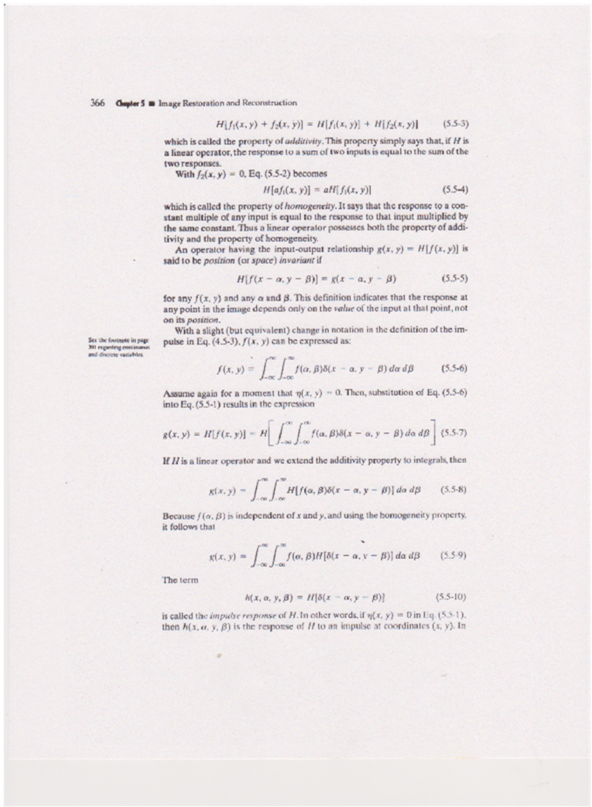
Document Layout Analysis

By: Garrett Hoch

Document Layout Analysis Overview

- What is Document Layout Analysis
  - Geometric layout analysis
  - Logical layout analysis
- Why is it useful?
  - Done before OCR
  - Gives meaning to text
  - Databases
- Algorithm: Docstrum



366 Chapter 5 ■ Image Restoration and Reconstruction

$$H[f_1(x, y) + f_2(x, y)] = H[f_1(x, y)] + H[f_2(x, y)] \tag{5.5-3}$$

which is called the property of *additivity*. This property simply says that, if H is a linear operator, the response to a sum of two inputs is equal to the sum of the two responses.

With $f_2(x, y) = 0$, Eq. (5.5-2) becomes

$$H[af_1(x, y)] = aH[f_1(x, y)] \tag{5.5-4}$$

which is called the property of *homogeneity*. It says that the response to a constant multiple of any input is equal to the response to that input multiplied by the same constant. Thus a linear operator possesses both the property of additivity and the property of homogeneity.

An operator having the input-output relationship $g(x, y) = H[f(x, y)]$ is said to be *position* (or *space*) *invariant* if

$$H[f(x - \alpha, y - \beta)] = g(x - \alpha, y - \beta) \tag{5.5-5}$$

for any $f(x, y)$ and any α and β. This definition indicates that the response at any point in the image depends only on the *value* of the input at that point, not on its *position*.

With a slight (but equivalent) change in notation in the definition of the impulse in Eq. (4.5-3), $f(x, y)$ can be expressed as:

See the footnote in page 391 regarding continuous and discrete variables

$$f(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta \tag{5.5-6}$$

Assume again for a moment that $\eta(x, y) = 0$. Then, substitution of Eq. (5.5-6) into Eq. (5.5-1) results in the expression

$$g(x, y) = H[f(x, y)] = H\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta\right] \tag{5.5-7}$$

If H is a linear operator and we extend the additivity property to integrals, then

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H[f(\alpha, \beta)\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \tag{5.5-8}$$

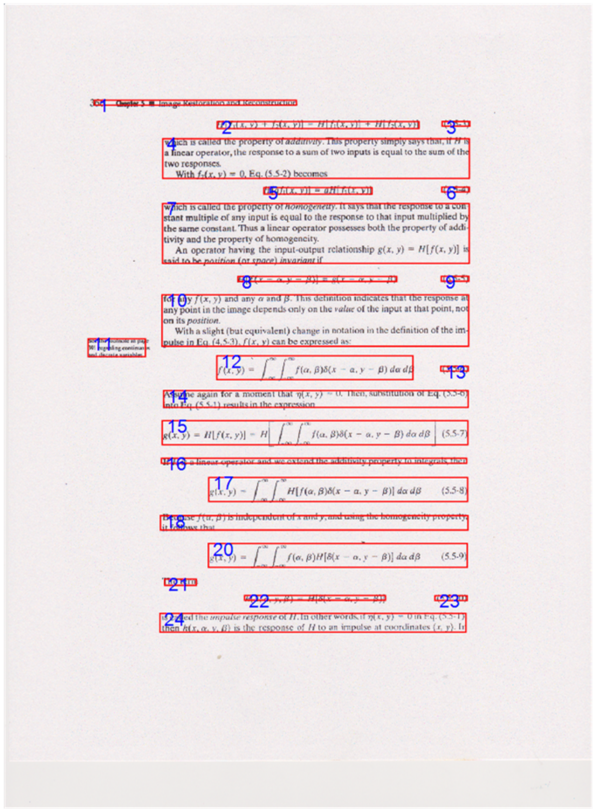
Because $f(\alpha, \beta)$ is independent of x and y, and using the homogeneity property, it follows that

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)H[\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \tag{5.5-9}$$

The term

$$h(x, \alpha, y, \beta) = H[\delta(x - \alpha, y - \beta)] \tag{5.5-10}$$

is called the *impulse response* of H. In other words, if $\eta(x, y) = 0$ in Eq. (5.5-1), then $h(x, \alpha, y, \beta)$ is the response of H to an impulse at coordinates (x, y). In

