

Elisa H. Barney Smith

## Contents

Introduction.....	12
Document Creation Materials.....	12
Writing Substrates.....	12
Inks.....	17
Writing and Printing Processes.....	20
HandHeld Writing Instruments.....	20
Machine Printing.....	23
Acquisition Methods.....	39
Flatbed Scanner and Fax Machine Acquisition.....	39
Cameras and Mobile Devices.....	44
Video.....	46
Other Specialty Modes.....	47
Document Quality.....	48
Factors Affecting Document Quality.....	48
Effects of Document Quality on Analysis and Recognition Results.....	50
Models of Document Degradations.....	51
Conclusion.....	59
References.....	60
Further Reading.....	60

---

## Abstract

A summary of materials used in creating documents, methods of creating the printed document, and methods to acquire a digital version of that document are presented. Current as well as historical methods, materials, and processes are presented. Along with this, a discussion of places where image degradations can enter the process is included. All this is related to how these aspects could affect document recognition ability.

---

E.H. Barney Smith  
Electrical & Computer Engineering Department, Boise State University, Boise, ID, USA  
e-mail: [EBarneySmith@boisestate.edu](mailto:EBarneySmith@boisestate.edu)

---

**Keywords**

Acquisition methods • Document degradations • Document defects • Document quality • Ink • Paper • Printing • Scanning

---

**Introduction**

Documents can be created by hand or by machine. In either case, several factors affect the final appearance, including the content, pigment, instrument transferring pigment to the paper, and the paper itself. How either people or machines perceive document appearance depends on how it is acquired. What is considered good quality on paper, when received directly by a human eye and processed by a human brain, is not always considered good quality when digitized and then viewed on a monitor. Likewise, what a person considers good perceptual quality on the original or digitized version is not always of a quality that can make the document content recognizable by a high precision machine.

To help explain the relationship between document sources and their quality, this chapter identifies junctures at which quality can decrease as it describes:

- Materials – materials, such as paper and ink; people: and machines use to create a document.
  - Processes – current and obsolete processes for creating printed text by hand or machine. Particular obsolete processes are noted for technologies archivists see in historical document collections.
  - Acquisition methods – methods for converting documents to digital form, facilitating automatic document image processing and recognition.
  - Models – document production models, quality measures, and how quality affects recognition results.
- 

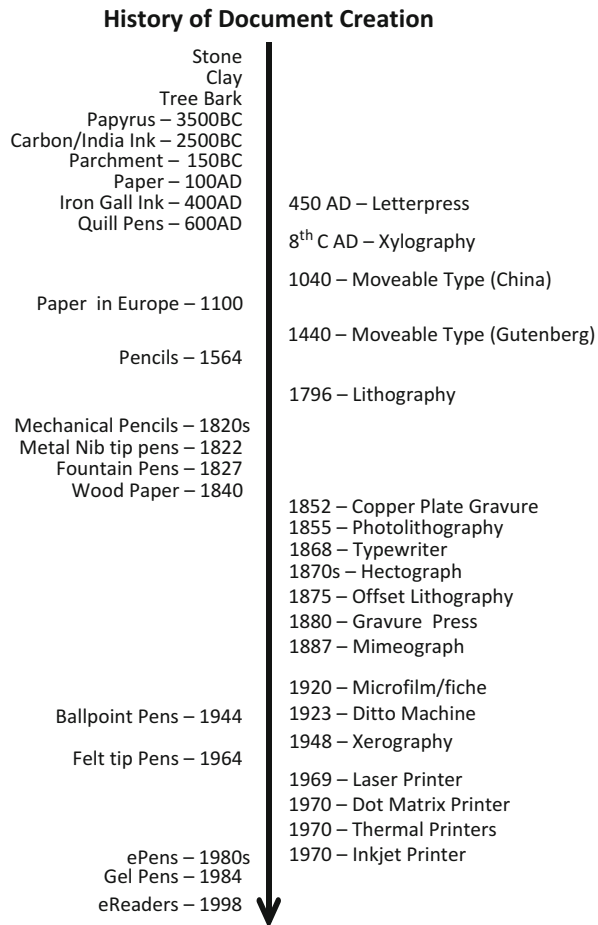
**Document Creation Materials**

This is an overview of some of the materials that have been used over time to create documents and which significantly shape their appearance. Materials include the substrate on which the document appears, usually considered a form of “paper,” and the ink that shows the written message. The choice of paper and ink is partially historical, set by available materials and technologies, and partially by the writing or printing process. Transferring the ink to the substrate can be done by hand with a writing instrument or by machine through a printing process. Figure 2.1 includes examples of different materials and methods for writing and printing and when they were first introduced. Each has introduced a new variable affecting the appearance of a final document.

**Writing Substrates**

Writing substrates are writing surfaces. Surprisingly, while there have been many since the earliest humans first began writing, many substrate fundamentals have not

**Fig. 2.1** Timeline of document creation materials and methods



changed all that much. The oldest writing available to study is preserved because it was written or carved on stone or because it was impressed into clay tablets. While these materials have longevity and in certain areas are plentiful, they are not particularly portable. Almost any portable substance that would retain the marks of a brush or pen was used as a writing substrate. This includes leaves, tree bark, boards, and cloth. In China, old writing has been found on bamboo sticks and in India writing on birch bark and palm leaves. The Mayan wrote on “paper” from the inner bark of the fig tree which was coated with a thin layer of a plaster like substance, and the text was painted onto the “paper”-like stucco painting. Not unlike today’s books, these were folded fanlike into a book form.

### From Papyrus to Parchment and Paper

The most well known of ancient writing substrates is *papyrus* from which the modern word paper derives. As early as 3500 BC, Egyptians used papyrus to form their paper. Papyrus is a type of reed called a sedge. The rind was removed to expose

the soft inner pith which was flattened, and strips were laid out in overlapping layers at right angles. This was cemented by beating the pith until the plant tissue ruptured, and the sap from the tissues formed a glue that would hold the strips together. The material was dried under pressure and polished on one side to form a smooth surface on which writing would occur. The standard writing unit or what we consider today to be a “page” evolved from the size of one of these units. Several of these units (on the order of 20) were combined by overlapping the edges from one unit to the next and cementing those in a similar fashion to form a roll, which became known as a *volumen* from the Latin word “to roll.” Each scroll contained about as much information as seven to ten pages of a modern handwritten book. The word *book* came from the name of the port, Byblos, through which the Greeks imported papyrus in the late Iron Age.

*Parchment* is a writing material made from stretched and untanned animal skins, particularly, calves, sheep, or goats. While leather was used for writing since 2000 BC, it didn’t store well and could only be written on one side. Parchment became commonly used for writing once a method was developed in the second century BC to allow both sides to be used for writing, although the interior side had a smoother surface. In Europe, it became the primary writing substrate from the fourth century AD until the Renaissance and the introduction of paper. Parchment made from fine skins from young calves or goats is called *vellum*. With the use of parchment, the writing substrate was no longer rolled but bound into codices, as we do with today’s printed books. To print, the Latin Bible required more than 500 calf hides. The quantity of hides necessary to create books made them quite expensive; thus, it became common to reuse parchment. The ink was scraped off the parchment or the writing was done at right angles to visually distinguish the new writing from the traces of the old. Books with this reuse are called *palimpsest*, from the Greek “to scrape again.” Even with this reuse, depending on parchment as a substrate limited book production.

*Paper* as we consider it today originated in China in the first century AD. The Chinese kept the process for creating paper a secret for many centuries, until they attacked the Arab city of Samarkand in the eighth century, and the Arabs took prisoner some of the Chinese skilled in making paper. Papermaking then moved west as the Arabs expanded their control in Europe. The first paper mill in Europe was founded in 1100 AD in Constantinople, and papermaking spread rapidly once in Europe until by the fourteenth century it was established throughout Europe. The introduction of paper in Europe led to an increase in the production of books and coincided with an increase in readers.

### **Paper Production**

Paper production begins by shredding and reducing vegetable fibers to a pulp in water. A thin layer of pulp is spread on a screen, and the water is drained off to form a felt. Pulp fibers are matted and dried. Process differences result in the variations among available papers. Paper was produced by manual methods for several centuries, but this limited the quantity or size of the sheet that could be

produced at any one running of the process. The first mechanized papermaking process was invented in 1798 by Nicolas Louis Robert and was made commercially practicable in 1805 by Henry and Sealy Fourdrinier.

Initially, the fibers used in making paper were made primarily from linen, jute, flax, and hemp. Fibers from cloth rags were a common source through the seventeenth century. Paper made from cotton fibers is called *rag paper*. These papers are used commonly today for high-quality documents and banknotes. Using fiber from straw was experimented with in the eighteenth century, and esparto grass was used frequently in England in the nineteenth century. A bleaching process was developed so white paper could be made from colored fibers. Starting in Saxony in the mid-1800s, most modern paper is made from tree cellulose, and the discovery that this was a suitable source greatly increased the paper supply. Wood is reduced to a pulp by either mechanical or chemical methods. The mechanical grinding of wood pulp between stone grindstones introduces many impurities, which decrease the paper's quality. Wood cell walls are constructed from a network of cellulose filled in with lignin. Mechanical grinding does not remove the lignin so the yield is higher, but the lignin over time yellows the paper and makes it brittle. An alternative was developed to add chemical reagents such as soda and sulfate, to break down the lignin that holds the cells together. If the chemical agent is caustic soda, soft fluffy fibers result that are good for cover and writing paper. Calcium bisulphate or magnesium bisulfate produces a stronger or harder fiber to create a paper better suited for printing. These acids can lead to deterioration of the paper and the inks over time, so sodium sulfate is an alternative that makes a very tough paper. The fibers will be longer in chemical pulping than in mechanical pulping, and thus, the paper will be stronger.

*Acid-free paper* has a neutral pH from being treated with a mild base (usually calcium or magnesium bicarbonate) to neutralize the natural acids occurring in wood pulp and in the production process. It is also lignin and sulfur free. It is suitable for archival purposes because it will not yellow or become brittle for a long period of time. If an adequate alkaline reserve is included, such paper will likely survive 1,000 years.

Until the late eighteenth century, paper was mostly *laid*. The fibers were laid on a chain surface with long parallel links, interrupted occasionally by a perpendicular gap. This pattern could be seen in the final paper. In the eighteenth century, *wove paper* was developed that was smoother for better printing. A woven wire mesh transports the pulp, and the grain of the paper is no longer prevalent. Today, wove paper is considered a cheaper paper. *Watermarks*, also known as papermarks, are intentional patterns pressed into the grain. They started appearing in paper in the thirteenth century to indicate origin.

Waste paper can be recycled and used instead of raw tree wood as the source of pulp. The paper must be shredded and then returned to a pulp state. Along the way the ink must be bleached out of it. As the paper is returned to pulp, the length of the fibers is reduced. This lowers the quality of the paper produced from this pulp. It is therefore usually mixed with virgin pulp.

## Finishing Procedures

The surface characteristics of the paper affect the visual characteristics of the writing trace. The ink can either sit on top of the paper fibers or be absorbed into them. Europeans were accustomed to using quill pens on parchment and needed a strong, scratch-resistant, non-absorbing paper. *Sizing* adds gelatin, rosin, starches, gums, or alum to the paper to make it harder and less absorbent and thus resistant to the water in water-based writing inks. Sizing can be done after the formation of sheets through *tub sizing* by placing the paper in a bath of gelatin essentially coating it with a thin layer of glue. Alternatively, with *engine sizing* the pulp is mixed with sizing materials during processing before the sheets are formed. Tub-sized paper is higher quality than engine sizing because the sizing material is located where it is most effective, but it is also more expensive. Sizing made paper durable enough that both sides of the paper could be used for printing.

There are several paper *finishes*. They are often *coatings* of pigments or vehicles (binders) such as calcium carbonate or china clay. Coatings can produce a *matte* (dull or muted), *semimatte*, or *glossy* finish. Paper was originally brush coated with a clay substance to produce a surface suitable for fine-screened halftones for use in the finest quality photographic reproduction. A *machine finish* will produce a smoother surface and is often used for magazines. *Coated* paper is usually white and in text weight. *Gloss* will lead to less dot gain when printing as the ink will not spread as much. Uncoated paper is found in white and colored versions. *Art paper* is paper glazed with a china clay coating then rolled to make it very smooth to better print halftones/screens for illustrated documents. However, the china clay coating reacts with the acids in the paper and makes the paper brittle so folds quickly become cracks. Coloring was first added to paper in 1687, and machine ruling lines first appeared in 1770.

*Calendering* is a finishing operation that passes the paper through a series of steel rolls to impart a glossy finish or increase the surface smoothness or opacity. Minimum calendering produces an eggshell or antique paper, which has a rougher texture and is very “non-glare,” which can increase readability. *Supercalendered* paper is given a smooth shiny finish by repeated rolling of the paper between hot and cold rollers. Machine finish papers have fairly extensive calendering and are used for magazines, because the finish enables the printing to reproduce very fine halftones.

