Document and Content Analysis

Lecture 06 - Document Image Analysis

Faisal Shafait

26.05.2011

How to convert a document image into editable text?

How to convert a document image into editable text?

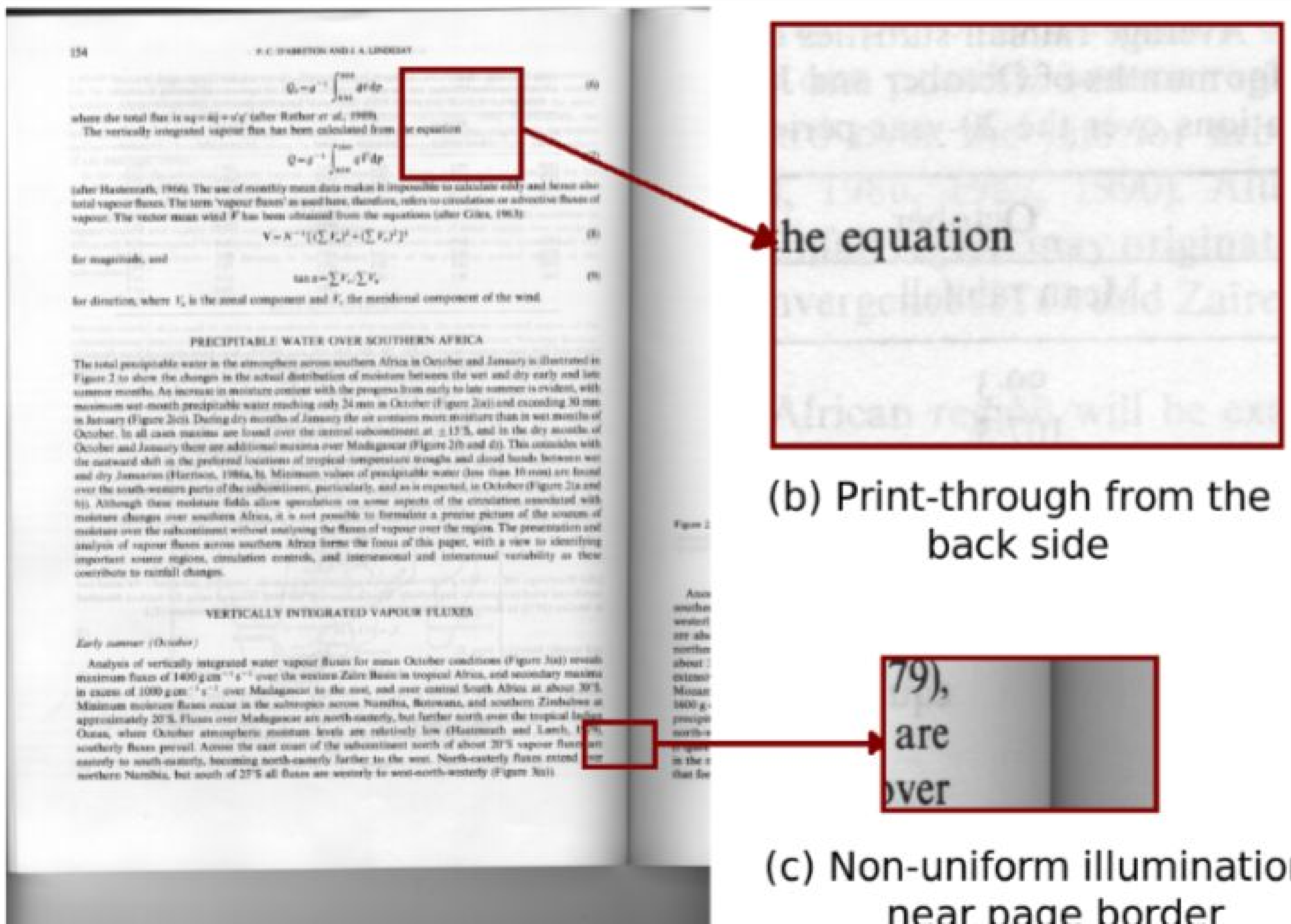
Optical Character Recognition (OCR)

How to convert a document image into editable text?

Optical Character Recognition (OCR)

We will learn how OCR works in the next four lectures!

A typical scanned book page



(a) Scanned book page

(b) Print-through from the back side

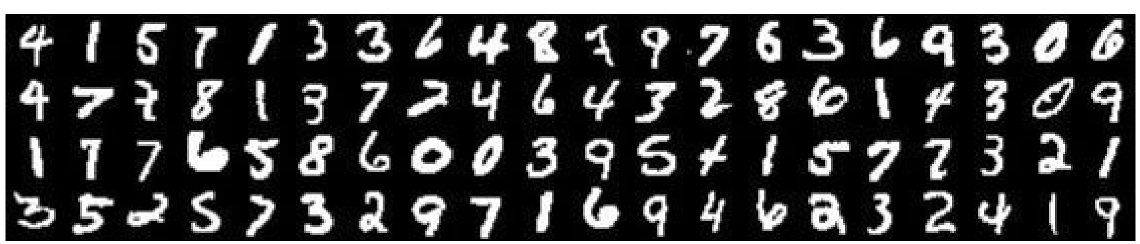
(c) Non-uniform illumination near page border