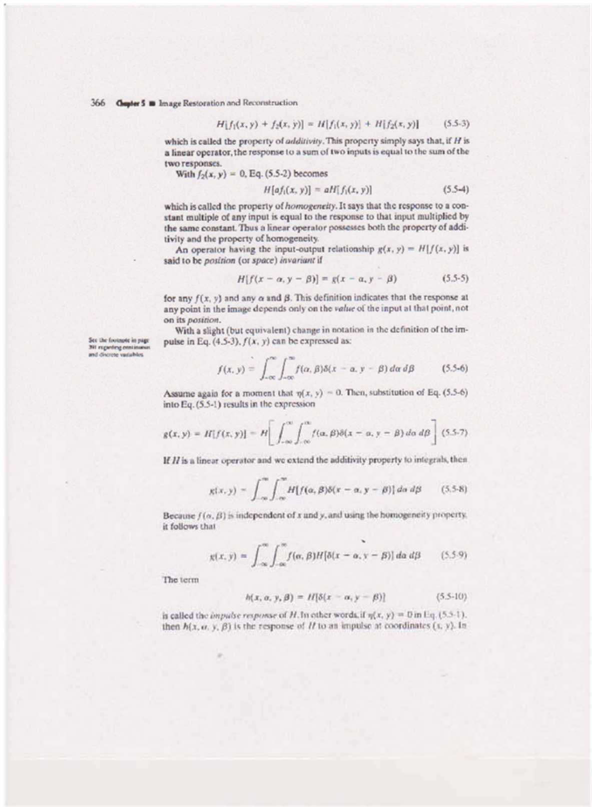


Algorithm - Docstrum

1. Preprocessing
2. Detect centroids
3. Determine k nearest neighbors
4. Estimate skew of image
5. Estimate in line and between line spacing
6. Find lines of text
7. Find blocks of text
8. Bounding box calculation

1. Preprocessing

- Convert image to gray scale
- Threshold
- Salt and pepper noise
  - Median Filter
- Morphological opening



366 Chapter 5 ■ Image Restoration and Reconstruction

$$H[f_1(x, y) + f_2(x, y)] = H[f_1(x, y)] + H[f_2(x, y)] \qquad (5.5\text{-}3)$$

which is called the property of *additivity*. This property simply says that, if H is a linear operator, the response to a sum of two inputs is equal to the sum of the two responses.

With $f_2(x, y) = 0$, Eq. (5.5-2) becomes

$$H[af_1(x, y)] = aH[f_1(x, y)] \qquad (5.5\text{-}4)$$

which is called the property of *homogeneity*. It says that the response to a constant multiple of any input is equal to the response to that input multiplied by the same constant. Thus a linear operator possesses both the property of additivity and the property of homogeneity.

An operator having the input-output relationship $g(x, y) = H[f(x, y)]$ is said to be *position* (or *space*) *invariant* if

$$H[f(x - \alpha, y - \beta)] = g(x - \alpha, y - \beta) \qquad (5.5\text{-}5)$$

for any $f(x, y)$ and any α and β. This definition indicates that the response at any point in the image depends only on the *value* of the input at that point, not on its *position*.

With a slight (but equivalent) change in notation in the definition of the impulse in Eq. (4.5-3), $f(x, y)$ can be expressed as:

See the footnote in page 391 regarding continuous and discrete variables.

$$f(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta \qquad (5.5\text{-}6)$$

Assume again for a moment that $\eta(x, y) = 0$. Then, substitution of Eq. (5.5-6) into Eq. (5.5-1) results in the expression

$$g(x, y) = H[f(x, y)] = H\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta\right] \qquad (5.5\text{-}7)$$

If H is a linear operator and we extend the additivity property to integrals, then

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H[f(\alpha, \beta)\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}8)$$

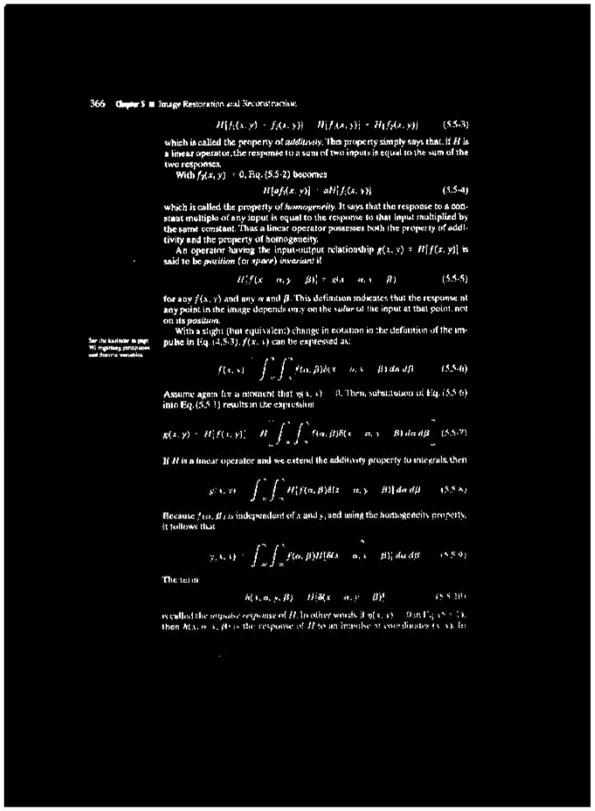
Because $f(\alpha, \beta)$ is independent of x and y, and using the homogeneity property, it follows that

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)H[\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}9)$$

The term

$$h(x, \alpha, y, \beta) = H[\delta(x - \alpha, y - \beta)] \qquad (5.5\text{-}10)$$

is called the *impulse response* of H. In other words, if $\eta(x, y) = 0$ in Eq. (5.5-1), then $h(x, \alpha, y, \beta)$ is the response of H to an impulse at coordinates (x, y). In



366 Chapter 5 ■ Image Restoration and Reconstruction

$$H[f_1(x, y) + f_2(x, y)] = H[f_1(x, y)] + H[f_2(x, y)] \qquad (5.5\text{-}3)$$

which is called the property of *additivity*. This property simply says that, if H is a linear operator, the response to a sum of two inputs is equal to the sum of the two responses.

With $f_2(x, y) = 0$, Eq. (5.5-2) becomes

$$H[af_1(x, y)] = aH[f_1(x, y)] \qquad (5.5\text{-}4)$$

which is called the property of *homogeneity*. It says that the response to a constant multiple of any input is equal to the response to that input multiplied by the same constant. Thus a linear operator possesses both the property of additivity and the property of homogeneity.