## Paper Classifications, Uses, and Quality

Paper production materials and processes influence paper quality. The paper options influence their use, characteristics, and quality. There are three factors to consider when buying paper today: *grade*, *whiteness*, and *opacity*. A higher grade paper has a more refined smoothness, whiteness, and greater opacity than other papers. In addition, there are four basic paper classifications to consider: *bond*, *book*, *cover*, and *cardstock*. Bond paper – lower grade paper that is used in most offices for printing and photocopying – has a semihard finish, and book paper comes in a range of textures. Rough papers will likely have ink dropouts where ink never reached the

**Table 2.1** Paper characteristics, uses, and quality


Classification	Use	Grade	Whiteness	Opacity	Thickness	Finish
Bond	Stationery, office copies, manuals	Writing, copy bond, digital, virgin pulp or recycled	Average commodity whiteness	More with thickness	Varies	Smooth or textured
Book	Text pages of books and booklets	Varies by whiteness, opacity, thickness, and finish properties	Varies with grade	Made with or without opacity properties	Varies	Smooth, vellum, coated, or uncoated
Cover	Pamphlets, book covers	Same as book/text	Same as book/text	More than book/text due to thickness	Medium to heavy	Same as book/text
Cardstock	Postcards, business cards, misc.	Varies	Varies	More due to thickness	Medium to heavy	Smooth or vellum

paper during the initial printing process. Ink spreads according to the porosity. Filler material, such as white chalks, clays, and titanium dioxide, is often added to the pulp to give it better opacity and surface finish. Cover and cardstock are not often used for producing documents (Table 2.1).

Paper is graded by thickness. In North America and Great Britain, this is indicated by measuring the weight of a ream of paper cut to the basis size for that grade of paper. A ream has 500 pages, although in times past a ream had 480–520 sheets. The basis size for bond paper is  $17 \times 22$  in. and for book paper is  $25 \times 38$  in.; thus, 20lb bond paper is equivalent in thickness to 50lb book paper. In Europe, paper grading is much simpler and uses the weight in grams per square centimeter ( $\text{g/m}^2$ ), sometimes abbreviated as gsm. 20lb bond paper is equivalent to 75.2 gsm paper. Paper thickness contributes to the likelihood of printed matter on the verso (back) side being visible on the recto (front) side. Calendering makes more dense paper. The choice of fillers also contributes. India paper is a very thin paper that is also opaque.

## Inks

Inks can be grouped into two categories, those used with handheld writing instruments and those used by mechanical printing processes. Inks are all made from *colorants* (pigments and dyes), *vehicles* (binders), *additives*, and *carrier substances* (solvents). The desired flow property depends on the printing or writing process that will be used as the ink has to match the transfer mechanism and drying or fixing process. Inks range from thin and watery to viscous and also exist in powders or solids. Ink must run freely but not spread. It must dry easily and not harm the paper or writing instrument.

Egyptians around 3000 BC used black ink made from carbon and red ink made from natural iron oxide suspended in water with gum or glue. Pictures of scribes and the hieroglyph for scribe always include a rectangle with two circles within it, , representing the wells for these two ink colors. At about the same time, the Chinese developed a similar black ink made from lamp or carbon black suspended in a dilute solution of water soluble gums. This kind of ink is called “India ink” as it was introduced to the west through India. This ink requires frequent stirring so the carbon remains in suspension. The carbon pigment remained on the paper surface instead of soaking into the paper. This ink is stable and shows minimal effects of age, but is water soluble.

Iron gall ink was invented in the fifth century AD and became the prominent writing material from the Middle Ages through the twentieth century. It was made from a mix of iron salts (often vitriol or iron(II) sulfate), tannin (an acid from oak galls from which gallotannins are extracted), and glue (gum Arabic, a vegetable gum from the acacia tree) to bind. Over time, the iron-tannin components would oxidize turning the ink black giving it the name “blue-black ink.” Eventually, this ink fades to a dull brown color. In the nineteenth century, indigo dye was first added to inks to produce a less acidic blue ink.

Colorants used in inks can be organic or inorganic pigments in soluble oil. Pigments have particle sizes 0.1–2  $\mu\text{m}$  and are held in suspension. They need a vehicle for binding them to the paper. Vehicles can also coat the pigments and protect against mechanical wear (abrasion) and are sometimes called varnishes. Pigments have a wide color absorption band. Dyes during application have higher color intensity, produce more luminous colors, and come in a wider range of colors. Dyes are organic compounds that are dissolved. Natural dyes were initially used for color but were replaced by aniline and synthetic dyes around 1900. Synthetic dyes are used almost exclusively today. Dyes can be transparent, and the particles are smaller than in pigment, but they are less light fast than pigments. Most printing methods use pigment, but inkjet printers predominantly use dyes.

Binders are usually resins dissolved in mineral oil. Additives depend on the printing process and influence drying time, flow behavior, and abrasion resistance. Carrier substances are thinning agents like mineral oil or solvents like toluene.

In the 1940s, the ball-point pen was commercially introduced which uses a viscous quick drying paste like ink. Ball-point pen ink colors come from synthetic dye and include methyl violet, Victoria blue, and luxol fast orange; nigrosine; copper phthalocyanine; and other organometallic dyes. Dyes and pigments compose about 25 % of the mass of a typical ball-point ink. The solvents or vehicles are made from a mixture of glycols such as ethylene glycol. Prior to 1950, oils such as linseed or mineral oil were used. The vehicle dissolves or suspends the dyes or pigments and promotes the smooth flow of the ink over the surface of the rotating ball. The vehicle dries quickly usually through evaporation leaving the color on the paper. Solvents make up 50 % of the mass of the ink. The remaining 25 % of the ink is resins which can be naturally occurring or synthetic materials and provide a viscosity to the ink.

In the 1970s and 1980s, felt-tip and roller writer pens were introduced which use a liquid ink that transfers through the tip and soaks the paper evenly. Fluid ink will



penetrate the paper fibers more than viscous inks. Gel pen inks introduced in the late 1980s are viscous, but not to the degree of ball-point ink. The gel is water based with biopolymers, such as xanthan and tragacanth gum, as well as polyacrylate thickeners. Gel ink contains pigment suspended in a viscous medium, so it has a thicker deposit of pigment making a bolder line. The pigments are opaque and come in a variety of bold colors. The pigments are typically iron oxides and copper phthalocyanine.

In addition to liquid and viscous inks, inks can also be solids. The Romans used lead rods for marking. When a large source of graphite was discovered in England in 1564, it was not understood that it was not a variety of lead, and the name remains today. *Pencil “lead”* consists of waxes, fillers (clay), powdered graphite, and water blended and extruded into rods which are dried and kiln fired. The result is porous and can be impregnated with wax to make the writing smoother. Colored pencils use colored pigments with clay and wax or fatty acid combined with water and an absorbent material like gum tragacanth. These are dried and the firing stage is omitted. Pencils come in several levels of hardness which come from varying the ratio of clay and graphite. In Europe, these range from 9H to H, F, HB, then 1B to 9B. H is a hard lead which deposits very little carbon on the paper making the mark very light, and B is a softer lead which writes very black. In North America, lead hardness is primarily indicated by numbers 1–4, with 1 corresponding to the European 1B, the most common hardness; 2 corresponding to HB; 3 corresponding to H; and 4 corresponding to 2H.

### **Inks for Machine Printing**

Printer ink is very different from pen ink. Ink characteristics are intertwined with machine printing technologies. This section focuses on the inks, with more details about the machine technologies identified in the section “[Machine Printing](#).” Letterpress inks are viscous, almost like paint. Historically, it is sometimes called black “treacle” as it was made from linseed oil boiled down until it attained a glue-like consistency after it was freed from the fats it contained when raw. The coloring came from lamp black particles that were ground and reground until they were very fine and would not clog the counters of the smallest letters. Modern inks are made from a combination of solvents and plastics. These inks dry through absorption into the paper. Offset printing is a commonly used printing technology that transfers (or “offsets”) an inked image from plate to rubber blanket and then to paper. It also uses an ink that is a highly viscous pasty ink. It is made of hard resin, sometimes alkyd resin; vegetable oil (linseed, soy, wood); or mineral oil and pigment. Gravure printing ink has a lower viscosity making it a liquid ink so it can fill engraved cells. Common solvents are toluene, xylene or petroleum spirits, ethanol, ethyl acetate, or water (sometimes mixed with alcohol).

The ink used for typewriters is contained on a ribbon. The ribbons are sometimes textile ribbons, and the weave of the ribbon is often visible in the character image, [Fig. 2.4b](#). Later development led to the production of a tape with a removable film of black ink that would transfer to the paper when pressure was applied. This tape was less susceptible to drying than the inked ribbons. Since it transferred a more

uniform coating of ink to the paper, it produced dark areas more uniformly than ribbons. It also kept the typeface from getting gummed up as the type only came in contact with the non-inked back of the tape (Fig. 2.5).

The toner used in *xerography*, as in laser printers or copy machines, is not restricted to a liquid ink and is most often a carbon-based powder mixture. The particles usually include magnetic carrier particles, often iron oxide, and a polymer that melts to affix the toner to the paper. The carrier is recycled and can be 80  $\mu\text{m}$  while the toner is 4–8  $\mu\text{m}$ . Toner without a carrier has particles 12–20  $\mu\text{m}$  in diameter. Liquid toner for xerography will contain particles of 1–2  $\mu\text{m}$  and allow colors to be directly mixed by mixing toner during the printing process.

Inkjet printers require a low viscosity ink that must be filtered so pigment agglomerates don't block channels in the nozzle of the print heads. Inkjet inks are usually water based. They tend to bleed or penetrate the substrate surface and can cause the substrate to warp or wave. Thus, specially coated paper is recommended for use with this printing method. Some inkjet papers melt a wax or plastic ink that remains on the surface of the paper.

---

## Writing and Printing Processes

The ink can be transferred to paper via a handheld device or a larger machine. This section describes the technologies in both categories. The inks used in these writing and printing methods were described in the section “[Inks](#).”

### HandHeld Writing Instruments

Before the advent of machine printing, all writing was done with a handheld writing instrument. Handwritten and hand-printed documents are all created by sliding a writing instrument across the writing substrate. There are many types of handheld writing instruments such as brushes, nib pens, ball-point pens, felt-tip pens, and pencils. The appearance of the stroke that results is determined by the shape of the writing tip, including how it deforms when in contact with the writing substrate, and the characteristics of the ink such as fluid type and opacity.

The Greeks used metal styli to mark on wax tablets. The Sumerians used reeds to imprint on clay tablets. In northern India people used a reed pen which led to the development of angular script forms, whereas in southern India, people used a metal stylus, and a more rounded script form evolved to not tear the paper. In Egypt, the stylus used for writing was a reed, whose end was chewed to make a kind of brush, so writing resembled painting, but the core of the reed held ink. In 1000 BC, the Chinese used a camel hair or rat hair brush. The medieval European scribes used a small brush for fine work called a *pencilus* (“little tail”) which led to the word pencil. Brushes are likely to have a variable stroke width and may have streaking in the stroke.



**Fig. 2.2** Samples of metal nibs and writing

*Quill pens* were introduced to Europe in the sixth century. The word pen comes from the Latin word *penna* which means feather. A nib, or tip, was cut into a feather from a large bird, usually a goose. This was then dipped in ink, usually water based, to form a reservoir of ink in the hollow shaft. Pressure between the nib and the paper caused ink to be transferred via capillary action to the paper. Through use, the point on a quill pen would wear down requiring the feather to be cut again. This could be done by the writer, or pen cutters often “stationed” themselves on streets offering their services and lending the word *stationary* to office supplies. In the 1800s, metal inserts (Fig. 2.2) were developed to remove the constant need to recut the nib. Early *metal nibs* had problems with lack of flexibility and corrosion, especially with the use of iron gall ink. When writing with a metal nib, the points of the tip often separate under pressure on the downstroke creating furrows in the paper that fill with extra ink called “nib tracks.” The shape of the nib influenced the writing styles and vice versa. The broad nib has a flat edge, and the thickness of the stroke depends on the angle of the stroke relative to the pen. Pointed nibs vary the stroke width by exerting variable levels of pressure to separate the tines different amounts. *Fountain pens* are nib pens that have an internal reservoir of ink. The first successful fountain pen was developed in 1884. This removed the gradual fading seen in writing as the reservoir of dip pens emptied.

*Ball-point pens* are the most common of all writing instruments today (Fig. 2.3a). The first patent for a ball-point pen was issued to an American named John Loud in 1888. He designed them to be able to write on rough surfaces. Improvements in ball grinding and measuring techniques enabled the pens to be constructed well.