Character Recognition

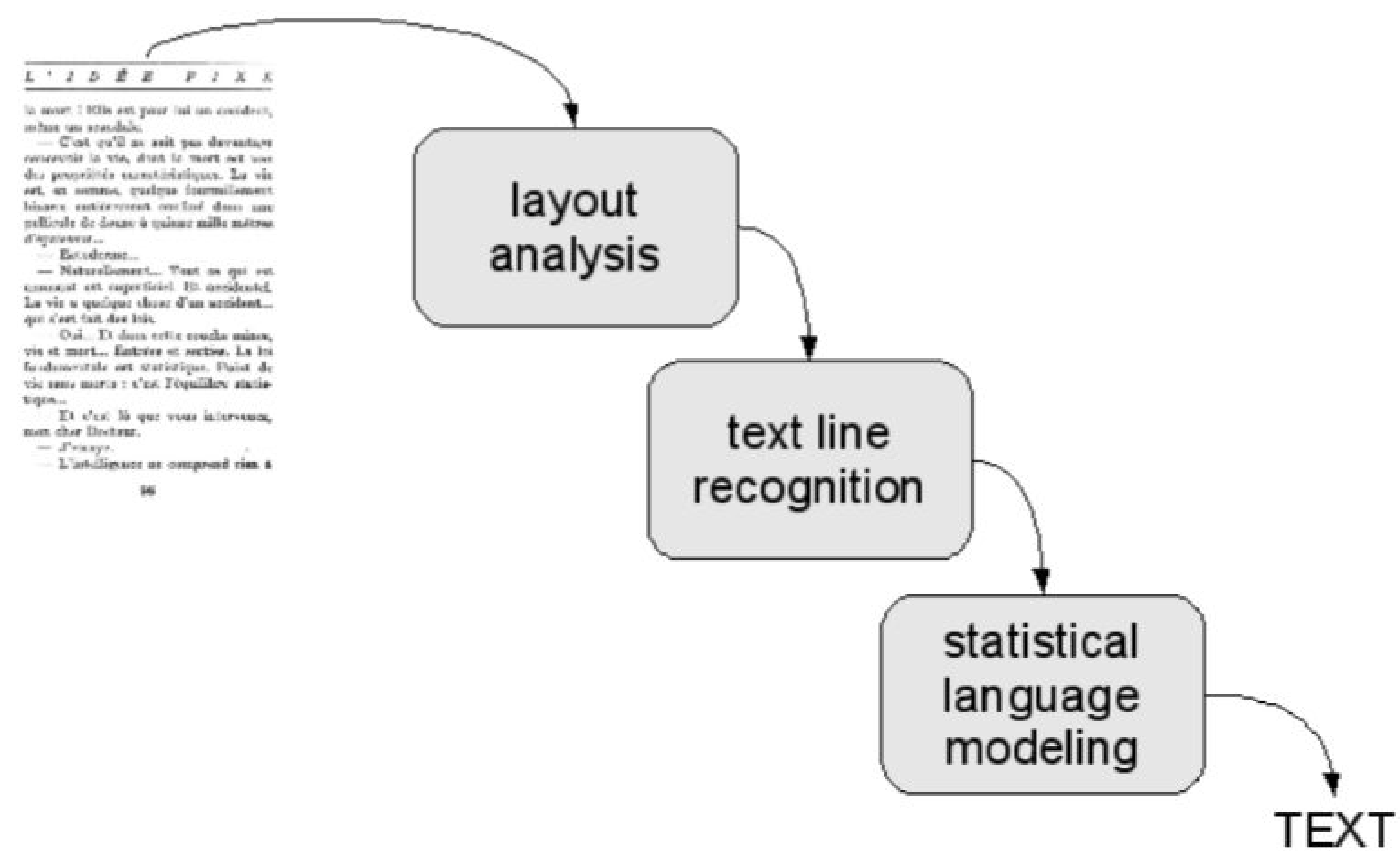
- Which character is this:
- What is this: Withsha its

Character Recognition

- Which character is this:
- What is this: Withsha its
- Isolated character recognition can be done as a standard pattern recognition problem, but a lot more needs to be done for a complete OCR system



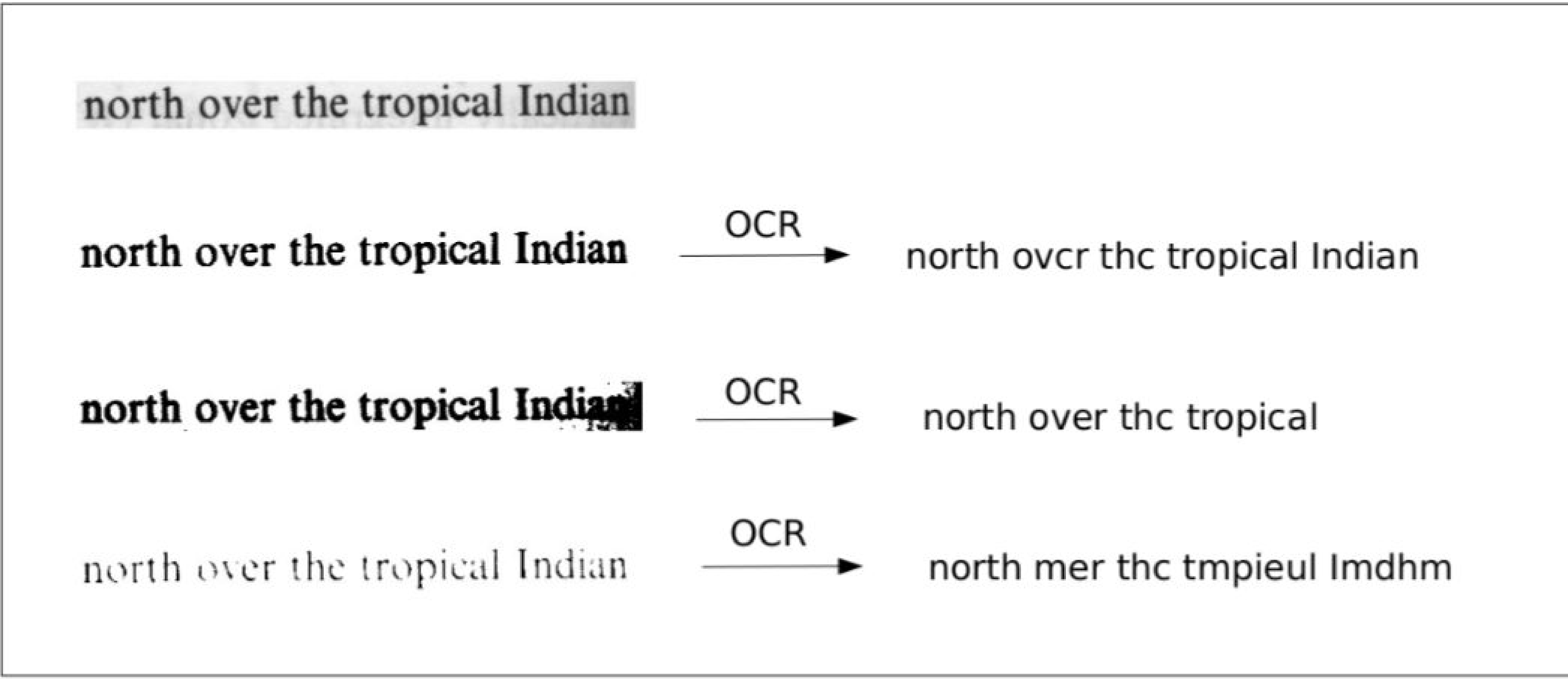
Flow chart for OCR



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



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

Global Binarization

- set

Otsu Global Thresholding

Let be the normalized histogram of the image

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

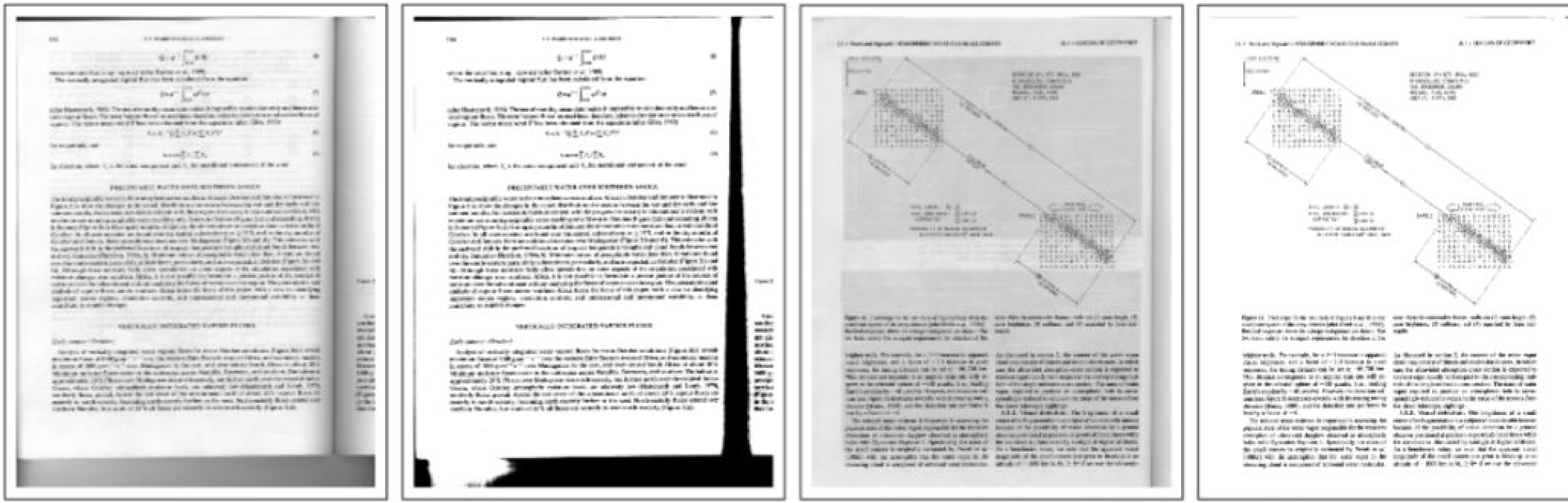
$$\mu_1 = \frac{1}{p_1} \sum_{g=0}^{T} g h_g$$