An operator having the input-output relationship $g(x, y) = H[f(x, y)]$ is said to be *position* (or *space*) *invariant* if

$$H[f(x - \alpha, y - \beta)] = g(x - \alpha, y - \beta) \qquad (5.5\text{-}5)$$

for any $f(x, y)$ and any α and β. This definition indicates that the response at any point in the image depends only on the *value* of the input at that point, not on its *position*.

With a slight (but equivalent) change in notation in the definition of the impulse in Eq. (4.5-3), $f(x, y)$ can be expressed as:

See the footnote in page 391 regarding continuous and discrete variables.

$$f(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta \qquad (5.5\text{-}6)$$

Assume again for a moment that $\eta(x, y) = 0$. Then, substitution of Eq. (5.5-6) into Eq. (5.5-1) results in the expression

$$g(x, y) = H[f(x, y)] = H\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta\right] \qquad (5.5\text{-}7)$$

If H is a linear operator and we extend the additivity property to integrals, then

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H[f(\alpha, \beta)\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}8)$$

Because $f(\alpha, \beta)$ is independent of x and y, and using the homogeneity property, it follows that

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)H[\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}9)$$

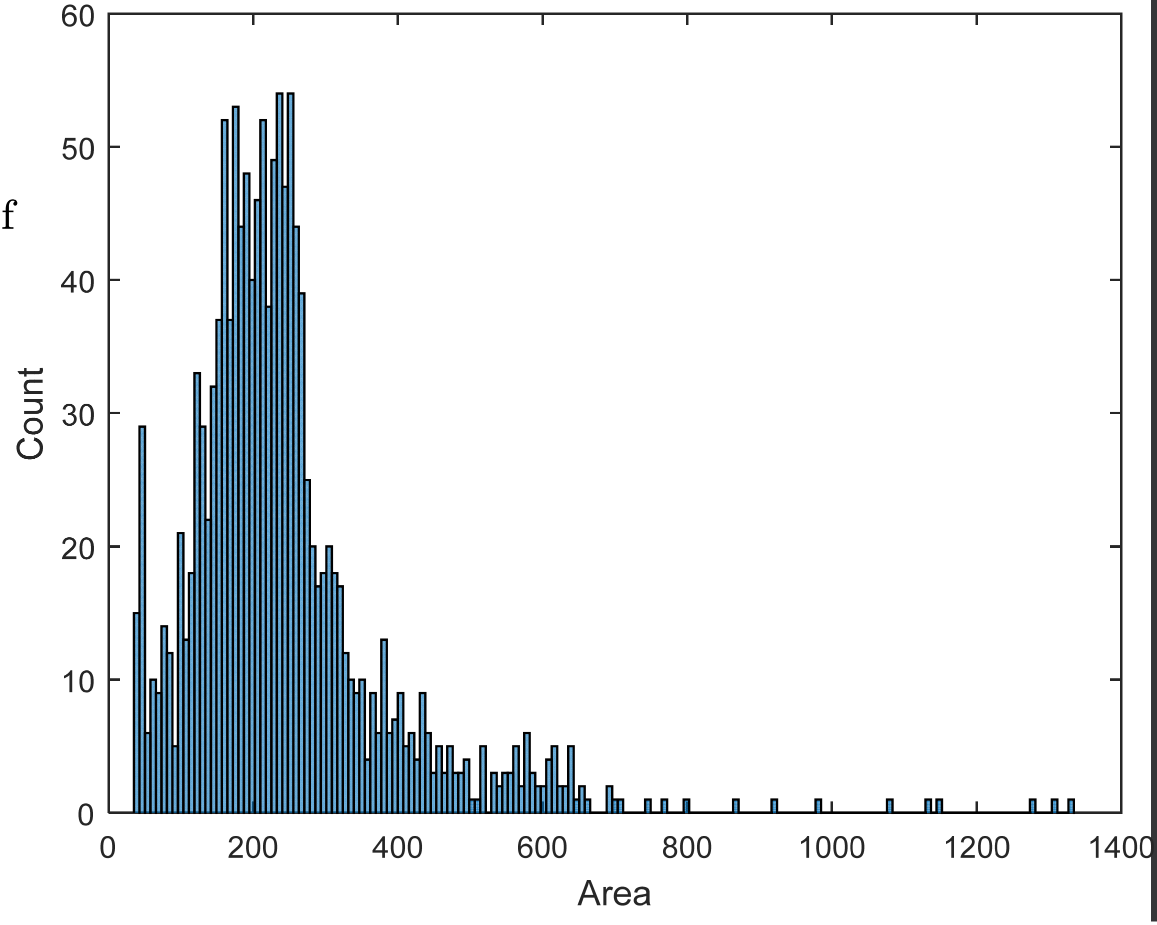
The term

$$h(x, \alpha, y, \beta) = H[\delta(x - \alpha, y - \beta)] \qquad (5.5\text{-}10)$$

is called the *impulse response* of H. In other words, if $\eta(x, y) = 0$ in Eq. (5.5-1), then $h(x, \alpha, y, \beta)$ is the response of H to an impulse at coordinates (x, y). In

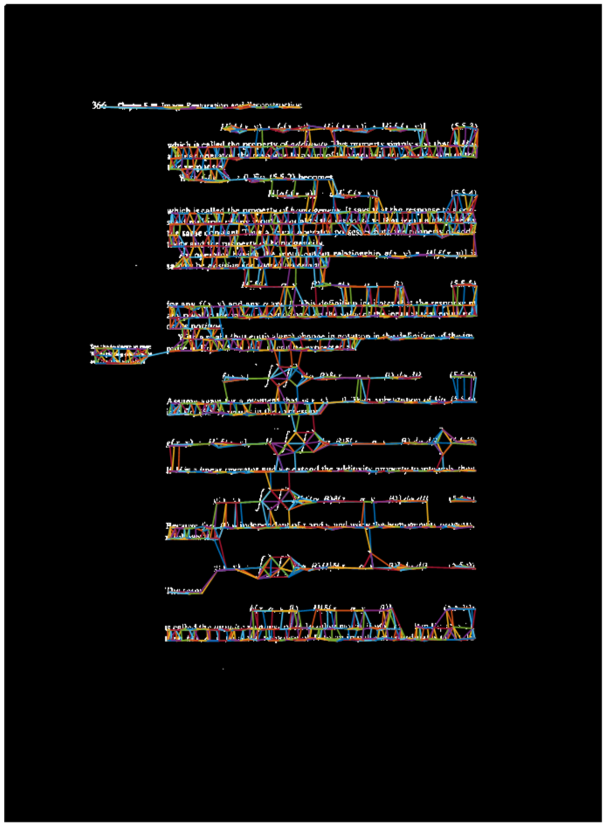
2. Detect Centroids

- 8-connected components
  - bwlabel
- Calculate area and position of centroids
- Filter out large and small centroids