**Fig. 2.3** Examples of writing tips and writing samples for (a) ball point, (b) gel (c) felt tip and (d) pencil

The Biro pen was introduced in England in 1944 and was widely adopted in the 1950s. Precision ground ball tips are 0.7–1.0 mm in diameter and fit in a socket with narrow ducts to channel ink from the reservoir. The ink is thick and gelatinous and is deposited by capillary action to the ball, thus requiring more pressure than when writing with a nib pen. Ball-point pens write best on rougher surfaces where friction can turn the ball and draw more ink. As the tip is used for propulsion, more ink will be on the edges of the stroke than in the center. *Rollerball pens* are similar to ball-point pens except the ink is more fluid. This ink soaks into the paper more than ball-point ink and produces a different stroke texture. Rollerball pens also require less pressure to write. *Gel pens* are ball-point or rollerball pens with a gelatinous ink (Fig. 2.3b).

*Felt-tip pens* or markers were introduced in Japan in 1964 (Fig. 2.3c). The tips are either hard felt or fiber bundles bound by resin. Felt-tip pens use a more liquid ink and thus are likely to have a less opaque stroke. In addition to black and dark blue, they contained brightly colored fluid inks. This type of pen is also used for highlighters which use an ink that intentionally produces a nonopaque mark. Permanent or indelible markers use a dye that is not easy to remove from a surface.

*Pencils* write by using the roughness of the writing substrate to grind off a thin layer of pencil “lead” or graphite which remains on the substrate as a mark (Fig. 2.3d). Variations in appearance result from the hardness of the lead and the pressure applied. Traditional wood pencils encase the lead in wood. In the first quarter of the nineteenth century, mechanical pencils which feed the lead through

a metal or plastic case became common. The common lead diameters are 0.5, 0.7, and 1.0 mm. Mechanical pencils have finer lead than wooden pencils. Wood pencils rely on a sharpening device to remove the surrounding wood and shape the tip to a finer point, usually smaller than the diameter of the fill lead. The degree of sharpness affects the writing appearance. This changes through use over time, and thus, a more consistent response is found with mechanical pencils.

The handheld writing instruments described so far transfer ink to a substrate like paper. The path the instrument takes is thus recorded and provides the image which is either directly viewed or ultimately digitized. *ePens* have been developed that allow this same pen trace to be recorded on a computer, either instead of marking the paper, or while marking the paper. They are often connected to displays where the trace is shown. To identify their position, some *ePens* are designed to work with special paper. The Anoto<sup>®</sup> paper has a dot pattern that varies across the paper. The pen, usually a ball-point pen, writes like a traditional pen but has an optical sensor built into it to see the dots and determine where the pen tip is hitting the paper. This information is delivered to a computer which decodes it to record the pen trace path. Other *ePens* have accelerometers to determine the motion directly which, when integrated, become positions. Alternatively, they rely on an electronic pad that detects the position of the pen. The *ePens* are the input device for online handwriting recognition and are discussed further in ►[Chap. 26](#) (Online Handwriting Recognition) and ►[Chap. 23](#) (Analysis of Documents Born Digital).

## Machine Printing

While ultimately the human hand participates in all printing processes, those where the hand is not directly involved are considered machine printing. Machine printing can be divided into two major subcategories, *impact* and *nonimpact*. Impact printing is the older of the technologies, while nonimpact printing has been enabled by the introduction of electronics. As the variety of new printing methods has expanded, all printing methods that do not require a printing plate master for the image or character have been categorized as nonimpact.





### Impact Printing

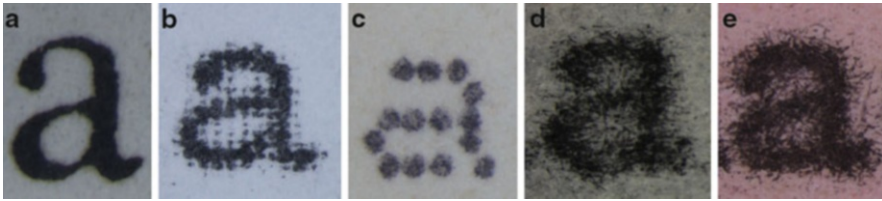
Impact printing uses four major methods to get ink onto paper. The documents produced by each of these methods have different image characteristics. Each has a preferred paper and ink. The combination of the ease of use and historical context of each has influenced how prevalent each type has been and how frequently they are the subject of document image analysis. [Table 2.2](#) summarizes the methods and types of impact printing.

### Relief Printing

One of the oldest forms of machine printing came from the text or the images being carved into wood blocks, which were then inked and pressed onto the paper in a process called *xylography*. The first samples of this type of printing date to the eighth

**Table 2.2** Impact printing methods and types. Black represents the printing device and gray is the ink

#	Method		Type(s)
1	Pressure against a relief (raised) surface		Letterpress, typewriter
2	Chemical action from a planographic (flat) surface		Lithography, photo lithography, offset printing, ditto machine, hectograph
3	Lifting ink from an intaglio (engraved or depressed) area		Gravure, rotogravure
4	Seepage through a stencil or screen		Silkscreen, mimeograph

**Fig. 2.4** Relief print samples: (a) letterpress, (b) typewriter, (c) dot matrix, (d) carbon copy, and (e) carbonless copy

century AD in China. The first printed book was the “Diamond Sutra” printed in China in 868 AD, first in scroll form, then as books. A collection of 130 volumes of classics were printed in 953 AD. This printing method became common in Europe in the fifteenth century. A full page of print was carved into each block. The production of the wood block was labor intensive, so the quantity of material printed with this method was small, but the blocks could be used for an extended time to produce many copies of each printed page.

The time needed to produce a page of type can be reduced by producing smaller blocks for each character or letter and then rearranging them for each page. This is called *letterpress* (Fig. 2.4a). *Movable type* was first invented in China by Pi-Sheng around the year 1040 AD. It used a ceramic font type. This was found to be brittle, and the font needed for the Chinese script was too large. Due to the value placed on Chinese calligraphy, it was not a viable invention. In the year 1234, a movable-type printer using a bronze font was invented in Korea and was used in parallel with xylography. The metal font allowed finer type detail than xylography. However, the bronze type produced sharp edges, and the preferred paper in Korea was soft. Koreans alleviated the problems by placing the paper on the type and rubbing it to contact the type and transfer the ink. From 1436, they also experimented with type made from softer lead. While the script in use in Korea at that time was also Chinese, the political will to disseminate large quantities of printed material made movable type a viable instrument such that a second font set was created in 1403 and a third in 1420. Ultimately, movable-type printing systems did not retain long-term viability due to changes in political leadership.



Although not its first inventor, the invention of movable type is usually credited to Johannes Gutenberg. In Mainz, Germany, in 1440, he created a movable-type printing press. As the Latin character set is relatively small, it is better suited than Chinese to this printing method and was adopted by others in the west. Printing reached Cologne, Basel, Rome, Venice, Paris, Nuremberg, and Utrecht by 1470 and England, Spain, and Portugal by 1480.

Text was laid out by a typesetter or compositor who would assemble individual pieces of type into a form held together by a frame or chase. Multiple instances of each letter are needed for each page. Care was needed to assure the typefaces would be planar and each piece would have equal pressure against the paper across the whole surface. The filled form was placed on the bed of the printing press and inked. A damp piece of paper was placed on top of it. The metal plate (platen) was lowered to bring the paper type in contact with the paper. To resist the scratching of quill pens, European paper was tougher than Korean paper. As a result, instead of rubbing the paper on the type, the printer used the pressure of olive or wine presses to complete the ink transfer. A tin and lead alloy was used for the type. This makes a firm enough material to press onto the paper to make the print, but not so hard as to cut the paper nor too soft where the type would deform after a limited amount of use. Still type was prone to being dented, and small variations in characters can often be seen in letterpress documents.

Creation of the type started by cutting a metal punch to produce a raised, reversed image of the letter carved onto the end of a hard metal bar. This was driven into a softer metal to form an indented, reversed version of the letter called the matrix. The matrix was placed into a mold over which molten metal was poured to create individual pieces of type, again raised and reversed. Many pieces could be made from one matrix. With the use of molds, the shapes of the characters in a given document became repeatable. This is beneficial to today's optical character recognition technology in that the pattern to be matched can be better predicted. Handwritten writing styles existed before the printing press, such as Uncial and Carolingian. Initially, the characters mimicked these writing styles, but over time, a greater variety of fonts were developed (see section in ►[Chap. 9](#) (Language, Script and Font Recognition)).

While movable type reduced the time spent carving blocks for each page, a significant amount of time was still needed to select individual-type pieces, place them in the frame, and then return them to the case after use. This process was partially automated with the introduction of the *Linotype* composing machine in 1886. The magazine did not hold type but instead matrices or molds to cast typefaces. Based on operator input, a whole line of type would be cast at a time which would then be set into the frame or chase. In 1897, a patent was issued for *Monotype* which also cast type upon demand, but cast individual-type pieces. For both methods, after printing the type would go back into the melting pot for reuse. These advances allowed for easier production, which increased production volume. Linotype was used primarily in newspaper and magazine production and Monotype mostly in book production. While these systems changed the composing method, they didn't change the printing method; however, every letter used in printing was new and sharp, thus increasing print quality.

a The quick brown fox slyly jumped over the lazy dog.

b The quick brown fox slyly jumped over the lazy dog.

c The quick brown fox slyly jumped over the lazy dog.

**Fig. 2.5** Typewriter samples: (a) manual typewriter, (b) electric typewriter, and (c) electric typewriter with ink tape instead of ribbon

The printing process was accelerated by the introduction of the *rotary press* which used cylindrical plates. With cylindrical plates, the step of lifting the plate was removed from the process. Rotary presses could roll the cylindrical plate across a flat surface, or two cylinders, one with type and one blank, could oppose each other. Individual sheets could be fed, or paper could be fed from a roll or web, into the rotary press, further accelerating the process. The cylindrical plates could not be made by placing discrete pieces of movable type into a form. Instead, methods of *stereotype* and *electrotype* were used to form the plate. With stereotyping, a duplicate was made from a flat plate of type by pressing a papier-mâché or fiberboard matte to the plate to make a mold from which molten lead could be cast. That mold could be bent allowing the production of cylindrical in addition to flat plate duplicates. For electrotype, an impression of a plate was made in a wax-coated foil or sheet of thermal plastic material. The mold was polished with graphite or sprayed with metallic silver to make it conduct electricity. It was then immersed in a solution of copper sulfate and sulfuric acid, and a layer of copper was deposited by electrolysis to a thickness of 0.015 in. or 0.38 mm. The copper is separated from the mold and backed with a lead-based alloy and mounted. Stereotype was used most often in the production of newspapers. Electrotype was used mostly for magazines.

*Typewriters* were initially developed in 1714 to enable the blind to read by embossing paper with letters. In their established form, typewriters produce printing for sighted people, Figs. 2.4b and 2.5. They use a premise similar to letterpress of inking a raised piece of type and pressing it to the paper, but the characters are pressed to the paper individually instead of a whole page at once. The units containing the typeface are located on the ends of rods. When a key is pressed, the rods rise and strike a ribbon impregnated with ink against the paper. The first American typewriter was patented in 1868 and contained 44 keys for the 26 uppercase letters, the numerals 0–9, and some punctuation. Typewriters became common in commercial offices in the 1870s. The carriage that holds the paper is moved horizontally after each character is struck. Initially, the carriage progressed a uniform amount after each keystroke, resulting in a uniform spaced type. After the user presses a carriage return key, the carriage will rotate vertically and return to the left margin to ready the machine for the next line.



A shift bar was added to allow two character symbols, usually an uppercase and a lowercase form of a letter, to occupy each type unit. This increased the character set to 88 symbols. In 1961, a model was introduced that replaced the type on long rods with a ball of type that would rotate to select the character symbol. Another variation used a disk of type. These balls and disks were easy to remove, enabling users to change fonts. Typewriters are used primarily with the Latin character set, but are also found for the Greek and Cyrillic characters. A typewriter that was capable of printing Japanese kanji-kana text was developed in 1915. Due to the larger symbol set, it was a much larger and more complicated machine and was not widely adopted.

The typeface is a flat surface which is pressed on a cylindrical surface. The cylinder is soft, so within a small area, the full shape will reach the paper. However, maintenance of the units varied, and when the calibration is extremely poor, or the machine is operating at a high speed, the type may not fully hit the paper, resulting in partial characters and a nonuniform baseline of the text (Fig. 2.5a). Vertical misalignment of the characters could result from poor calibration of the keys or from the key being pressed when the type tray was still in transit from one position to the other. Early typewriters were fully manually powered so different amounts of pressure would be imparted on each stroke. The introduction of electric typewriters resulted in a constant amount of pressure on each character giving greater uniformity to the resulting image (Fig. 2.5b, c). Also, the loops in the characters were prone to filling either because the typeface was gummed up with ink from contact with the ink ribbon or because the type pressed down the ribbon far enough that not only the ribbon in front of the type came in contact with the paper, but from being stretched the ribbon within a loop and near the border also made contact (Fig. 2.4b).