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

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

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

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 window around each pixel

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

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

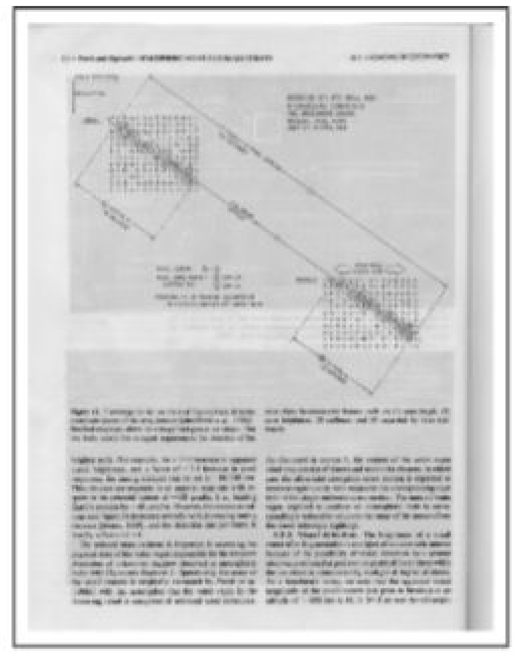
Niblack (1986): $t(x,y) = m(x,y) + ks(x,y)$

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

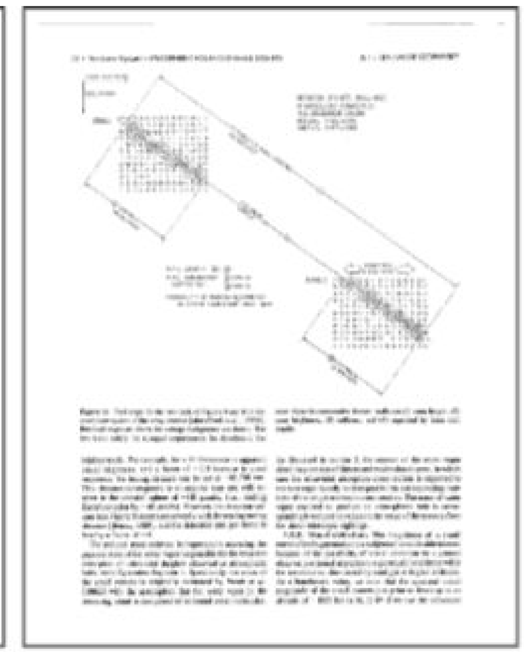
Sauvola Local Thresholding

(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

$$\begin{aligned} m(x,y) \quad = \quad & (I(x+w/2, y+w/2) + I(x-w/2, y-w/2) - \\ & I(x+w/2, y-w/2) - I(x-w/2, y+w/2)) \, / w^2 \end{aligned}$$

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

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

Connected Component Analysis

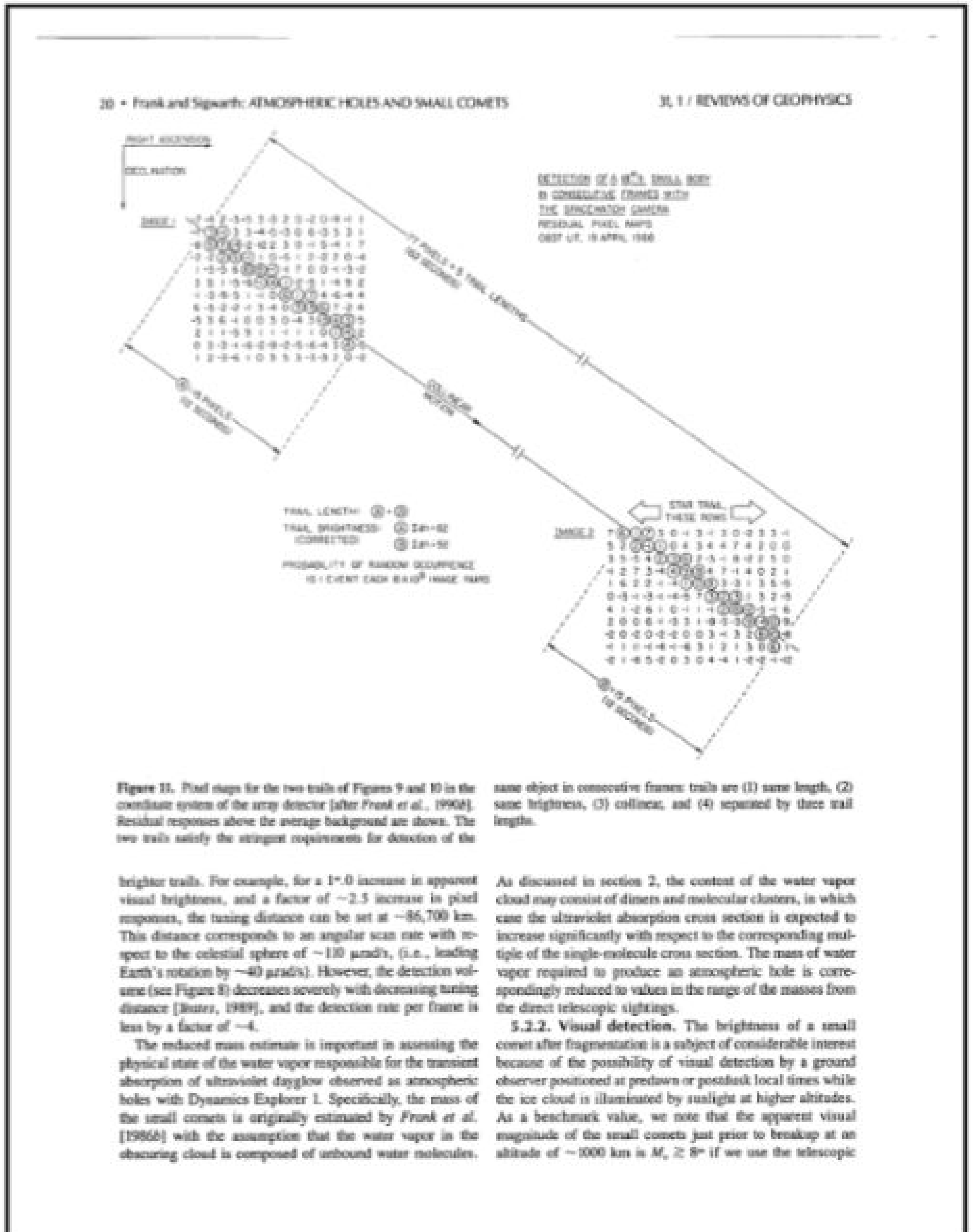


Figure 11. Pixel maps for the two trails of Figures 9 and 10 in the coordinate system of the array detector [after *Frank et al.*, 1990*b*]. Residual responses above the average background are shown. The two trails satisfy the stringent requirements for detection of the same object in consecutive frames: trails are (1) same length, (2) same brightness, (3) collinear, and (4) separated by three trail lengths.