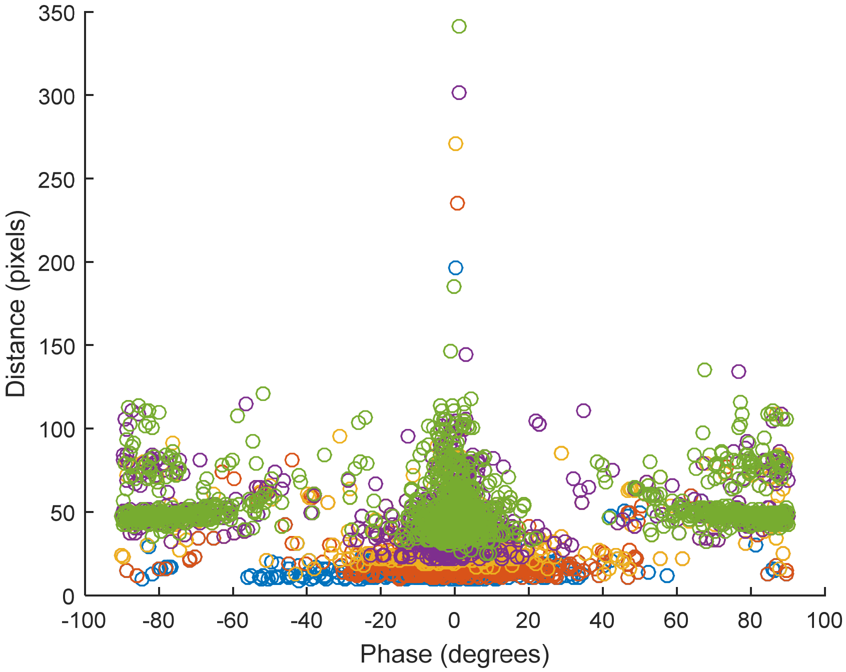


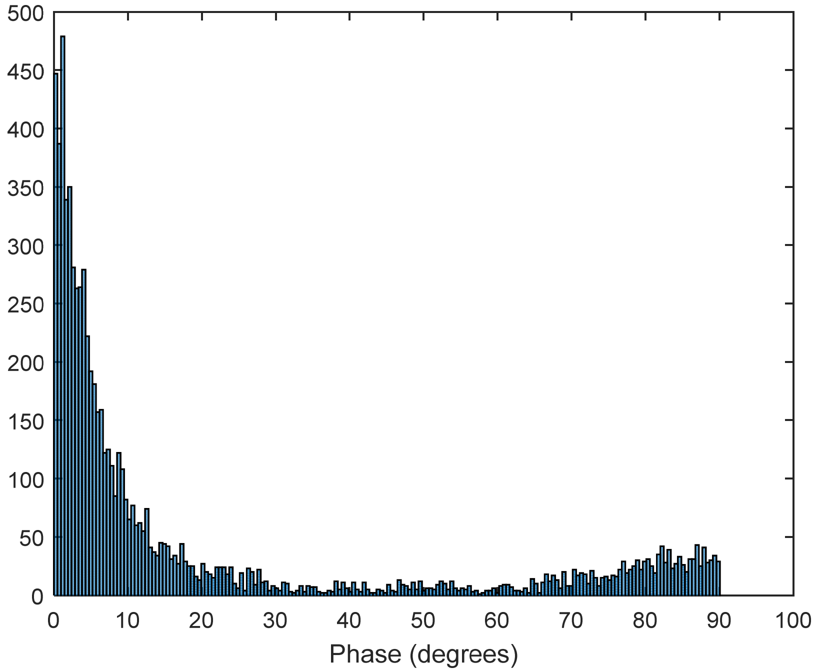
3. K-Nearest Neighbors

- 5 Nearest Neighbor
  - knnsearch
- Calculate Phase
- Calculate Distance
- Longest Part of Computation