Corrections could be made by painting the paper with an opaque white paste correction fluid and retyping on top of the corrected area. A surface that was not smooth would result. Alternatively, a slip of tape or paper with a film of white “ink” could be inserted between the ribbon, and the paper and the character could be retyped covering the ink in the character “zone” with white, then a different character could be typed on top of this corrected area. Impressions of both characters would result on the paper, even though only one would be optically visible on the image surface.

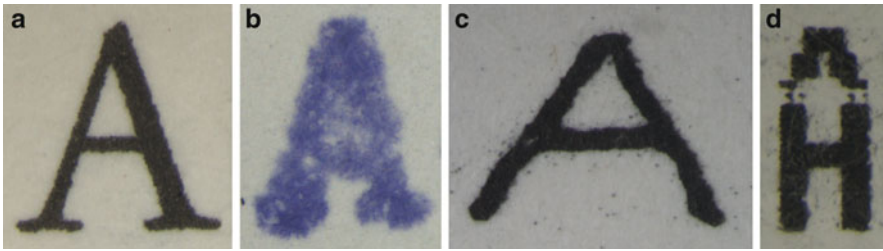
Later developments allowed the input to the typewriter to be remotely located, making typewriters early output devices for computers. *Dot matrix printers*, invented in 1964, were created to be output devices for computers. In functionality, they are a variation of a typewriter. The information is transferred to the paper via a rectangular grid of pin-shaped type that impacts an ink ribbon. A digital control signal from a computer determines which of the pins are activated. Because of the addressability, the shape of the character can be changed through software, and a greater variety of character shapes are possible than with a traditional typewriter. Initially, the printers had a grid of  $5 \times 7$  pins (Fig. 2.4c). This grew to 24 rows of pins. The dot density ranged from 60 to 360 dpi (dots per inch).

*Carbon paper* was a common way of producing two to four copies of a document. Pieces of paper coated on one side with a thin film of carbon black, initially bound to the paper by a thin film of oil, would be placed between pieces of blank paper. When the paper was pressed with high pressure, such as from a typewriter or pen, the carbon would transfer to the blank paper beneath it (Fig. 2.4d). Carbon paper was made at least as early as 1823 but was only made common with the introduction of typewriters because it did not work well with nib pens. At one time, the technology was so prevalent that the labels “cc” and “bcc” (for carbon copy and blind carbon copy) still appear in many email applications, where there is no carbon in sight. Because the carbon was prone to transferring to surfaces not intended, a similar process called *carbonless copies* could be made by using autocopy or NCR (no carbon required) paper, a set of papers containing encapsulated dye or ink on one piece and a reactive clay on the next on the facing side (Fig. 2.4e). When the two came in contact with enough pressure, such as that from a pen or typewriter, the microcapsules would break and spill their dye. Both these techniques produce lighter copies with broader strokes at layers further down the stack. The width of the stroke in the copies can be wider causing touching and filled characters.

### Planographic Printing

*Lithography* is a planar process where the printing surface is neither raised nor recessed. Lithographic printing started with the application of an oily substance on a stone printing plate, from which the process received its name (litho, stone; graphy, writing). The stone is ground to a smooth even finish. Grease is applied where color or ink is desired in the final image. An artisan can draw directly on the stone with a grease pencil. During printing, the stone passes beneath a dampening roller to apply water to the nongreasy parts, then beneath inking rollers where greasy ink is attracted to the grease of the template but repelled from the water. The stone is then pressed to a piece of paper where the ink is transferred to the paper. Zinc or aluminum plates can be substituted for the stone. These plates are lighter and take less space to store. They can also be bent into cylinders allowing use in a rotary press configuration, which accelerates the printing process.

In *photolithography*, a photographic negative of the image to be printed is made. The negative can be made from any source or collection of sources of images, including handwritten sources, or printed samples from any printing process. These can be composed as a single unit or by placing several different pieces from different sources together to form a single page of print. This process is useful to reprint books that are out of print, since the publisher would not need to repeat the time or expense of typesetting. The plate is sensitized with a bichromated emulsion which is hardened when exposed to light usually from passing light through a film negative. The design appears only where light passed through the film. The plate is rolled with a greasy ink and then washed to remove the unexposed emulsion and its ink layer, also filling in the non-inked areas with water. Printing mirrors the approach used in lithography, dampening and inking the plate at each revolution. The very thin film of ink that is used requires a very smooth paper. It can also lead to low depth of tone in screened images. An alternative deep-etched photolithography method exists that



**Fig. 2.6** Planographic printing methods (a) Offset lithography, (b) ditto machine, (c) hectograph, and (d) thermal printing

produces etching deeper than with regular lithography. Because this is a chemical rather than mechanical process, it is referred to as “cold type,” because no molten lead is involved.

In direct lithography, because the plate comes in direct contact with the paper, the image on the plate must be reversed from the desired final document image. *Offset printing* or indirect lithographic printing works on a similar premise as in direct lithography printing, but the ink is first transferred from the metal plate to a rubber plate and then transferred from the rubber plate to the paper (Fig. 2.6a). The rubber surface squeezes the image into the paper rather than placing it on the surface of the paper. This offers the advantage of enabling the use of cheaper papers since the rubber can press more firmly to the paper, opening up possibilities beyond using paper with more costly glossy or polished finishes. The extra transfer step can, however, introduce minor distortion to the character shapes, especially on interior and exterior corners. Conversion from letterpress to offset in larger print houses occurred in the 1970s.

Another planographic printing technique, *spirit duplicators* or *ditto machines*, was used often for small batch printing. Invented in 1923, a two-ply master was used that included one layer that would be written or typed on and the second which was coated with a layer of colorant impregnated wax. When the top sheet was written on, the wax was transferred to the back of that sheet. In the printing process, a thin coating of solvent (typically ammonia or methylated spirits, an even mixture of isopropanol and methanol) was applied to the master sheet, and a thin layer of wax would then transfer to the blank paper when contact was made. The wax, and thus the resulting print, was usually a purple color (Fig. 2.6b). The machines were limited to 500 copies for a single master.

The *hectograph* machine, also known as the gelatin duplicator or jelly graph, was invented in the late 1870s. Ink containing an aniline dye was used to write on a sheet of paper. The document image was transferred to a shallow tub of gelatin mixed with glycerin by placing the inked side of the paper onto the bed of jelly. A clean piece of paper would be pressed onto this gelatin, and some of the ink would transfer to the paper (Fig. 2.6c). Twenty to eighty copies could be made from each master, with later copies being lighter.

*Thermal printers* use special paper that when heated, turns black. Invented in the early 1970s, the printers function similar to dot matrix printers, except the grid of dots are heated and placed in contact with the paper and are not forcefully pressed into the paper. The heat usually results from an applied electric current. The paper is impregnated with a mixture of a dye and a matrix, so no external ink is used. When the matrix is heated above its melting point, the dye reacts with the acid and changes to a colored form. Thermal printers are faster than dot matrix printers and are usually smaller and consume less power. Thermal printing was commonly used in fax machines and is still frequently used in retail receipts (Fig. 2.6d).

### **Gravure/Intaglio Printing**

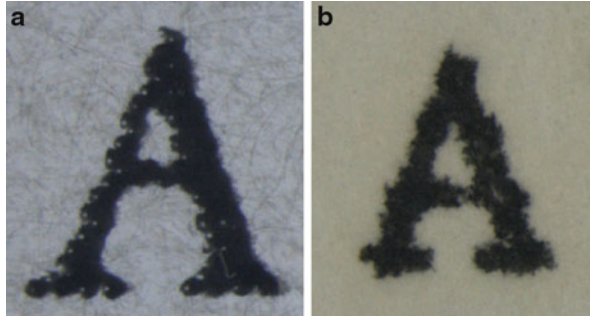
In *gravure*, also known as *intaglio* printing, the printing surface is recessed, and ink that fills the recessed areas is transferred to the paper. A plate is etched, or cut; coated with ink; and then wiped clean. When the paper is pressed against the plate, the ink in the wells is transferred to the paper. The amount of ink can be controlled by varying the depth of the engraving, the width of the engraving, or both. Similar to in offset printing, plates were etched with acid that eroded the metal in the desired printing locations to depths proportional to the hardness of the gelatin. *Photogravure* results when a photographic process makes the engraving. Photogravure has a higher initial cost for plates than photo-offset printing but produces higher quality documents over a longer run batch and was commonly used for high distribution magazines containing high-quality images. When the printing surface is on a cylindrical form, it is called *rotogravure*.

When a photograph or other continuous tone image will be included, the entire drum is pre-etched with a screen pattern (see section “[Multitone and Color Printing Dithering and Screens](#)”). Then, the desired text and images are etched on top of this. The variable size of the ink wells can produce a greater variety of tones achieving very high image quality. However, there is a tradeoff in quality in that a serrated or jagged edge typically appears on lines and text (Fig. 2.7a). Text also loses sharpness because of the watery consistency of the ink.

### **Silkscreen Printing**

*Silkscreen* is a printing method that fits into the fourth impact printing method of applying ink to substrate. Here, ink is passed through a stencil. The stencil has a stabilization mesh made of either silk or sometimes a metal screen to support the gutters. The screen is stretched across a frame producing a printing form. Material is applied to portions of the screen to block the pores. This can be done by cutting a pattern from a lacquer backed sheet to remove the backing and placing that on the screen. A screen can be coated with an ultraviolet cured varnish that is marked with a photomechanical stencil design that is washed away exposing the screen. Photographic development or lithographic crayons can also be used to create the stencil. A squeegee is passed over the screen to force the paint or ink through

**Fig. 2.7** Printing methods  
(a) Gravure and (b)  
mimeograph



the porous portions of the screen. The paint or ink has the consistency of syrup. Silkscreening is used primarily for posters, displays, and fine art reproductions, as well as on diverse surfaces such as bottles, wall paper, and fabric.

In offices, the *mimeograph* or *stencil duplicating machine* was used to print documents that needed more copies than a typewriter with carbon paper could produce, but not enough that turning to other techniques was reasonable (Fig. 2.7b). The typical capacity for one stencil was 200 copies. This machine, invented in 1890, was based on Thomas Edison's patent for autographic printing that is used to make the stencil. The stencil was usually made on a typewriter with the ink ribbon removed, so the type would directly impact the stencil and displace the wax. Ink is pressed through the stencil master and then onto the paper. Over the course of the run, the interiors of closed loops would often fall out, assuring that the loops of those characters would be filled with ink in subsequent copies. This technique produces documents with a higher quality than the ditto machine that was common at the same time for small batch printing.

### Nonimpact Printing

In the twentieth century, methods of printing that did not involve physically pressing a template onto a substrate were developed. These were made possible by the development of computers and electronics able to move the ink relative to the paper. The development of these technologies also increased the flexibility of the printing system, removing the need to create masters for everything that would be printed. The term nonimpact came when computer-controlled printing methods of dot matrix and electrophotographic (EP) printing were compared. Dot matrix printing impacted a pin to the paper and EP printing did not, even though in EP printing contact is made between a roller and the paper. Nonimpact printing came to refer to all printing methods that were masterless and does not always refer to printing processes that do not involve any contact with the paper. An advantage of nonimpact technologies is their ability to customize documents between rotations. Since these technologies don't have a fixed master, they can produce a greater variability within a print run even if a constant image is desired.

## Electrophotographic Printing

*Electrostatic* printing or *xerography* (from the Greek “dry writing”) was first publically demonstrated in 1948 as an image copying process. The process starts by charging the photoconductor (PC) plate. Light discharges part of the PC. Oppositely charged toner particles are attracted to the remaining charged areas. Paper is charged, and as it comes in contact with the PC, the toner particles are transferred from the PC to the paper. Application of heat and pressure affixes the toner to the paper. Excess toner particles are removed from the PC, and the PC is neutralized for the next cycle.

Computer-generated images are transferred to the PC by a row of LEDs (light-emitting diodes) or by a rotating mirror reflecting a laser beam onto the PC. In this case, the light source “paints” the PC with charge that attracts the toner particles. The use of a laser gave rise to the name *laser printer*, even though a laser is not always the light source involved. The print quality depends on the resolution or addressability of the imaging system. The basic procedure has the PC discharging one row at a time. The row is divided into addressable units which, in theory, are either on or off. This has the tendency to make diagonal lines “jagged.” When a laser is used, varying the laser intensity in finer gradations than the base addressable resolution allows smooth appearing lines to be produced at just about any angle (Fig. 2.8a).

