Project the image on the **x-axis**.
 - The length of the projection is equal to the **width** of the image
 - The value at each index of projection is equal to the number of black pixels in that **column** of the image

Horizontal Projection



Recursive X-Y Cut Algorithm

- **Recursive analysis of projection profiles**
- **Compute horizontal and vertical projection profiles of the image.**
- **Compute largest (zero-)valleys in the horizontal (v_y) and vertical (v_x) projections**
- **Split the image in the direction of larger valley into two images if $v_{larger} \geq T$**
- **Stop when the image can not be split further**

Things to remember

- **Otsu Thresholding**
- **Sauvola Thresholding**
- **Connected Component Analysis**
- **Run-Length Smearing Algorithm**
- **Recursive X-Y Cut Algorithm**