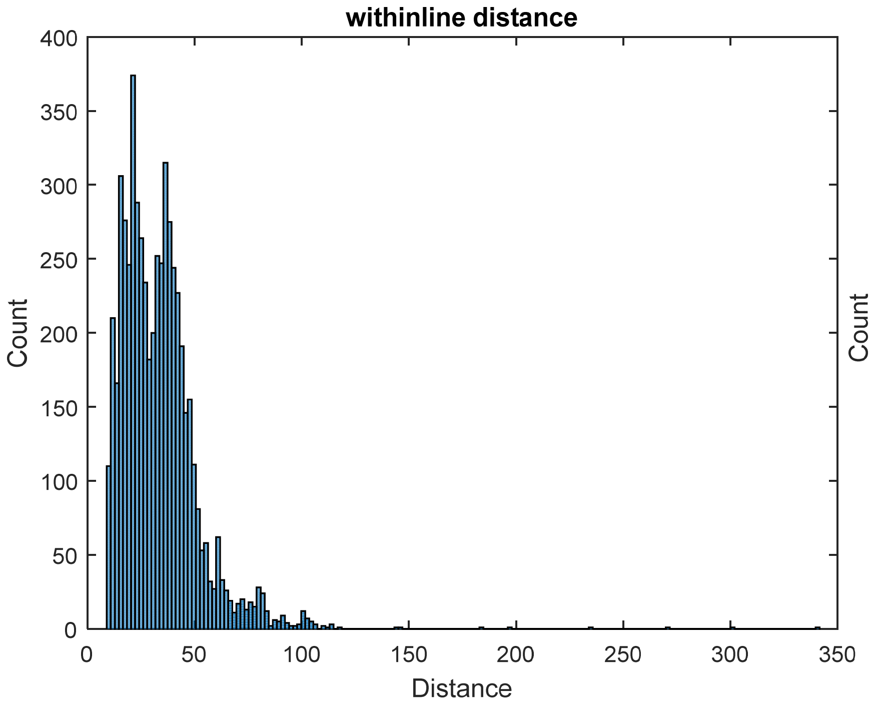
4. Estimate Phase

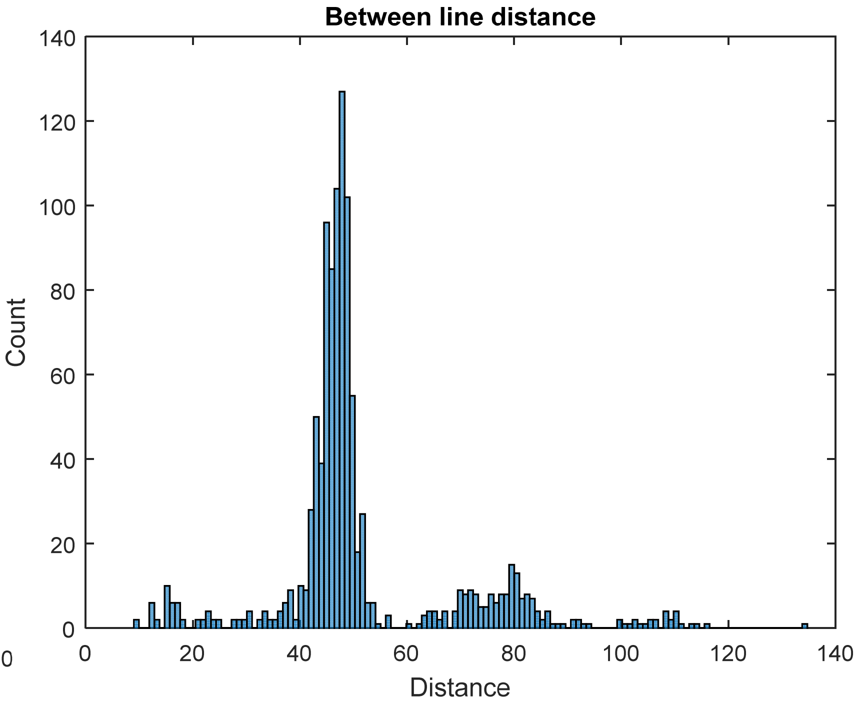




5.Estimate inline and between line distance

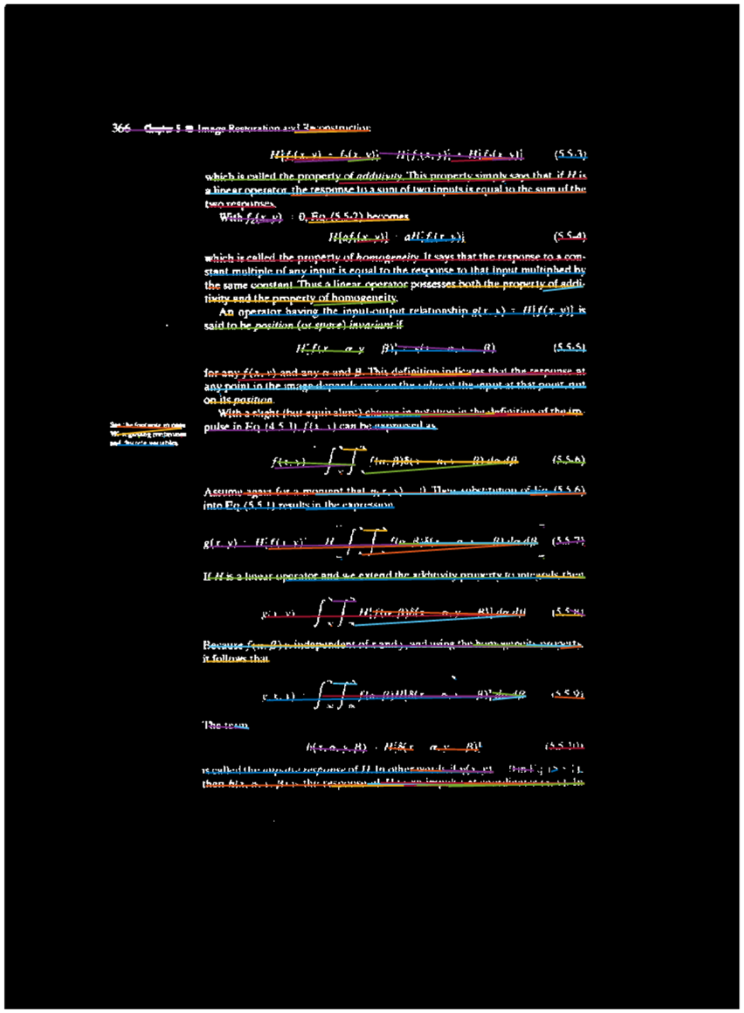
- Based on the phase
  - Nearest neighbors that have phase around 0 degrees are inline
  - Nearest neighbors that have phase around 90 degrees are between line


withinline distance


Between line distance

6. Find text lines

- Threshold centroids based on phase
- Transitive Closure
- Linear regression



366 Chapter 5 ■ Image Restoration and Reconstruction

$$H[f_1(x,y)+f_2(x,y)] = H[f_1(x,y)] + H[f_2(x,y)] \qquad (5.5.3)$$

which is called the property of *additivity*. This property simply says that, if H is a linear operator, the response to a sum of two inputs is equal to the sum of the two responses.

With $f_2(x,y) = 0$, Eq. (5.5.2) becomes

$$H[af_1(x,y)] = aH[f_1(x,y)] \qquad (5.5.4)$$

which is called the property of *homogeneity*. It says that the response to a constant multiple of any input is equal to the response to that input multiplied by the same constant. Thus a linear operator possesses both the property of additivity and the property of homogeneity.

An operator having the input-output relationship $g(x,y) = H[f(x,y)]$ is said to be *position (or space) invariant* if

$$H[f(x-\alpha, y-\beta)] = g(x-\alpha, y-\beta) \qquad (5.5.5)$$

for any $f(x,y)$ and any α and β. This definition indicates that the response at any point in the image depends only on the value of the input at that point, not on its *position*.