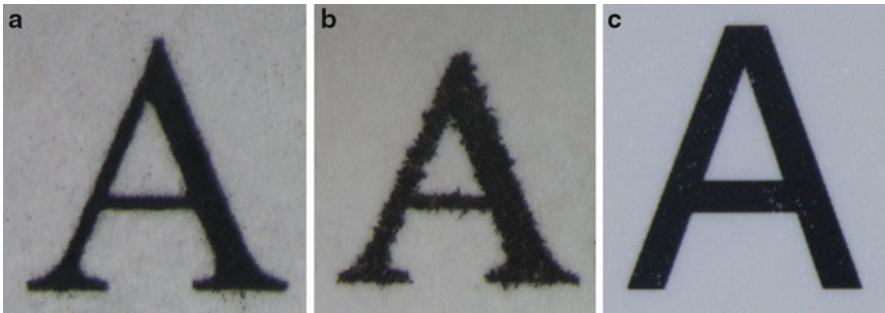
For a *photocopier*, it is easiest to use the light reflected off the white part of the paper to discharge the PC. The same principle applies too when xerography is used to print microfilm or microfiche images. When using xerography as the initial printing mechanism, as with a laser printer, the PC is discharged where the ink is to accumulate. These two processes use the same basic principles, but opposite charges. Each cycle will deteriorate the quality of the resulting image, Fig. 2.9a–d, because degradations from both the printing process and the scanning process contribute to the resulting image; see the section “[Acquisition Methods](#).”

In a *facsimile* or *fax machine* images are spatially sampled in a location remote to the photoconductor. The samples are sent by telephone line to control a light source in a printer at a remote location. Because of the low bandwidth of the telephone line, facsimile machines usually sample the image at a lower resolution to transfer fewer samples. The printed image will thus have a lower quality than printing of an original on a laser printer, even though they use the same printing technology, Fig. 2.9e.

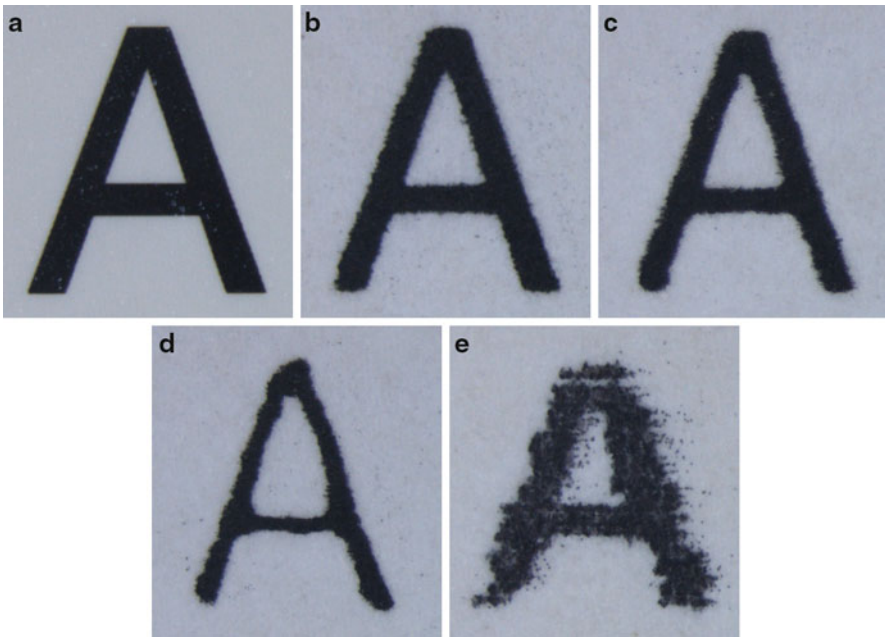
## Inkjet Printing

*Inkjet* printing is another nonimpact printing technology (Fig. 2.8b). Inkjet printing sprays a quantity of ink from a fine high-pressure nozzle onto the paper where it is absorbed. Inkjet printers use either a thermo- or piezoelectric method to control the ink flow. In the thermal method, the ink is heated to the boiling point, where built-up pressure in the nozzle tube ejects the ink. Alternatively, a piezoelectric-controlled pump-like device manually ejects the ink. In some instantiations, the ink will form a mist and produce a spattered appearance. Modern inkjet printers have developed techniques that reduce this effect. Inkjet printers may have a single nozzle for each





**Fig. 2.8** Examples of nonimpact printing methods. (a) Xerographic, (b) inkjet, and (c) image setter



**Fig. 2.9** (a) Original image. (b)–(d) Effect of repeated photocopying, copies 1, 2, and 7. (e) Output of fax machine

color that moves back and forth across the sheet of paper depositing ink or an array of nozzles that span the paper width depositing ink in parallel. This array-based process increases the inkjet printing speed to the point that is practical to customize the product from piece to piece in mass production runs by changing the document content.

### Other NIP Methods

An *imagesetter* produces ultrahigh resolution images of documents by combining computer output with photographic technology (Fig. 2.8c). Imagesetters are often used to get high-quality proofs before another high-resolution printing mechanism is used that requires more time and materials in the setup process. Operating similarly to laser printers, imagesetters use a laser light source to expose film instead of (dis)charging a PC. The film is a high-sensitivity continuous tone film or a light room contact film. The film has a protective layer on top of a light-sensitive silver halide layer. Below this is a carrier base and then antihalation backing. After development, a high-quality black and white document image is visible. With a spot size of 7–45  $\mu\text{m}$ , resolutions up to 8,000 dpi are possible.

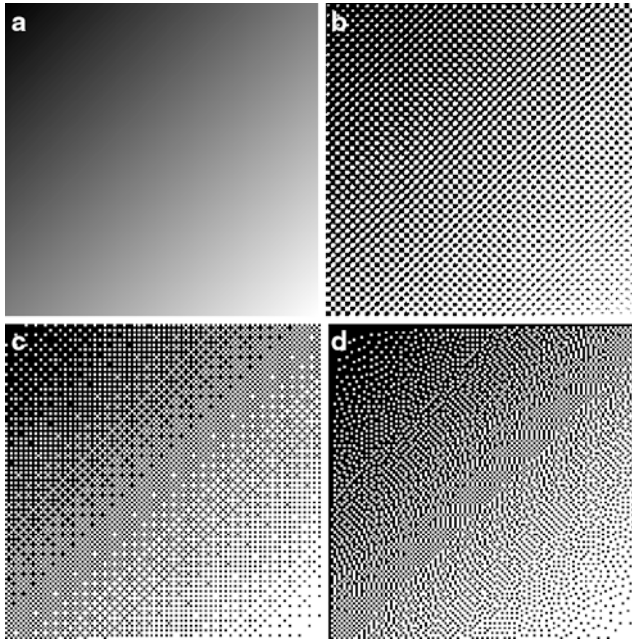
*Microfilm* and *microfiche* are two types of microforms. The images of documents are reduced to about 4% of the original document size through a photographic process onto a photographic film substrate. Microfilm stores the images on reels of film. Microfiche stores the images on flat sheets of film. This storage mechanism was used for documents that were not often referenced but that had a large number of pages. This could significantly reduce the physical storage space required for these documents. The American Library Association endorsed this method of information storage in 1936. It is relatively inexpensive to make multiple copies of the films, and the films allow access to some documents that might be too rare or fragile for general access. This is similar to the reasoning behind many document digitization projects today. Pages from a microform could be viewed on readers, and selected pages could then be printed at original size. Microfilm itself gets scratched very easily. This deteriorates the image quality. Most microfilms are made with the silver halide process. These films if properly stored have a life expectancy of hundreds of years. However, in reality they have a much shorter shelf life than paper documents. After a few decades under certain environmental conditions, the old acetate films begin to slowly chemically decompose and produce a vinegar scent. Charts exist for acetate film that explain the relationship between temperature, relative humidity (RH), and the “vinegar syndrome.” For example, a fresh acetate film that is kept in a room maintained year round at 70°F (21°C) with a RH of 50% will last for 40 years before the film exhibits the “vinegar syndrome.” At this point, the film is still useful, but deterioration begins to accelerate, requiring imminent action to avoid the loss of content.

### Multitone and Color Printing, Dithering and Screens

When images in more than the bi-level black and white are desired, the multitone or color effect is achieved by using a *dithering* or *screening* process. Here, the appearance of multiple levels of lightness or darkness is achieved by varying the percent of a viewing zone containing ink. Visually, the eye averages the image to produce the appearance of a range of tones related to the fill percentage.

Historically, this was achieved in monochrome by an engraving process where the image was converted to a texture with varying thickness lines. This only allowed coarse approximations, and creation of these drawings was very time consuming.





**Fig. 2.10** Examples of dithering methods. (a) Original image, (b) AM halftoning, (c) FM halftoning, and (d) error diffusion halftoning

In 1881, Georg Meisenbach introduced a process where a fine mesh or screen of diagonal lines crossed at right angles on a sheet of film or glass was placed in front of a photosensitive film. As the light from the image is passed through the screen, the lens breaks the image into thousands of small points of light which are stronger where the image is lighter and weaker where the image is darker. This is known as *AM halftoning* as the dots are uniformly spaced but have variable size and thus a variable brightness amplitude (Fig. 2.10b). Screens range from 65 lpi (lines per inch) for newspapers to 600 lpi for fine art and photography reprints. Most illustrated books are printed with 133 or 150 lpi line screens. At 150 lpi (60 l/cm), the eye can no longer detect the individual dots from a normal viewing distance of 12 in. (30 cm).

The percentage filled, and thus apparent darkness, can also be controlled by changing the number of dots per unit area. This technique is known as *FM halftoning*. The dots are of constant size and can be arranged in an orderly, uniformly distributed pattern or stochastically arranged to avoid the appearance of certain patterns that are noticeable to the eye and detract from the uniform field (Fig. 2.10c).

More advanced methods have been developed to produce more pleasing results. However, these come with a higher computational cost. One common method is *error diffusion*, which is based on sigma-delta modulation from one-dimensional

signal processing. The image is scanned and the pixel value is predicted based on linear combinations of the neighboring pixels. The sign of the error in the intensity value estimate determines whether or not to print a black dot at each location (Fig. 2.10d). Other methods make extra passes to remove certain dot combinations that are known to stand out to the human visual system. A process called direct binary search checks the dot arrangement and makes small changes so no specific texture artifacts are visible.

The halftoning algorithms assume that the dots will have a certain size and color when the algorithm is implemented to arrange them. The printing processes can cause the dot size to vary from the addressable resolution or the dot shape to differ from that modeled. This can make the total area appear brighter or darker in an artifact known as *dot gain*.

More than two gray values (on and off) are possible in many print technologies by varying the ink film thickness in a process referred to as *density modulation*. Combining density modulation with halftoning produces a greater number of apparent gray levels than halftoning with just two primary colors.

## Color

The visible spectrum includes light with wavelengths in the range of 350–750 nm. Light of all wavelengths is incident on a surface, and some wavelengths are absorbed while others are reflected. Those that are reflected then stimulate the cones in the human eye producing the appearance of light. There are three types of cones in the eye, each of which has a different sensitivity to, or “weighting” of, each light wavelength. How each cone is excited determines the color we perceive.

Similar to being able to create a range of gray tones from two base colors, black ink, and white paper, a range of colors can be created by combining a small set of color primaries with the white paper. To replicate the perception of a gamut of color, color documents are printed with three reflective color primaries, cyan (C), magenta (M), and yellow (Y). By mixing these three colors in small areas, the perceived color is the color that is *not* absorbed by the inks on the paper in a given region. Because the primaries are imperfect, when mixing all primaries not all the colored light will be absorbed, and usually a dark grayish brown results instead of the desired black. Thus printing systems usually also print black (K) as a fourth primary producing the CMYK color system. The space of colors spanned by a set of primaries is called a *color gamut*, and due to different ink primary properties, each printing and display system will have its own color gamut.

To break a color image into its color primaries for analog printing processes, the original image is photographed through a series of color filters that absorb colors of which it is composed and permits the remaining colors to be recorded. Color digital images come from scanning the image with a grid of sensors sensitive to a single color primary through smaller scale color filters.

When color images are printed, each color layer is screened separately. Use of AM halftoning is most common. The four colors are printed each at a different angle to avoid introducing interference or *moiré* artifacts. The most apparent color, black, is printed at an angle of 45° to be less visibly distracting. The least apparent color,

yellow, is printed at an angle of 10°. The remaining colors, cyan and magenta, are printed at angles of 75° and 15°, respectively.

### **Digital Printing, Electronic Books and Displays**

Computers and electronic technology have allowed document creation to take on many new forms. Some of these have increased the automation of the document creation process, so increased speed and precision are possible. Some have evolved the physical printing methods or allowed the document to be printed from a remotely created source. Another development has made the image itself digital from the start. ePens, discussed in the section “[Inks](#),” allow direct input of the pen trace into the computer. Displays allow the image of the document to be seen either directly from the word processing equipment (see [▶Chap. 23](#) (Analysis of Documents Born Digital)) or after being digitized from hardcopy.