brighter trails. For example, for a $1^m.0$ increase in apparent visual brightness, and a factor of ~2.5 increase in pixel responses, the tuning distance can be set at ~86,700 km. This distance corresponds to an angular scan rate with respect to the celestial sphere of ~110 μrad/s, (i.e., leading Earth's rotation by ~40 μrad/s). However, the detection volume (see Figure 8) decreases severely with decreasing tuning distance [*Yeates*, 1989], and the detection rate per frame is less by a factor of ~4.

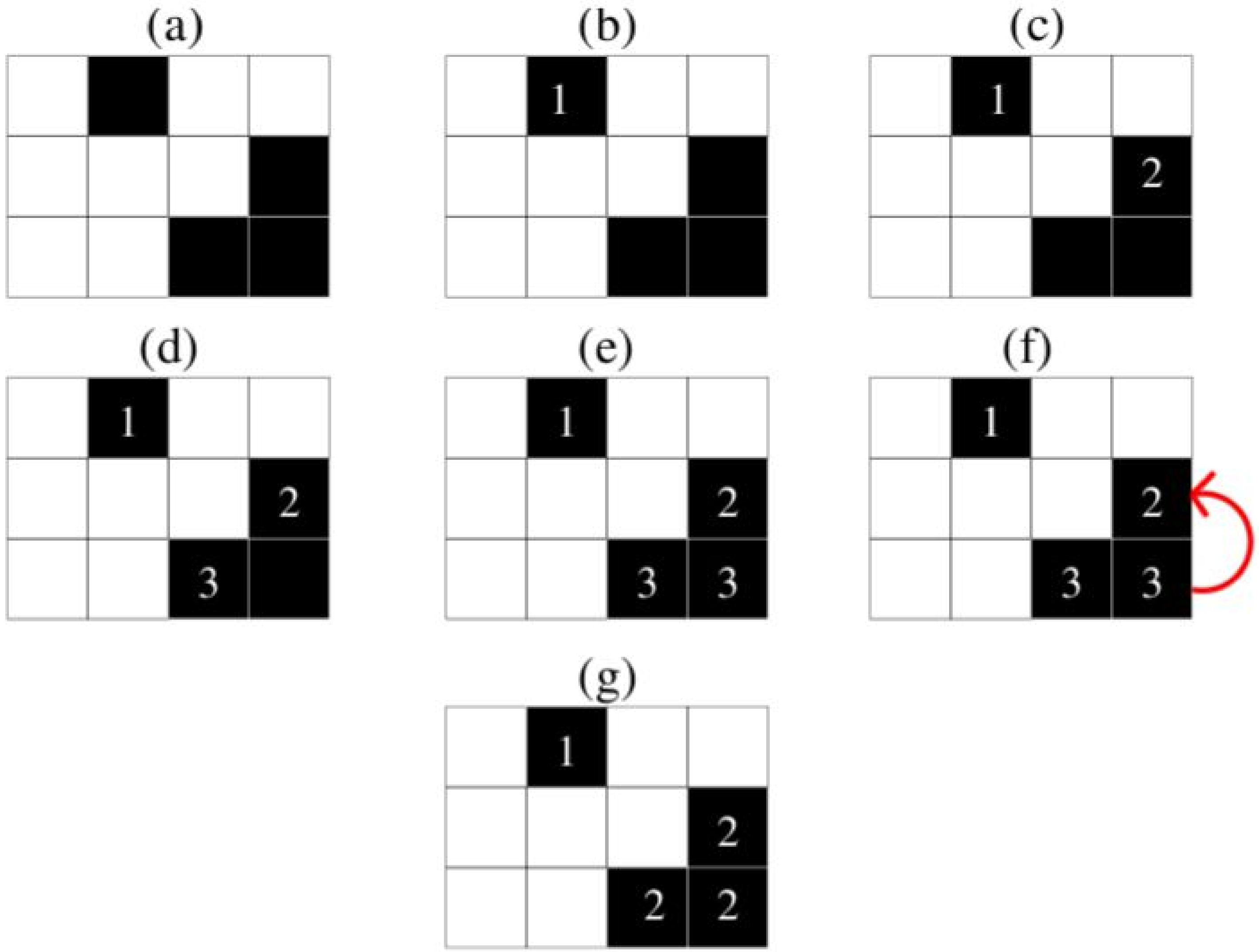
The reduced mass estimate is important in assessing the physical state of the water vapor responsible for the transient absorption of ultraviolet dayglow observed as atmospheric holes with Dynamics Explorer 1. Specifically, the mass of the small comets is originally estimated by *Frank et al.* [1986*b*] with the assumption that the water vapor in the obscuring cloud is composed of unbound water molecules. As discussed in section 2, the content of the water vapor cloud may consist of dimers and molecular clusters, in which case the ultraviolet absorption cross section is expected to increase significantly with respect to the corresponding multiple of the single-molecule cross section. The mass of water vapor required to produce an atmospheric hole is correspondingly reduced to values in the range of the masses from the direct telescopic sightings.

5.2.2. Visual detection. The brightness of a small comet after fragmentation is a subject of considerable interest because of the possibility of visual detection by a ground observer positioned at predawn or postdusk local times while the ice cloud is illuminated by sunlight at higher altitudes. As a benchmark value, we note that the apparent visual magnitude of the small comets just prior to breakup at an altitude of ~1000 km is $M_v \geq 8^m$ if we use the telescopic