With a slight (but equivalent) change in notation in the definition of the impulse in Eq. (4.5.1), $f(x,y)$ can be expressed as

[illegible]

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha,\beta)\delta(x-\alpha, y-\beta)\,d\alpha\,d\beta \qquad (5.5.6)$$

Assume again for a moment that $\eta(x,y) = 0$. Then substitution of Eq. (5.5.6) into Eq. (5.5.1) results in the expression

$$g(x,y) = H[f(x,y)] = H\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha,\beta)\delta(x-\alpha, y-\beta)\,d\alpha\,d\beta\right] \qquad (5.5.7)$$

If H is a linear operator and we extend the additivity property to integrals, then

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H[f(\alpha,\beta)\delta(x-\alpha, y-\beta)]\,d\alpha\,d\beta \qquad (5.5.8)$$

Because $f(\alpha,\beta)$ is independent of x and y, and using the homogeneity property, it follows that

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha,\beta)H[\delta(x-\alpha, y-\beta)]\,d\alpha\,d\beta \qquad (5.5.9)$$

The term

$$h(x,\alpha,y,\beta) = H[\delta(x-\alpha, y-\beta)] \qquad (5.5.10)$$

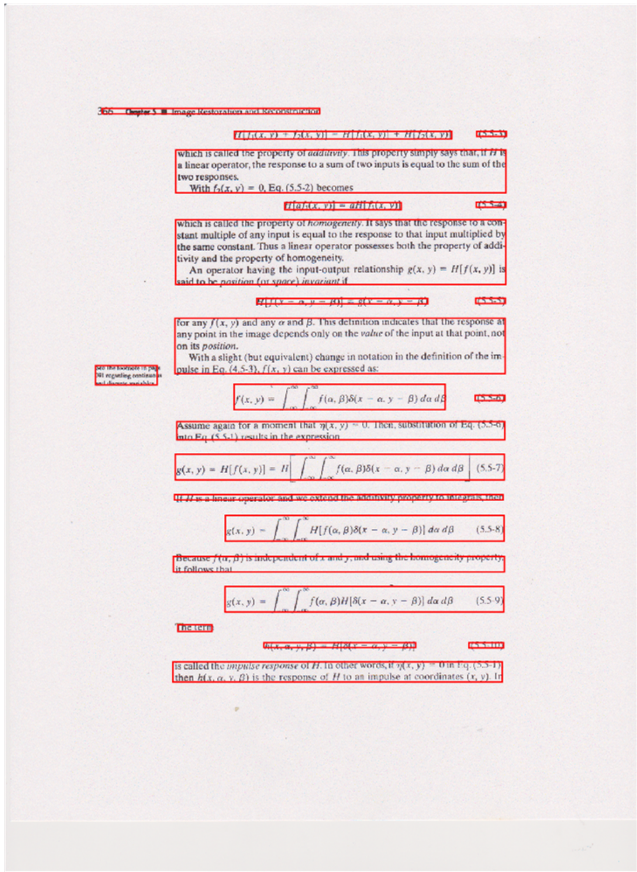
is called the *impulse response* of H. In other words, if $\eta(x,y) = 0$ in Eq. (5.5.1), then $h(x,\alpha,y,\beta)$ is the response of H to an impulse at coordinates [illegible]

7. Find text blocks

- Each line is compared to each other
  - If it meets the criteria to be in block then add it to the block
  - Else start a new block
- Sort text lines by:
  - Approximately parallel
    - based on estimated phase
  - Perpendicular distance
    - Based on between lines distance
  - Overlap or parallel distance
    - Based on inline distance
- Customized based on a document to document basis

8. Bounding box

- From the Previous step bounding boxes are drawn for each text block.
- Based on the position and size of a box each box can be labeled as text, equation, equation number, section heading, and etc.



366 Chapter 5 ■ Image Restoration and Reconstruction

$$H[f_1(x, y) + f_2(x, y)] = H[f_1(x, y)] + H[f_2(x, y)] \qquad (5.5\text{-}3)$$

which is called the property of *additivity*. This property simply says that, if H is a linear operator, the response to a sum of two inputs is equal to the sum of the two responses.

With $f_2(x, y) = 0$, Eq. (5.5-2) becomes

$$H[af_1(x, y)] = aH[f_1(x, y)] \qquad (5.5\text{-}4)$$

which is called the property of *homogeneity*. It says that the response to a constant multiple of any input is equal to the response to that input multiplied by the same constant. Thus a linear operator possesses both the property of additivity and the property of homogeneity.

An operator having the input-output relationship $g(x, y) = H[f(x, y)]$ is said to be *position (or space) invariant* if

$$H[f(x - \alpha, y - \beta)] = g(x - \alpha, y - \beta) \qquad (5.5\text{-}5)$$

for any $f(x, y)$ and any α and β. This definition indicates that the response at any point in the image depends only on the *value* of the input at that point, not on its *position*.

See the footnote in page 381 regarding continuous and discrete variables.

With a slight (but equivalent) change in notation in the definition of the impulse in Eq. (4.5-3), $f(x, y)$ can be expressed as:

$$f(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta \qquad (5.5\text{-}6)$$

Assume again for a moment that $\eta(x, y) = 0$. Then, substitution of Eq. (5.5-6) into Eq. (5.5-1) results in the expression

$$g(x, y) = H[f(x, y)] = H\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)\delta(x - \alpha, y - \beta)\, d\alpha\, d\beta\right] \qquad (5.5\text{-}7)$$

If H is a linear operator and we extend the additivity property to integrals, then

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H[f(\alpha, \beta)\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}8)$$

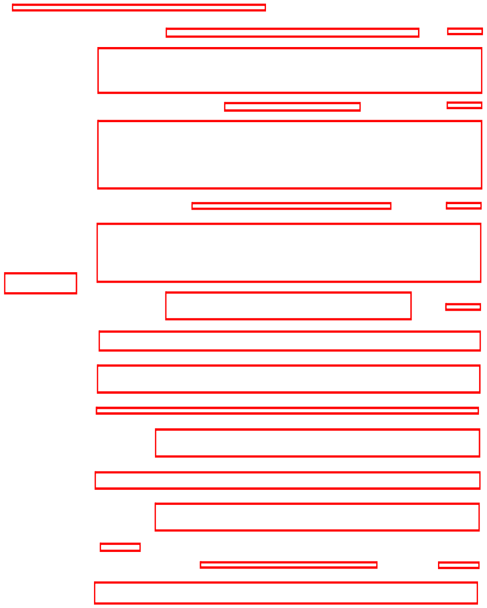
Because $f(\alpha, \beta)$ is independent of x and y, and using the homogeneity property, it follows that