Electronic books have existed almost as long as computers, but the ready access to the material, and its adoption by society, is recent. Books and other documents have been available at libraries or through distribution on CD ROMs for viewing on computer displays. The jump to portability has come through the introduction and successful marketing of *eReaders* such as the Amazon Kindle, Barnes & Noble Nook, Sony Reader, and Apple iPad. eReaders are small electronic devices with a display about the dimensions of a medium paperback book. eReaders have enabled users to read traditional book content as eBooks. Some people appreciate being able to carry the content of several books in a physically small device. The increased speed with which new content can be accessed or updated, whether locally or remotely created is another advantage. They are also able to reduce waste from printing documents, like newspaper, that are not intended for long duration use. For these document types, the content analysis becomes the research focus. Further discussion of this type of document is covered in [▶Chap. 23](#) (Analysis of Documents Born Digital).

### **Displays**

Images of documents before printing and after acquisition from their printed form are made visible through *electronic displays*. Displays can be emissive, transmissive, or reflective. The *emissive display* technology that dominated the market in the 1900s is the *Cathode Ray Tube (CRT)*. CRTs rely on cathodoluminescence. An electron beam is scanned across a phosphor screen, which then luminesces. The eye can retain memory of the light for a period of time, so the screen does not need to be imaged over the whole surface at once. The vertical information is refreshed 60 times per second in North America, and 50 times per second elsewhere. The signal is usually interlaced by scanning odd and even numbered lines in separate passes to increase the vertical speed, while avoiding flicker. For color images three electron beams are used which hit phosphor dots that luminesce either red, green, or blue when struck by the beam. The dots are arranged in triads and at angles such that only the correct electron beam will strike it. The control electrode strengthens or weakens the electron beam to vary the brightness level. CRT displays are characterized by

large bulky vacuum tubes that contain the electron beam, with a depth close to the size of the screen width. Flat panel displays have a much smaller depth profile and often use less energy. *Plasma displays* operate like fluorescent lights with a series of RGB pixels in long horizontal strips. The alternating current generates a plasma. The gas in the discharge radiates ultraviolet light when excited. This UV light then excites color phosphors via photoluminescence. *Light-emitting diodes* (LEDs) are another emissive display technology. LEDs of red, green, and blue colors are arranged in pixel clusters and produce the desired image. LED displays are most predominant as score boards and large advertising boards. Organic LEDs (OLEDs) are developing with smaller sizes and often flexible displays. These are used in some cell phones and PDAs (personal digital assistants) and are occasionally found in displays up to 19 in. (43 cm).

The most common *transmissive displays* are *liquid crystal displays* (LCDs). The liquid crystal is polarized, and based on an applied electric field, the direction of polarization can be changed. This is placed on top of another polarized layer. When the liquid crystal is polarized in the same direction as the base layer, light is permitted to pass. When the liquid crystal is polarized perpendicular to the static layer, light is blocked. Red, green, and blue color filters are applied in sets of three small vertical rectangles for each pixel, so the transmitted light is colored. LCDs require a back light source. This is most commonly cold cathode fluorescent light operating at high frequency, although sometimes the light source is a panel of white LEDs. The occasional use of LEDs as the light source sometimes leads those displays to be incorrectly called LED displays.

*Reflective displays* are used on many eReaders such as the Amazon Kindle and the Barnes & Noble Nook. They rely on an external light source, either the sun or a lamp, to provide the light. This light bounces off the display which in some areas absorbs the light and in others reflects it. This is the same principle used with paper documents, so these displays are sometimes referred to as *ePaper* or *eInk* depending on whether the white or the black part of the display provided the name. Reflective displays generally take less energy because they do not need to produce the light and because they do not need to be constantly refreshed to maintain the image but can be readdressed to change the image content. The reflective displays often produce less eye strain because the light is less harsh, and they are low glare and so can be used in a variety of lighting conditions. Another characteristic of these displays is that they can be much thinner and lighter than other displays and sometimes flexible like paper. Two main approaches to ePaper have been developed. One uses technology developed by Xerox PARC in the 1970s, which was spun off to Gyricon Media Inc. and 3M. This consists of spheres of diameter approximately 100  $\mu\text{m}$  that are painted half white and half black. This gives an effective resolution of about 250 dpi. The spheres change their orientation based on the application of an electrical field. The other technology, developed by Massachusetts Institute of Technology (MIT) and eInk Corp., fills microcapsules with a dark liquid and many charged white pigment particles. Depending on the electrical field the charged particles will move either to the top of the capsule near the surface of the paper to produce white or to the bottom so the dark fill is visible.

## Acquisition Methods

People use digital computers to perform document image analysis on digital images. There are many methods possible for converting a physical spatially continuous document image into a digital representation. These methods have developed to accommodate the broad range of document types that are of interest.

### Flatbed Scanner and Fax Machine Acquisition

Documents are traditionally printed on flat paper which led to the development of the *flatbed scanner* as the primary method of acquisition. For acquisition on flatbed scanners, a page is placed face down on the glass platen. The scanner has a scanning bar, which extends across the width of the page and moves down the length of the page taking a series of pictures, one for each row of the total image. In some scanner configurations, the light and sensor array are stationary, and the platen or paper moves. To take a picture of one row in the page image, a light source shines on the image. A series of mirrors redirects the light reflected from the original image to a focusing lens. This lengthens the effective focal length of the lens. The light is then directed to the sensor array, which collects charge as the bar moves continuously along the page. The optical resolution in the scanning bar direction depends on the distance the scanning bar moves during the exposure time of the sensors. The time spent traversing the distance corresponding to 1 pixel length is set such that the appropriate amount of charge can be accumulated. Resolution in the transverse direction is determined by the number of sensor elements in the linear array and the magnification of the optical system. The readings from each row are accumulated to form the two-dimensional (2D) bitmap for the image, which is then passed from the processor in the scanner to the computer. Some scanners have a 2-D array of sensors, so data for the whole image is acquired in one step, usually without moving parts.

The sensors can be either *charged couple devices (CCDs)* or *complementary metal-oxide silicon (CMOS)*-based devices. Both sensor types have a charged silicon layer that when exposed to light will dislodge electrons from the silicon lattice that when “counted” by the electronics is converted to a voltage that has a monotonic relationship with the amount of light reaching the sensor area. This voltage is then converted into a digital signal. CCD sensors pass the charge reading out along a row, cell by cell. CMOS sensors can be addressed in two dimensions so all values can be read essentially at the same time since they acquire an image in a 2-D array of sensor elements, removing the need to move a 1-D array of CMOS sensors relative to the page to acquire a 2-D image. CCDs are found in older equipment and higher end equipment. Until recently, the quality of the CCD-based sensors was notably higher than the quality of the CMOS sensors. CMOS sensors were preferred for low-cost applications because the manufacturing process for CMOS sensors could be done at the same time as processing for the accompanying circuitry.

The quality of the CMOS sensors has increased to the point where they are being used across a broad range of products and are becoming common even in high end devices.

The physical 2-D document image is converted into a two-dimensional array of numbers (pixels) during the scanning process (Fig. 2.11). The actual darkness of the writing substrate (paper) and the ink together with the brightness of the illumination produce the recorded pixel values image. When the document is acquired on a flatbed scanner, the illumination is carefully controlled to produce a near uniform value to get a better representation of the document. Misfeeding of a page through a scanner can cause image distortion, such as skew.

During acquisition, light reflected off the document passes through lenses and is focused on an array of sensors. Each sensor returns a scalar value which is the average of the light over a region in the document. This can be mathematically modeled as 2-D convolution with impulse sampling. The light received at each sensor will come from a limited region on the paper and will also usually come from that area nonuniformly. The *point spread function* (PSF) describes how the light received at the sensor is spatially weighted. This can be measured using a “knife edge” image, which is a 2-D step input, and differentiating the response. This procedure has been standardized in the International Standard Organization’s ISO 12233 [1] procedure.

Noise is often unintentionally added to the sampled image. The noise can come from the variations in the paper and ink intensities as well as sensor noise. This is usually modeled as being additive i.i.d. Gaussian noise. Sometimes the noise that results has more of the characteristics of shot or impulse noise.

The acquired signal has continuous valued intensities. These are quantized to convert it to a digital signal. Most scanners initially store 16–48 bits of gray levels but then convert it to 2 or 256 levels per color channel based on the user’s choice of acquisition settings. An efficient quantization scheme is needed to retain the greatest amount of information. To determine the level reduction, either the scanner or user selects *brightness* and *contrast* values. When the scanner performs this function, it uses an automatic algorithm based on data sampled from the image.

A high (low)-brightness setting will map the input reflectances to a higher (lower) numerical value in the gray-level digital image. Contrast is the magnitude of the difference in output gray level between the brightest white and the darkest black. Therefore, a high-contrast setting will allow a larger range of gray values in the image than a low-contrast setting. This is used to accentuate highlights and shadows (Fig. 2.12). To avoid saturation, the scanner brightness and contrast settings should be adjusted such that the range of reflectances from black ink to white paper does not exceed the 0–255 range. The brightness and contrast can also be changed in post processing (see ►Chap. 4 (Imaging Techniques in Document Analysis Processes)), but if the scanner settings cause over- or undersaturation because the reflectances map to a gray level outside the 0–255 range, this cannot be compensated for later. The gray-level mapping function for a given scanner can be determined by scanning a test chart with calibrated gray-level reflectance patches and noting their gray-level value in the resulting image.

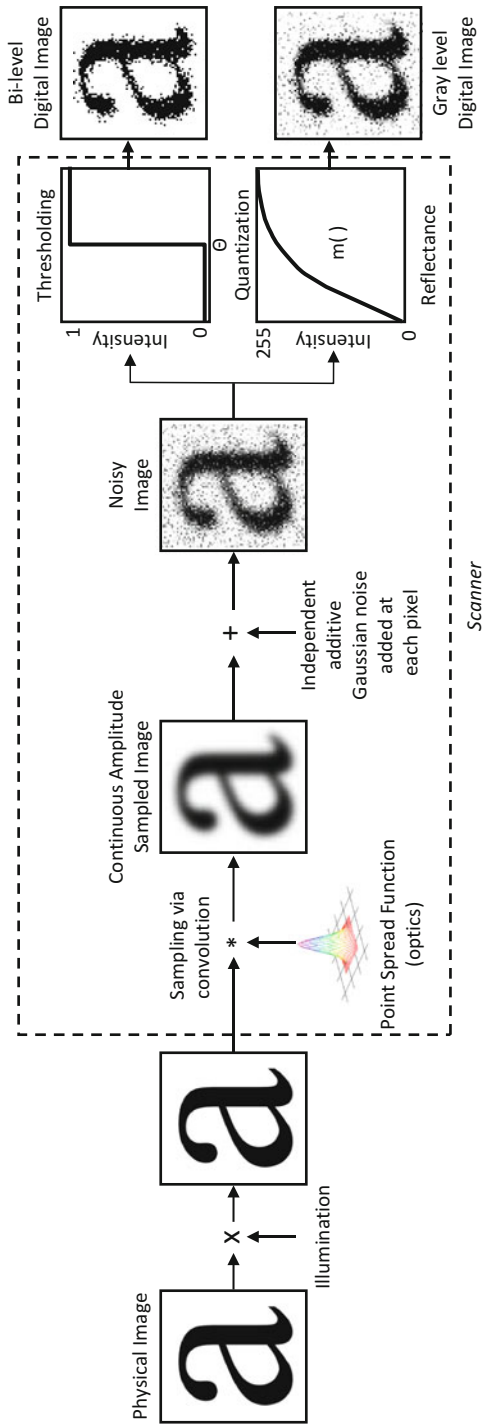


Fig. 2.11 Model of the scanner acquisition process



a

used  $\tau_1$  value to the  $\tau_1$  value  
 displacement curve will yield

$$(4.22)$$

$$(4.23)$$

to determine  $\Theta$ :

$$(4.24)$$

and  $\Theta$ , for which the theoret-  
 n  $f_{r_{max}}(\tau)$ , data. Best fit is

$$)))^2 \quad (4.25)$$

4.2.1. Then the parametric  
 regression. As described in  
 a form for which a unique  
 are 4.9. Within the shaded  
 describe  $f(\tau)$ . The equations  
 d star, there is a single scan-  
 be constant. In the test tar-  
 4.9, this corresponds to a

pend on the value of  $\Theta$  and  
 functional forms depends on  
 h horizontal slice, the MSE

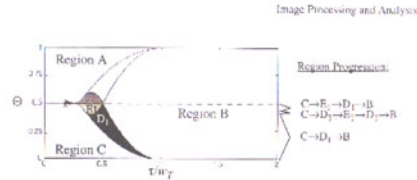


Figure 4.9: Fraction regions for triangular PSF.

the functional form of region  $E_1$ , then to the functional form of region  $D_1$  and finally to the functional form of region B. For  $\Theta$  values between these two ranges, the progression would be  $1 \rightarrow D_1 \rightarrow E_1 \rightarrow D_1 \rightarrow B$ . A similar set of progressions occur for  $0.5 < \Theta < 1.0$ .