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

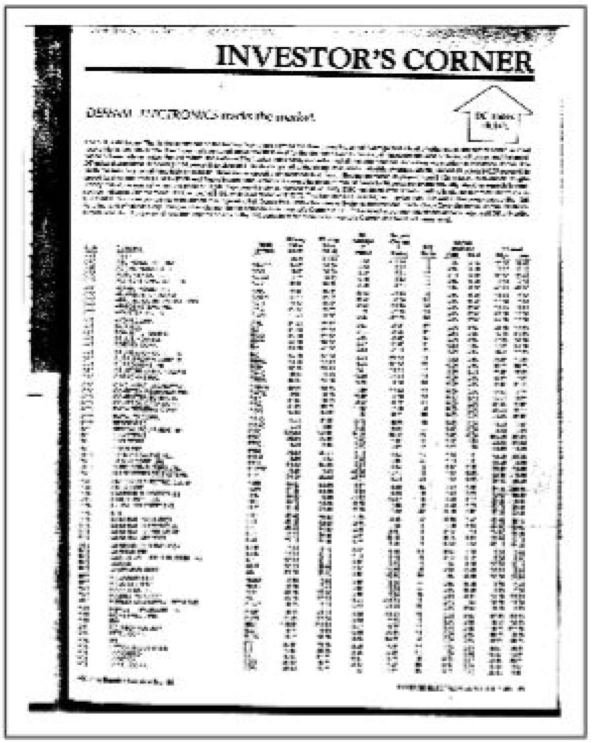
Connected Component Analysis Example

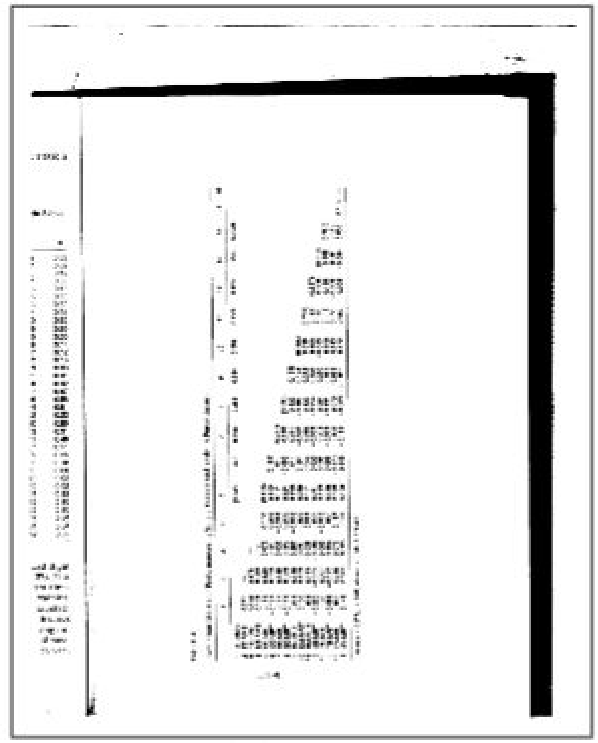


Other Pre-processing Tasks

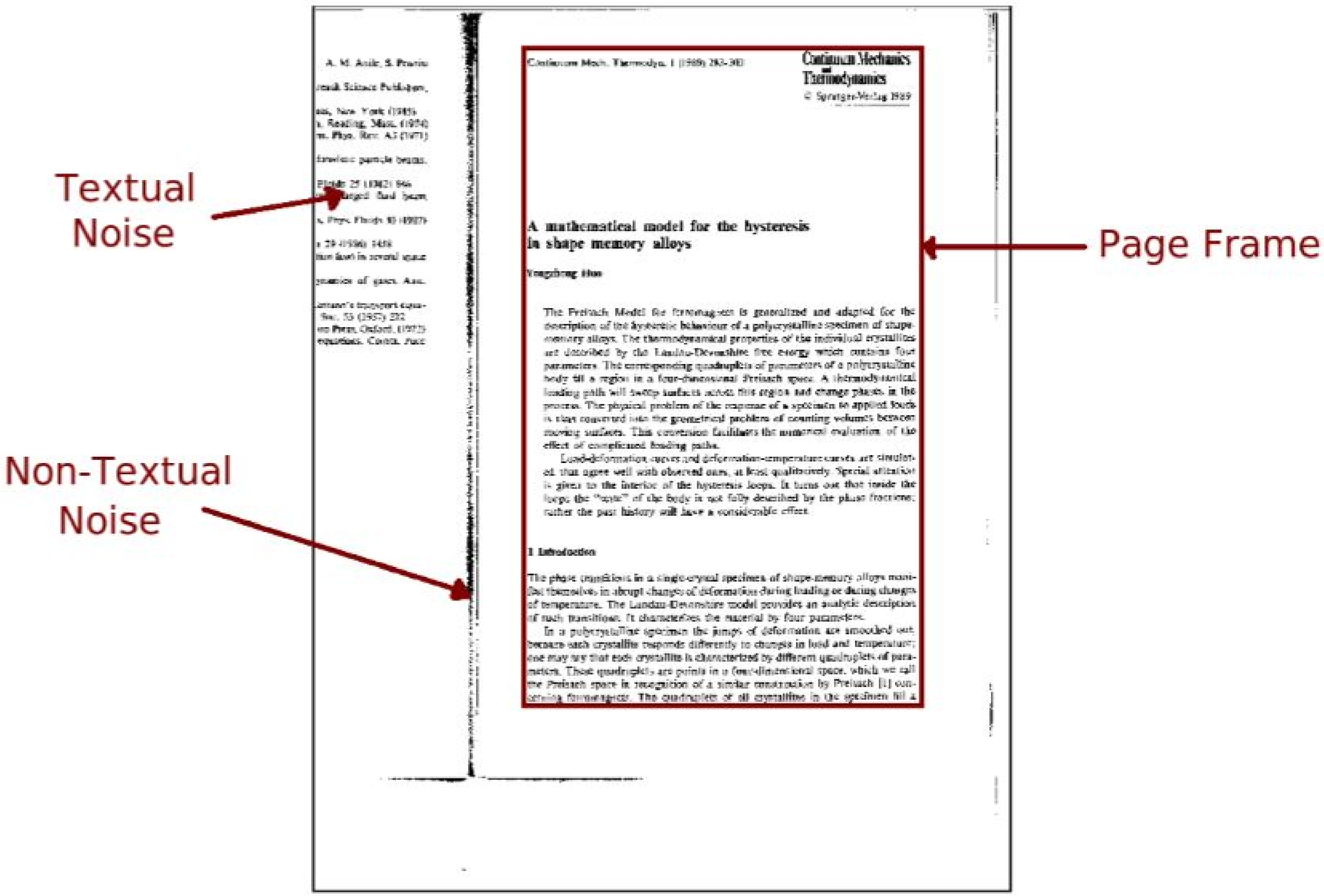
- Orientation detection
- Marginal noise removal
- Skew correction