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha, \beta)H[\delta(x - \alpha, y - \beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}9)$$

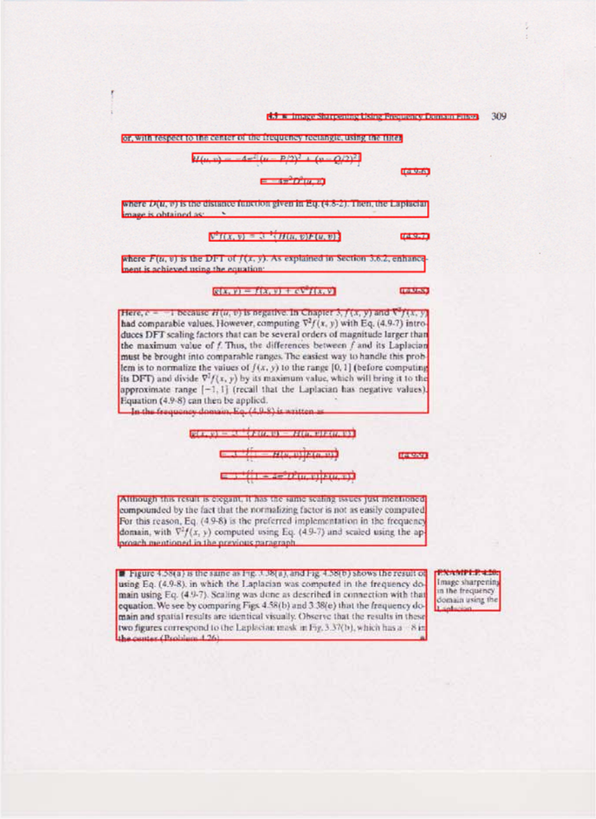
The term

$$h(x, \alpha, y, \beta) = H[\delta(x - \alpha, y - \beta)] \qquad (5.5\text{-}10)$$

is called the *impulse response* of H. In other words, if $\eta(x, y) = 0$ in Eq. (5.5-1) then $h(x, \alpha, y, \beta)$ is the response of H to an impulse at coordinates (x, y). In



Similar and Dissimilar Document Structure



4.9 ■ Image Sharpening Using Frequency Domain Filters 309

or, with respect to the center of the frequency rectangle, using the filter

$$H(u,v) = -4\pi^2[(u-P/2)^2 + (v-Q/2)^2]$$
$$= -4\pi^2 D^2(u,v) \qquad (4.9\text{-}6)$$

where $D(u,v)$ is the distance function given in Eq. (4.8-2). Then, the Laplacian image is obtained as:

$$\nabla^2 f(x,y) = \mathfrak{I}^{-1}\{H(u,v)F(u,v)\} \qquad (4.9\text{-}7)$$

where $F(u,v)$ is the DFT of $f(x,y)$. As explained in Section 3.6.2, enhancement is achieved using the equation:

$$g(x,y) = f(x,y) + c\nabla^2 f(x,y) \qquad (4.9\text{-}8)$$

Here, $c = -1$ because $H(u,v)$ is negative. In Chapter 3, $f(x,y)$ and $\nabla^2 f(x,y)$ had comparable values. However, computing $\nabla^2 f(x,y)$ with Eq. (4.9-7) introduces DFT scaling factors that can be several orders of magnitude larger than the maximum value of f. Thus, the differences between f and its Laplacian must be brought into comparable ranges. The easiest way to handle this problem is to normalize the values of $f(x,y)$ to the range [0, 1] (before computing its DFT) and divide $\nabla^2 f(x,y)$ by its maximum value, which will bring it to the approximate range [−1, 1] (recall that the Laplacian has negative values). Equation (4.9-8) can then be applied.

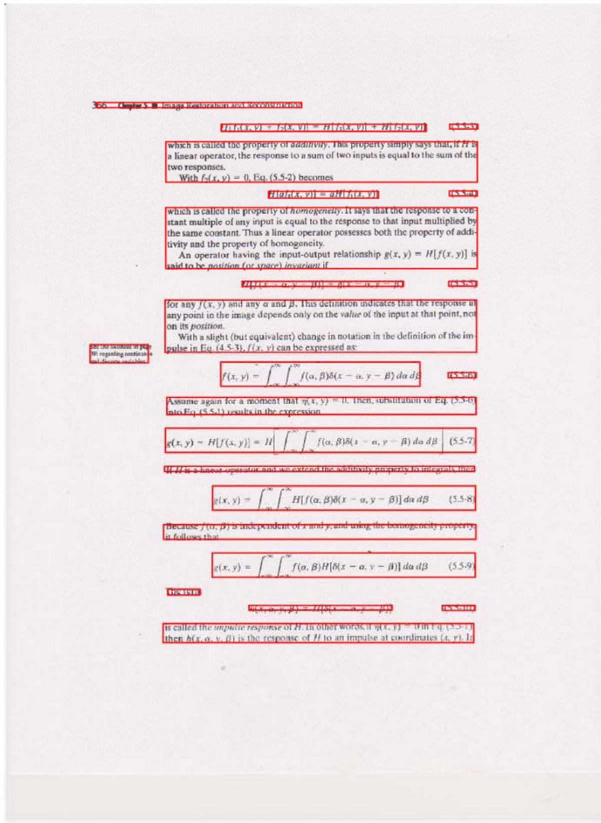
In the frequency domain, Eq. (4.9-8) is written as

$$g(x,y) = \mathfrak{I}^{-1}\{F(u,v) - H(u,v)F(u,v)\}$$
$$= \mathfrak{I}^{-1}\{[1 - H(u,v)]F(u,v)\} \qquad (4.9\text{-}9)$$
$$= \mathfrak{I}^{-1}\{[1 + 4\pi^2 D^2(u,v)]F(u,v)\}$$

Although this result is elegant, it has the same scaling issues just mentioned, compounded by the fact that the normalizing factor is not as easily computed. For this reason, Eq. (4.9-8) is the preferred implementation in the frequency domain, with $\nabla^2 f(x,y)$ computed using Eq. (4.9-7) and scaled using the approach mentioned in the previous paragraph.