To fit the parametric fraction to the data, assumptions about which of these six categories of progressions is valid must be made. Then the location of the boundaries between regions with respect to the data must be hypothesized. Each data region will be tested separately by regression to find the best parameters. The partition location whose estimate of  $(w_T, \Theta)$  gives the minimum MSE is considered correct and the corresponding  $(w_T, \Theta)$  estimates are designated to be the correct parameter values.

The regions F, G, etc. (regions where  $0 < f(\tau) < 1$  and  $\tau/w_T < 1/4$ ) have been eliminated in the implementation because they are very small and the effect of the star being in two dimensions is likely to obscure the data in these regions any ways. If necessary, they could be added to the implementation.

4.2.4 Implementation of Estimation from Wedges

The final estimation method uses isolated wedges. The location of the two straight edges and the apex of the blurred wedge must be found from the image and together they determine the PSF width and threshold value.

The pixels denoting the top and bottom edges of the wedge are identified by applying a MATLAB contour function to the image and these coordinates are fit to straight

b

used  $\tau_1$  value to the  $\tau_1$  value  
 displacement curve will yield

$$(4.22)$$

$$(4.23)$$

to determine  $\Theta$ :

$$(4.24)$$

and  $\Theta$ , for which the theoret-  
 n  $f_{r_{max}}(\tau)$ , data. Best fit is

$$)))^2 \quad (4.25)$$

4.2.1. Then the parametric  
 regression. As described in  
 a form for which a unique  
 are 4.9. Within the shaded  
 describe  $f(\tau)$ . The equations  
 d star, there is a single scan-  
 be constant. In the test tar-  
 4.9, this corresponds to a

pend on the value of  $\Theta$  and  
 functional forms depends on  
 h horizontal slice, the MSE

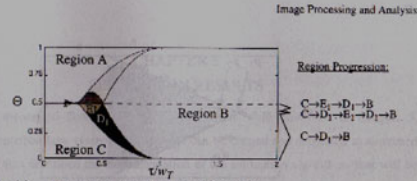


Figure 4.9: Fraction regions for triangular PSF.

the functional form of region  $E_1$ , then to the functional form of region  $D_1$  and finally to the functional form of region B. For  $\Theta$  values between these two ranges, the progression would be  $1 \rightarrow D_1 \rightarrow E_1 \rightarrow D_1 \rightarrow B$ . A similar set of progressions occur for  $0.5 < \Theta < 1.0$ .

To fit the parametric fraction to the data, assumptions about which of these six categories of progressions is valid must be made. Then the location of the boundaries between regions with respect to the data must be hypothesized. Each data region will be tested separately by regression to find the best parameters. The partition location whose estimate of  $(w_T, \Theta)$  gives the minimum MSE is considered correct and the corresponding  $(w_T, \Theta)$  estimates are designated to be the correct parameter values.

The regions F, G, etc. (regions where  $0 < f(\tau) < 1$  and  $\tau/w_T < 1/4$ ) have been eliminated in the implementation because they are very small and the effect of the star being in two dimensions is likely to obscure the data in these regions any ways. If necessary, they could be added to the implementation.

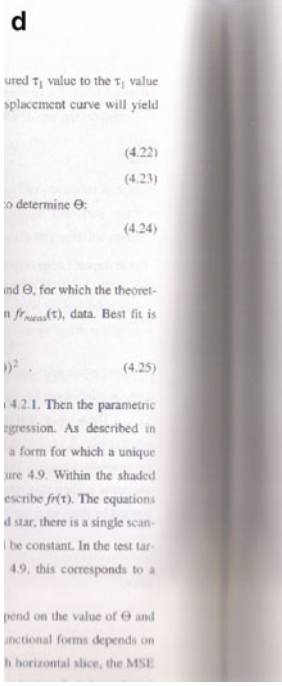
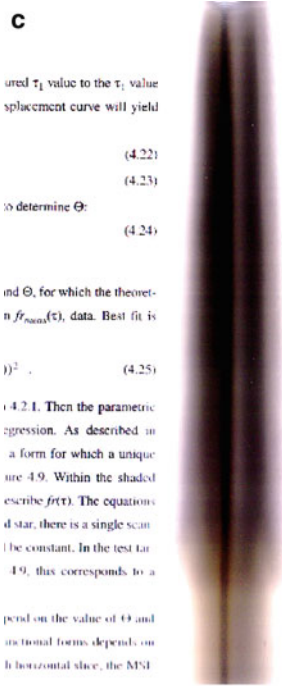
4.2.4 Implementation of Estimation from Wedges

The final estimation method uses isolated wedges. The location of the two straight edges and the apex of the blurred wedge must be found from the image and together they determine the PSF width and threshold value.

The pixels denoting the top and bottom edges of the wedge are identified by applying a MATLAB contour function to the image and these coordinates are fit to straight

Fig. 2.12 (continued)





**Fig. 2.12** Page scans showing (a) high brightness (shows saturation), (b) low brightness, (c) high contrast, and (d) low contrast

**C**

used  $\tau_1$  value to the  $\tau_1$  value  
 displacement curve will yield

$$(4.22)$$

$$(4.23)$$

to determine  $\Theta$ :

$$(4.24)$$

and  $\Theta$ , for which the theoret-  
 n  $f_{r_{max}}(\tau)$ , data. Best fit is

$$(4.25)$$

4.2.1. Then the parametric  
 regression. As described in  
 a form for which a unique  
 are 4.9. Within the shaded  
 describe  $f(\tau)$ . The equations  
 d star, there is a single scan-  
 be constant. In the test tar-  
 4.9, this corresponds to a

pend on the value of  $\Theta$  and  
 functional forms depends on  
 h horizontal slice, the MSE

**d**

used  $\tau_1$  value to the  $\tau_1$  value  
 displacement curve will yield

$$(4.22)$$

$$(4.23)$$

to determine  $\Theta$ :

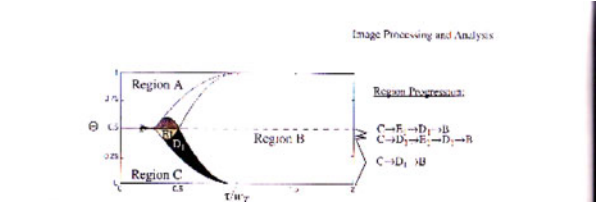
$$(4.24)$$

and  $\Theta$ , for which the theoret-  
 n  $f_{r_{max}}(\tau)$ , data. Best fit is

$$(4.25)$$

4.2.1. Then the parametric  
 regression. As described in  
 a form for which a unique  
 are 4.9. Within the shaded  
 describe  $f(\tau)$ . The equations  
 d star, there is a single scan-  
 be constant. In the test tar-  
 4.9, this corresponds to a

pend on the value of  $\Theta$  and  
 functional forms depends on  
 h horizontal slice, the MSE



**Figure 4.9: Fraction regions for triangular PSF.**  
 the functional form of region  $E_1$ , then to the functional form of region  $D_1$  and finally to the functional form of region B. For  $\Theta$  values between these two ranges, the progression would be  $1 \rightarrow D_1 \rightarrow E_1 \rightarrow D_1 \rightarrow B$ . A similar set of progressions occur for  $0.5 < \Theta < 1.0$ .

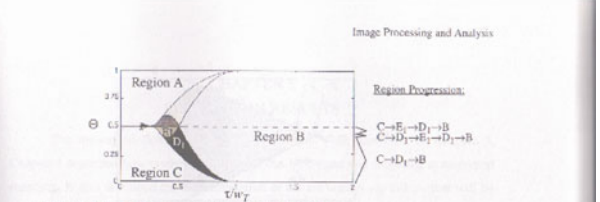
To fit the parametric fraction to the data, assumptions about which of these six categories of progressions is valid must be made. Then the location of the boundaries between regions with respect to the data must be hypothesized. Each data region will be tested separately by regression to find the best parameters. The partition location whose estimate of  $(w_T, \Theta)$  gives the minimum MSE is considered correct and the corresponding  $(w_T, \Theta)$  estimates are designated to be the correct parameter values.

The regions F, G, etc. (regions where  $0 < f(\tau) < 1$  and  $\tau/w_T < 1/2$ ) have been eliminated in the implementation because they are very small and the effect of the star being in two dimensions is likely to obscure the data in these regions any ways. If necessary, they could be added to the implementation.

**4.2.4 Implementation of Estimation from Wedges**

The final estimation method uses isolated wedges. The location of the two straight edges and the apex of the blurred wedge must be found from the image and together they determine the PSF width and threshold value.

The pixels denoting the top and bottom edges of the wedge are identified by applying a MATLAB contour function to the image and these coordinates are fit to straight



**Figure 4.9: Fraction regions for triangular PSF.**  
 the functional form of region  $E_1$ , then to the functional form of region  $D_1$  and finally to the functional form of region B. For  $\Theta$  values between these two ranges, the progression would be  $1 \rightarrow D_1 \rightarrow E_1 \rightarrow D_1 \rightarrow B$ . A similar set of progressions occur for  $0.5 < \Theta < 1.0$ .

To fit the parametric fraction to the data, assumptions about which of these six categories of progressions is valid must be made. Then the location of the boundaries between regions with respect to the data must be hypothesized. Each data region will be tested separately by regression to find the best parameters. The partition location whose estimate of  $(w_T, \Theta)$  gives the minimum MSE is considered correct and the corresponding  $(w_T, \Theta)$  estimates are designated to be the correct parameter values.

The regions F, G, etc. (regions where  $0 < f(\tau) < 1$  and  $\tau/w_T < 1/2$ ) have been eliminated in the implementation because they are very small and the effect of the star being in two dimensions is likely to obscure the data in these regions any ways. If necessary, they could be added to the implementation.

**4.2.4 Implementation of Estimation from Wedges**

The final estimation method uses isolated wedges. The location of the two straight edges and the apex of the blurred wedge must be found from the image and together they determine the PSF width and threshold value.

The pixels denoting the top and bottom edges of the wedge are identified by applying a MATLAB contour function to the image and these coordinates are fit to straight

If a bi-level image is to be produced, only the brightness setting is variable. This controls the *threshold level*,  $\Theta$ , used to convert the analog reflectance to a bi-level value

$$\text{bilevel}[i, j] = \begin{cases} 1 & \text{analog}[i, j] > \Theta \\ 0 & \text{analog}[i, j] < \Theta. \end{cases}$$

Values at every pixel will be converted by the same, global, threshold. Images acquired in gray scale usually are converted to bi-level through binarization before recognition algorithms are applied. These can be global thresholding algorithms or one of a variety of adaptive thresholding algorithms that target cases where the contrast between text and background is not uniform across the page or has significant noise. These are described further in (►[Chap. 4](#) (Imaging Techniques in Document Analysis Processes)).

There is a tradeoff between spatial sampling rate and gray depth. Gray scale acquisition can be completed at a lower resolution and still produce the same optical character recognition (OCR) accuracy as bi-level acquisition.

*Facsimile*, or fax, machines incorporate a type of flatbed scanner, but scan at much lower resolutions. The resolution is most commonly  $100 \times 100$  or  $100 \times 200$  dpi, Fig. 2.9e. The low resolution was chosen from earlier transmission bandwidth constraints. The fax machine sends the acquired image to a remote location for printing, rather than saving it for local processing.

Color scanners employ the same acquisition process, but divide the reflected light signal into three channels. Each channel is passed through a color filter. A process equivalent to gray-level acquisition is used on each color channel.

## Cameras and Mobile Devices

With camera-based acquisition, traditional paper documents are no longer the only content that can be acquired. Billboards, business signs, posters, and meeting room white boards are among the new material that is acquired and analyzed (see ►[Chap. 25](#) (Text Localization and Recognition in Images and Video)). Camera-based acquisition also removes the restriction that the object being captured must be a planar object. Text on bottles or round sign pillars can also be acquired. Cameras usually acquire the image in color as they are not designed specifically for bi-level document image acquisition.

Cameras and mobile devices operate on principles similar to those of flatbed scanners, but the position of the sensor relative to the object is now variable and so is the illumination. This increases the range of image distortions. In camera-based acquisition, the light is often a near field point source so the illumination pattern can be nonuniform. It is common for it to have a maximum intensity near the center decreasing with distance from the center, usually with radial symmetry. This often results in the corners of the image being noticeably darker than the center, sometimes making paper at the corners appear darker than the text at the center (Fig. 2.13a). Glare from flash, or from ambient light, can oversaturate parts of the image, obscuring parts of the image.