Orientation Detection

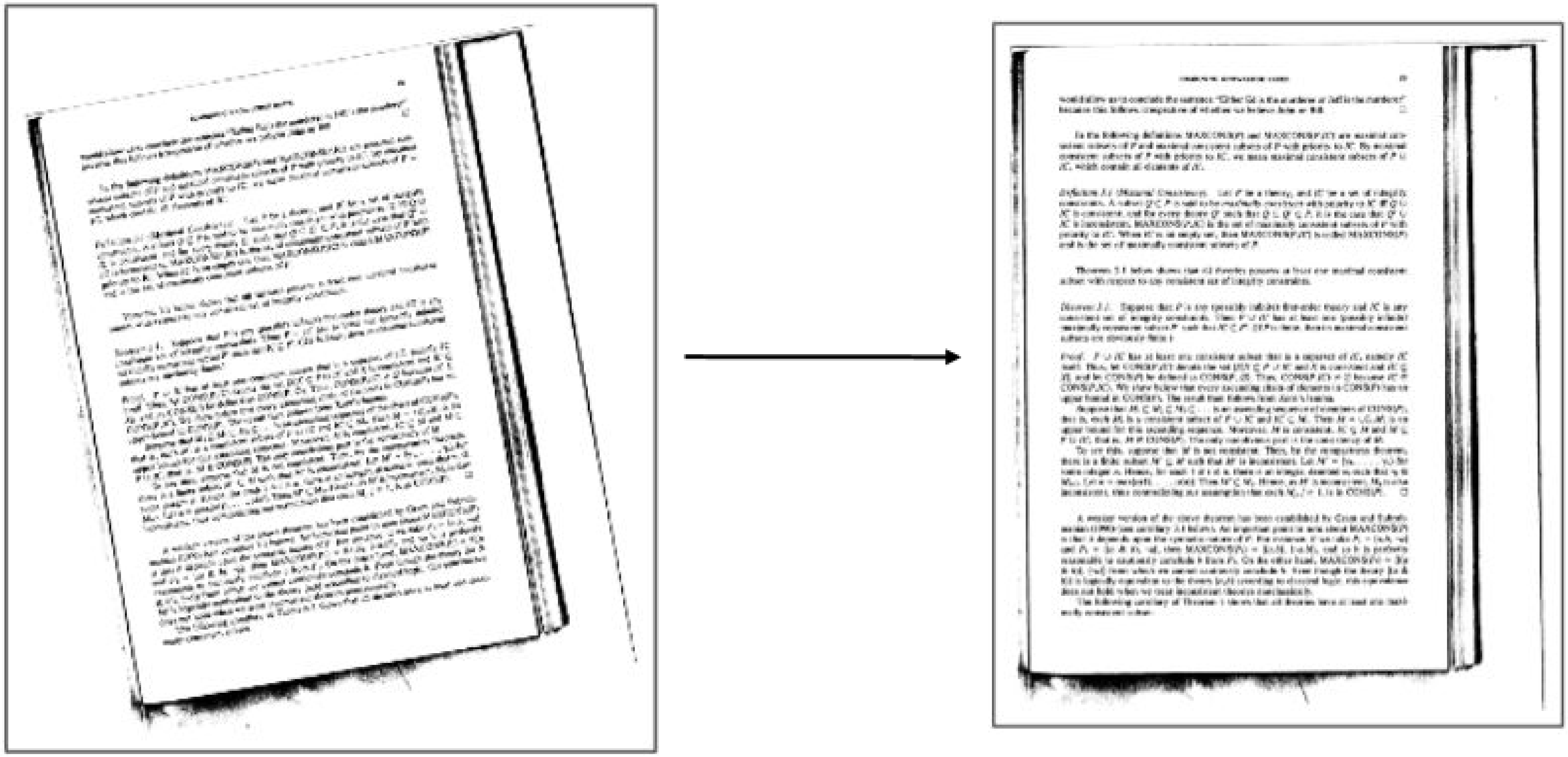

INVESTOR'S CORNER



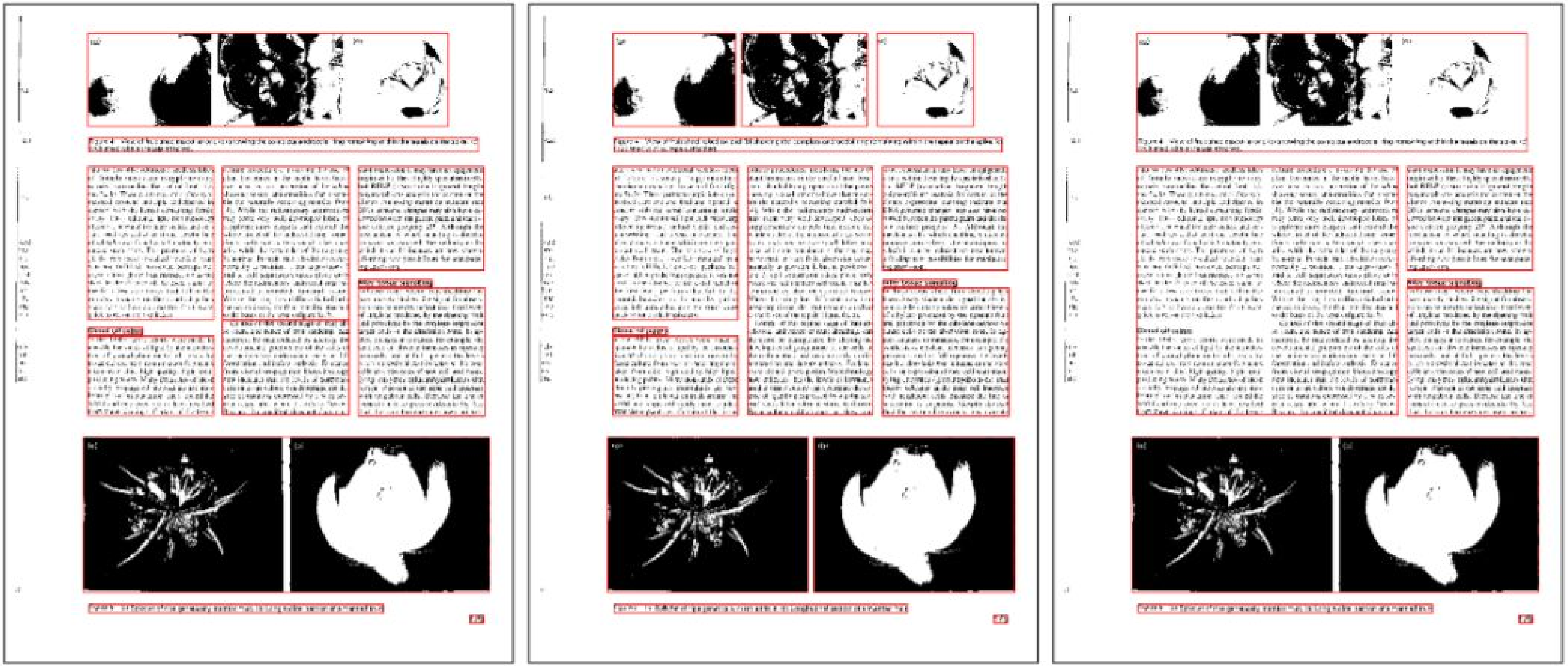
Marginal Noise Removal



Skew Correction

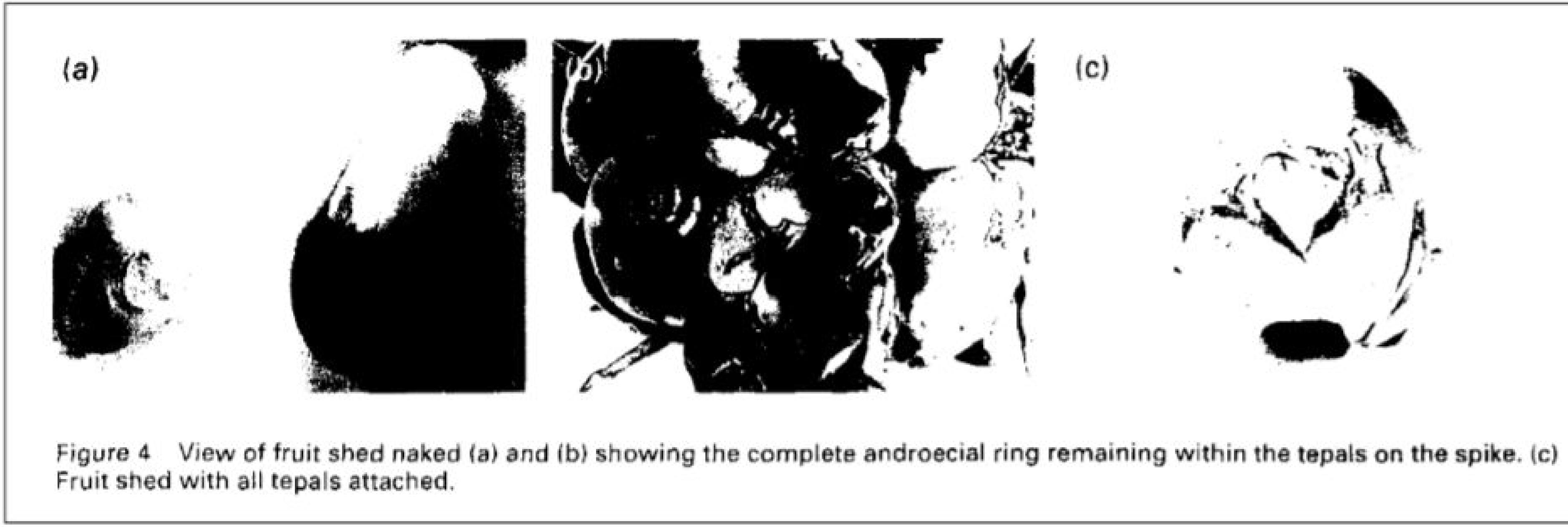


Page Segmentation



(a) Segmentation A (b) Segmentation B (c) Segmentation C

Incorrect Page Segmentation



(a) Input page segment

```
tv \_ _. -1 4
la)  `  )âoњ  ,  `  (C)
Â»        ` fr /.1 Â·r , ` WM.  âoњ
7 if Â¢âo Â¥,  `    _`  {    Ã©>Ã©_j l; I     X I I
K;~Â·  -  _V . Â·  ` _ ~ M" JF Q} {  fÂ·',; Ai ' V _ I \' g,
 âo   _ it ~ _y, i * t
  f jgj .   __it  ` l âoњ   f
  1< .Â· âoњ Â·. V Â·- âoњÂ¢  âoњ âoњ âoњ F {U
Â·~ . 7 ~   Ã© . i/`I  / _âoњ * âoњ K Â»
Figure 4 View of fruit shed naked la) and lb) 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-

(a) Input page segment

into six (usually) additional seedless lobes culture procedures. involving the use of
of female mesocarp (supplementary plant hormones in the media have. how-
carpels) surrounding the central fruit (fig- ever. also led to a proportion of the palms
ure 5a. b). These parthenocarpic lobes syn- showing sexual abnormalities that resem-
thesised carotene and lipid and ripened in ble the naturally occurring mantled fruit
concert with the kernel-containing fertile [4]. While the rudimentary androecium
ovary. This additional lipidâ„rich mesocarp may form very well-developed lobes of