EXAMPLE 4.20: Image sharpening in the frequency domain using the Laplacian.

■ Figure 4.58(a) is the same as Fig. 3.38(a), and Fig. 4.58(b) shows the result of using Eq. (4.9-8), in which the Laplacian was computed in the frequency domain using Eq. (4.9-7). Scaling was done as described in connection with that equation. We see by comparing Figs. 4.58(b) and 3.38(e) that the frequency domain and spatial results are identical visually. Observe that the results in these two figures correspond to the Laplacian mask in Fig. 3.37(b), which has a −8 in the center (Problem 4.26). ■



366 Chapter 5 ■ Image Restoration and Reconstruction

$$H[f_1(x,y) + f_2(x,y)] = H[f_1(x,y)] + H[f_2(x,y)] \qquad (5.5\text{-}3)$$

which is called the property of *additivity*. This property simply says that, if H is a linear operator, the response to a sum of two inputs is equal to the sum of the two responses.

With $f_2(x,y) = 0$, Eq. (5.5-2) becomes

$$H[af_1(x,y)] = aH[f_1(x,y)] \qquad (5.5\text{-}4)$$

which is called the property of *homogeneity*. It says that the response to a constant multiple of any input is equal to the response to that input multiplied by the same constant. Thus a linear operator possesses both the property of additivity and the property of homogeneity.

An operator having the input-output relationship $g(x,y) = H[f(x,y)]$ is said to be *position* (or *space*) *invariant* if

$$H[f(x-\alpha, y-\beta)] = g(x-\alpha, y-\beta) \qquad (5.5\text{-}5)$$

for any $f(x,y)$ and any α and β. This definition indicates that the response at any point in the image depends only on the *value* of the input at that point, not on its *position*.

See the footnote in page 99 regarding continuous and discrete variables.

With a slight (but equivalent) change in notation in the definition of the impulse in Eq. (4.5-3), $f(x,y)$ can be expressed as:

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha,\beta)\delta(x-\alpha, y-\beta)\, d\alpha\, d\beta \qquad (5.5\text{-}6)$$

Assume again for a moment that $\eta(x,y) = 0$. Then, substitution of Eq. (5.5-6) into Eq. (5.5-1) results in the expression

$$g(x,y) = H[f(x,y)] = H\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha,\beta)\delta(x-\alpha, y-\beta)\, d\alpha\, d\beta\right] \qquad (5.5\text{-}7)$$

If H is a linear operator and we extend the additivity property to integrals, then

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H[f(\alpha,\beta)\delta(x-\alpha, y-\beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}8)$$

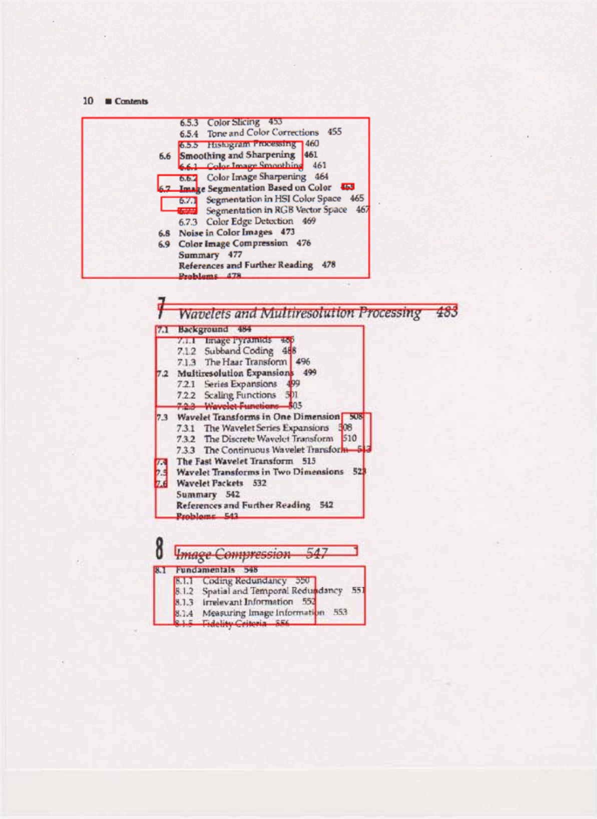
Because $f(\alpha,\beta)$ is independent of x and y, and using the homogeneity property, it follows that

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\alpha,\beta)H[\delta(x-\alpha, y-\beta)]\, d\alpha\, d\beta \qquad (5.5\text{-}9)$$

The term

$$h(x,\alpha,y,\beta) = H[\delta(x-\alpha, y-\beta)] \qquad (5.5\text{-}10)$$

is called the *impulse response* of H. In other words, if $\eta(x,y) = 0$ in Eq. (5.5-1), then $h(x,\alpha,y,\beta)$ is the response of H to an impulse at coordinates (x,y). In



10 ■ Contents

Discussion

- Pros
  - Can separate analysis into subsection for more accurate results
  - Analysis independent of skew
- Cons
  - The algorithm needs to be customized based on the document
    - Current area of research
  - Nearest neighbor computation is computational heavy
- Future work
  - Need to implement skew estimation
  - Explore more advanced techniques
  - Use in conjunction with OCR

References

[1]O'Gorman, L., "The document spectrum for page layout analysis," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.15, no.11, pp.1162-1173, Nov 1993

[2]Simon, A.; Pret, J.-C.; Johnson, A.P., "A fast algorithm for bottom-up document layout analysis," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.19, no.3, pp.273-277, Mar 1997

[3]Lawrence O’Gorman, Rangachar Kasturi, Document Image Anlysis, ISBD 0-8186-7802-X, Library of Congress Number 9717283.

[4]Cattoni, R., Coianiz, T., Messelodi, S., Modena, C.M.: Geometric layout analysis techniques for document image understanding: a review. Technical report, IRST, Trento, Italy (1998)

[5]Seong-Whan Lee; Dae-Seok Ryu, "Parameter-free geometric document layout analysis," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.23, no.11, pp.1240-1256, Nov 2001

Questions?