offered a potential for high yields. and cer- supplementary carpels that extend the
tain seedlings and at least one genetic line whole circlet of the androecial ring. some-
of oil palm was found which routinely pro- times. only one or two small lobes may
duced such fruit. The protnise of high arise while the remainder of the ring may
yields from these so-called `mantled` fruit be normal. In such fruit. ahscission occurs
was not fulfilled. however. perhaps be- normally at position l. but at positions 2
cause. although the fruit ripened. it was not and 3. cell separation takes place only
shed. ln the absence of the usual signal of where the rudimentary androecial ring has
the first few ripe fruits that fall to the remained as aborted staminal tissue.
ground. bunches on the inandled paltns Where the ring has differentiated into
were left unheeded and the fruit were tnesocarp tissue. the fruit remains attached
quick to rot on their spikelets. 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 198()s. great efforts were tnade to therefore be manipulated by altering the
upgrade the yields of lipid by the introduc- developmental programme of the cells of
tion of clonal plant material raised by the rudimentary androecium early in dif-
tissue culture from root or shoot fragments ferentiation and before anthesis. Evidence
taken from elite. high quality. high lipid- from clonal propagation biotechnology
producing palms. Many thousands of these now indicates that the levels of hormones
clonally propagated individuals are now used in tissue culture can determine the de-
bearing fruit in plantation trials around the gree of mantling expressed by a palm sev-
world and improved yields have resulted eral years later when it starts to flower.
from these plantings. Certain of the tissue Because the condition does not show con-

(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*

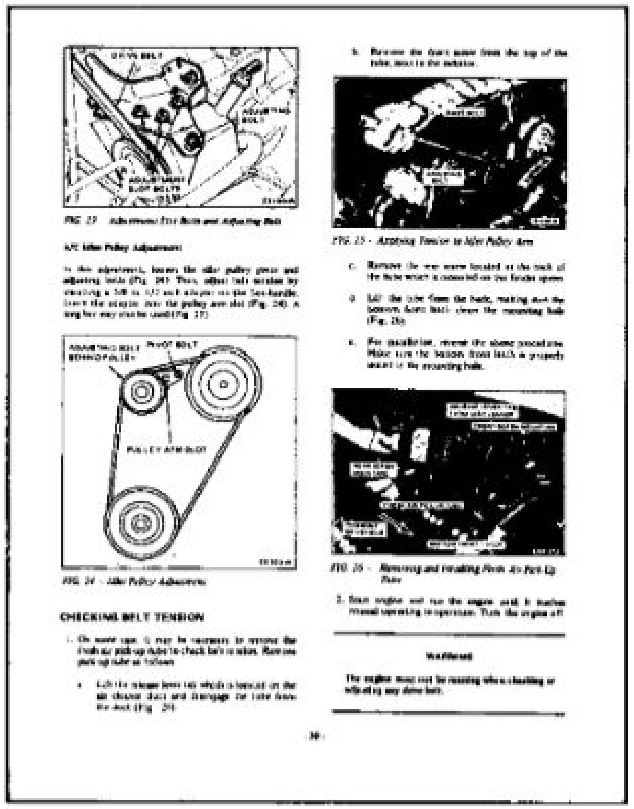
Run-Length Smearing Algorithm

- 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

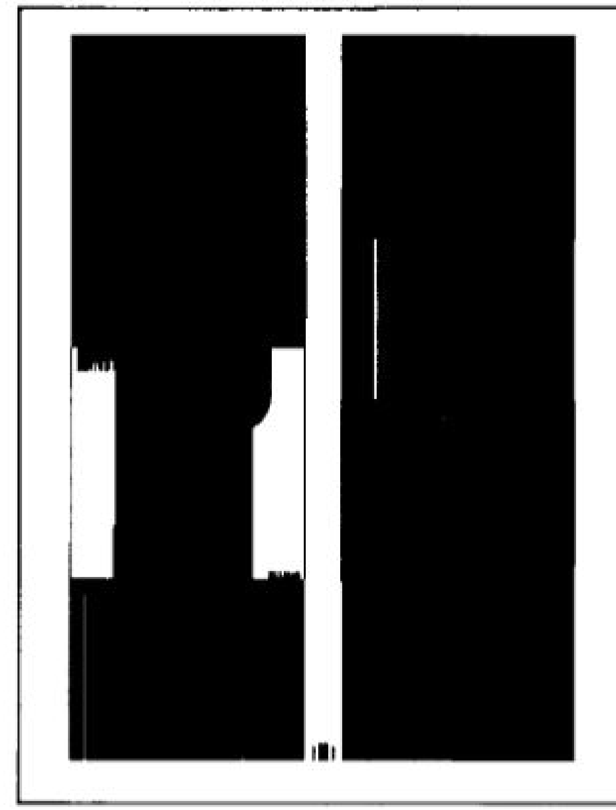
- A smooth final bitmap is obtained by again smearing in horizontal direction.
- Connected components in the final bitmap correspond to segments in the image.

Run-Length Smearing Algorithm



(a) Original Image

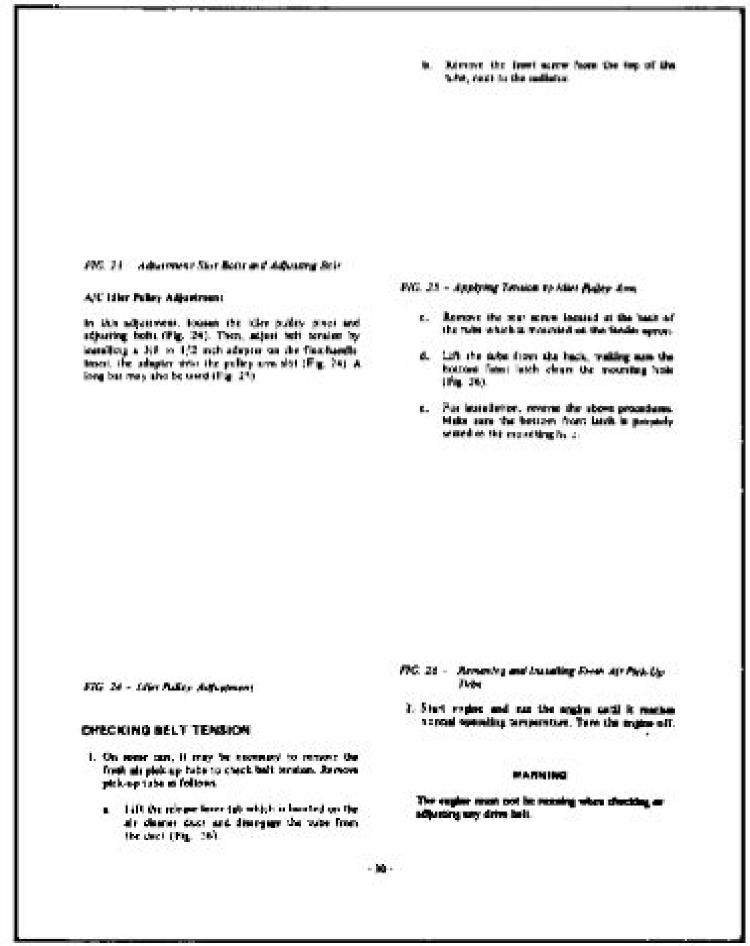
(b) Horizontally Smeared Image



(c) Vertically Smeared Image

Run-Length Smearing Algorithm

(d) Final Image after Smoothing

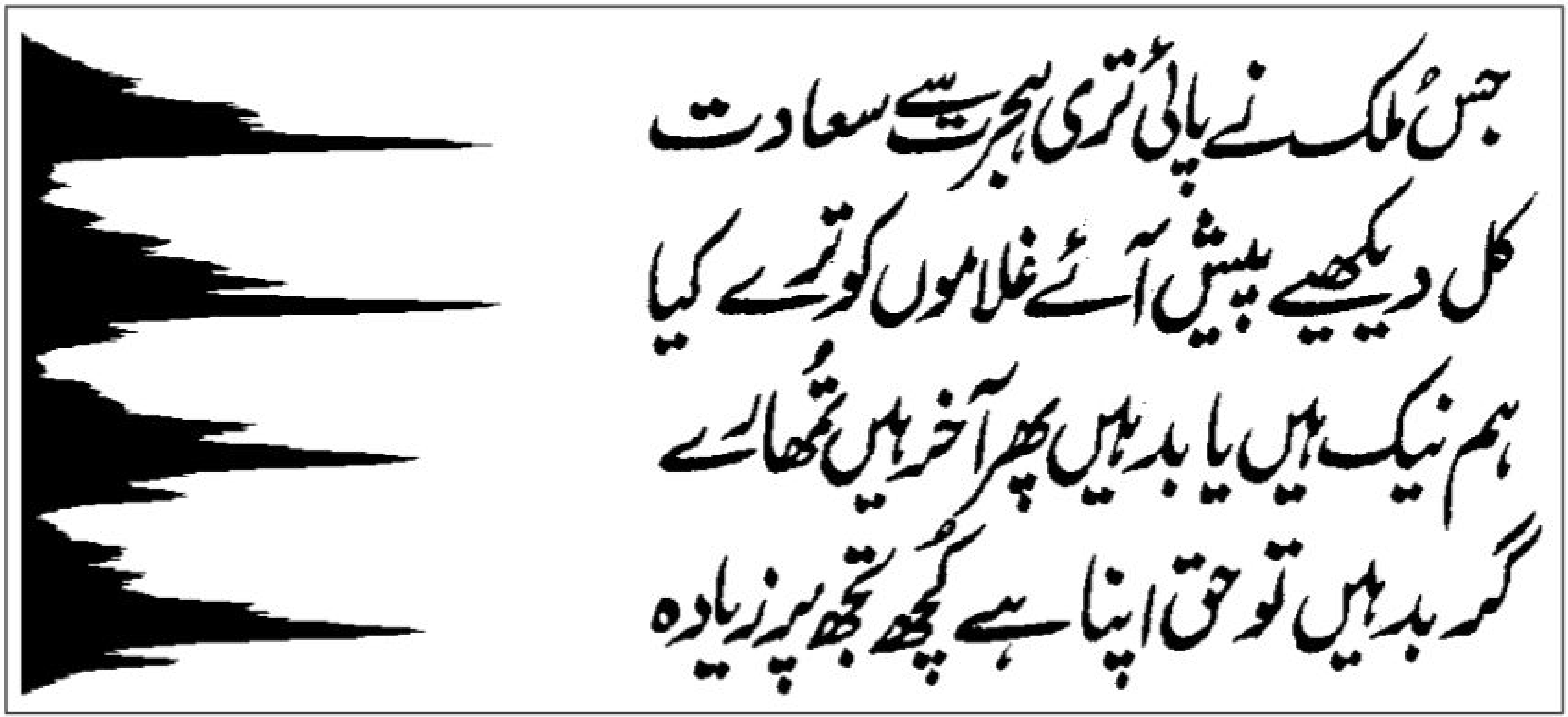


(e) Identified text regions

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 and vertical projections
- Split the image in the direction of larger valley into two images if
- 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