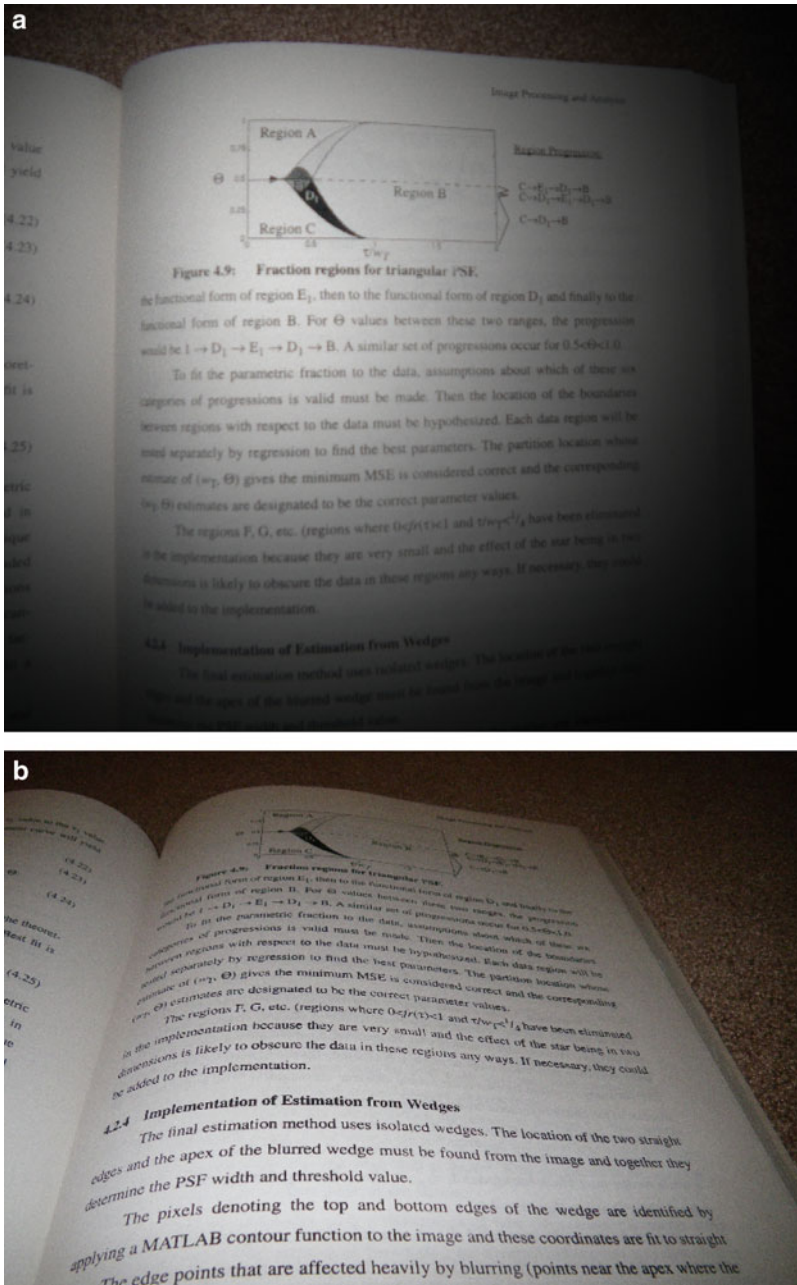


Fig. 2.13 Camera-based acquisition. (a) Nonuniform lighting and glare (b) perspective distortion

Cameras and mobile devices often have lower effective resolution than flatbed scanners. Both the camera optics (lenses) and sensor sensitivity contribute to this. Newer model cameras are able to acquire document images (A4 and US letter) at close to 300 dpi resolution. Some specialized cameras and document platforms have been developed to acquire book images. The Xerox Book Scanner was one of the early ones. The Google Million Book project utilizes camera-based imaging.

If the camera is not placed directly parallel to the image plane, perspective image distortions will result (Fig. 2.13b). These can be corrected for in post processing if the angle can be correctly determined (see ►Chap.4 (Imaging Techniques in Document Analysis Processes)). Variable positioning of the camera can result in part of the image being out of focus, requiring more correction than just perspective or the introduction of motion blur. These defects can be corrected to some degree by methods described in ►Chap.4 (Imaging Techniques in Document Analysis Processes).

## Video

Video in its most basic form is a temporal sequence of camera-based images. The object being captured can now exhibit motion relative to the sensor. This can cause motion blurring degradations and the need for object tracking. Text designed to provide information to the viewer that is computer generated and not acquired through the video device can be superimposed on the image such as with closed captioning.

For digital processing, analog video must be converted to a digital sequence. This involves spatial resampling and intensity quantization. In an analog recording, the vertical resolution is quantized, but along the horizontal axis, the signal is traced and an analog recording of the observed reflectance is acquired. Video is usually a color signal, and analog video is stored to be compatible with analog televisions. The color is therefore stored as an intensity channel and two color channels, instead of RGB as the image is acquired or as digital color images are often stored. The color channels are usually stored at a lower quality level to take advantage of compression possible based on the human visual system. As much of the world is converting to digital television, analog to digital video conversion will be less of an issue in the future, although digital video compression will still convert the colorspace and introduce artifacts.

Other image processing steps are usually needed when recognizing text in a video sequence. Even if the document is stationary, in video the camera may be moving so there is a possibility that the image is blurred. The document will appear in multiple image frames, so tracking of the document from frame to frame is needed. How these and other artifacts are handled is discussed in more detail in ►Chap. 25 (Text Localization and Recognition in Images and Video).

## Other Specialty Modes

While the majority of acquisition is done in the visible light band, there are applications that need other acquisition modes, such as historical document OCR and forensic OCR. For historical documents, the ink may have faded, or the paper may have yellowed or been damaged by watermarks or mold. These all reduce the contrast between the text and the background, sometimes to the point where they cannot always be distinguished in the visible spectrum. In the special historical document category of palimpsests, where the ink was deliberately removed to reuse the parchment or vellum to create a new book, trace particles from the original layer of ink may still be present and of interest to researchers, but not in quantities that will allow adequate detection in the visible band. Likewise, when trying to determine origins of questioned documents, details about the document beyond those observable in the visible band may be necessary. Differentiation between similar but different inks is often sought to identify forgeries. Identification of the type of ink or paper can provide a date of origin and determine authenticity of the document as a whole or certain parts of the text. The inks may appear similar in the visible band, but due to different chemical composition or decay rates, other imaging techniques may show a distinction.

Filters can be applied to cameras to accentuate color or spectral differences. Likewise, light at different wavelengths can be used to illuminate the document. Infrared or ultraviolet light sources are often used to show different inks. The same acquisition procedures discussed for visible band scanner or camera acquisition apply to these methods as well, but the light will be in a different frequency band. In addition, if there is a significant enough difference between the reflected light for text and substrate regions at those wavelengths, the text can become visible or show characteristically different properties, allowing different ink samples to be distinguished. Microfilm and microfiche systems need a projection system to project visible light through the film and receive it at a sensor.

Sometimes, the light changes frequency when interacting with the ink or paper, and this change in frequency is measured at the sensor, such as in Raman spectroscopy. Another method uses the incident light to cause the materials in the document to fluoresce. Ultraviolet illumination can cause organic material, such as in parchment, to fluoresce, emitting longer wavelength light that may be in the visible band. The ink will block some of this fluorescence, making it visible, although the document is not reflecting visible light. Several acquisition modes can be used in conjunction if the images are properly registered after imaging, and their combination can highlight document features.

Documents are usually thought of as being 2-D planar or nearly planar surfaces, and 2-D imaging techniques are most often used. Occasionally, the document is rolled or bound and has become fragile so transforming the document to a physically planar object is not possible. In these cases, 3-D imaging techniques, such as MRI and PET as used in medical imaging, can be called upon to provide images of

the document. The 2-D paper surface will be a planar curve in the interior of the acquired volume “cube.” The 2-D surface can be extracted and “flattened” digitally.

---

## Document Quality

The factors affecting quality are many, and ways to specify or quantify them exist. Some defects have been modeled. Models help understanding and can improve algorithms that work with images containing defects. When the image quality is low, it can affect analysis and recognition results.

It is always desirable to have high document quality for both reading and analysis. However, humans and computers do not always define image quality the same way. For humans, the ability to comfortably read the content is the deciding goal, but more than the individual characters affect this ability. Three levels of document quality exist:

- *Decipherability* – the ability to identify and read correctly a certain letter form
- *Legibility* – the ability to identify and read correctly a certain word form
- *Readability* – the ability to identify and read a certain meaningful chain of words correctly

Image quality can also be determined by the computer’s ability to correctly recognize characters or words, as indicated in character and word error rates. Physiological comfort and familiarity are only human factor issues, not a machine readability issue, but both computers and humans look for contrast, uniform font, and variations between character classes.

Unfortunately, quality can decrease based on several factors: the quality of the materials used in document production, the method used to produce the document, the content, and the acquisition method. Knowledge of the document content and lexicographic clues influence these items. This is also influenced by typography: the choice of typeface, size, boldness, length of line, and margins (size and justified/even or uneven) or paper gloss or glare. For a computer, these clues can be included in its recognition method.

## Factors Affecting Document Quality

There are many sources of document image degradations. They can be categorized into three areas:

1. defects in the paper (yellowing, wrinkles, coffee stains, speckle)
2. defects introduced during printing (toner dropout, bleeding and scatter, baseline variations, show-through)
3. defects introduced during digitization through scanning (skew, mis-thresholding, resolution reduction, blur, sensor sensitivity noise, compression artifacts)

To provide an objective and quantifiable measure of the quality of newly printed documents, an international standard ISO 13660 [2] was developed. The standard provides a set of device-independent image quality attributes, along with



**Table 2.3** Summary of attributes used in print quality standard ISO 13660

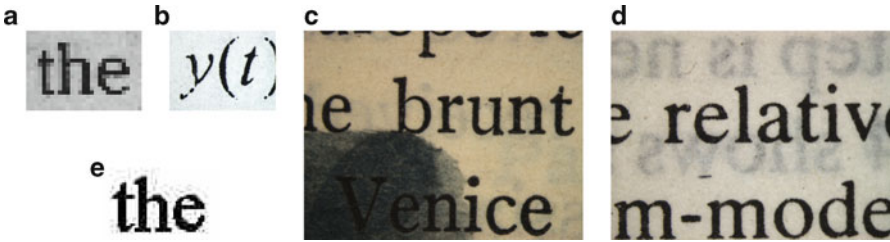
Large area density attributes	
Darkness	How dark the large inked areas are in terms of optical density
Background haze	Whether colorant/ink is visible in the background field but not as specific individual marks
Graininess	Standard deviation of aperiodic fluctuations of optical density at spatial frequencies greater than 0.4 cycles/mm
Mottle	Aperiodic fluctuations of optical density in printed zones at spatial frequencies less than 0.4 cycles/mm
Extraneous marks	(Count of visible unintended regions of colorant in the background area and voids (holes or gaps) within a solid image area)
Character and line attributes	
Blurriness	Rate of transition from black to white
Raggedness	The amount the path of an edge boundary varies locally
Line width	Average stroke width
Darkness	Optical density
Contrast	The relationship between the darkness (reflectance) of a line segment or character image and the background
Fill	Appearance of homogeneity of darkness within the boundary of a line segment or character image
Extraneous marks	Unintended regions of colorant in the background area near a character or line segment
Background haze	Colorant near a character that is visible but not resolvable as distinct marks at standard viewing distance to the unaided eye

measurement methods, and analytical procedures to describe the quality of high-contrast documents across a range of printers including impact printers, nonimpact printers, and photocopiers. The standard includes attributes to measure features over large areas such as the background or large solid-filled areas and to measure text and line quality attributes, Table 2.3. All of these factors intended for the printer development community also affect document image analysis.

The attributes in ISO 13660 target newly printed documents. In document image analysis, document quality is also affected by photocopying, digitization, binarization, compression artifacts, aging, fading, yellowing, bleedthrough, and show-through.

When digitizing an image, spatial resolution is lost, and this may obscure some image features (Fig. 2.14a). During printing or copying, ink may drop out (Fig. 2.14b). Photocopying in addition to digitization involves reprinting, a process that may cause additional loss of features, Fig. 2.9. If the image is high contrast and on a uniform background, binarization is not an overly difficult problem, but if the document has aged, or was written in a lighter ink, there may be fading, yellowing, and staining which complicate the binarization problem (Fig. 2.14c).

Thin or backlit paper, or documents written with a very penetrable ink, can face problems where the text on the verso side becomes visible on the recto side as bleedthrough or show-through, again decreasing the document quality (Fig. 2.14d).



**Fig. 2.14** Examples of degradations affecting document image quality. (a) Low-resolution scanning, (b) ink dropout, (c) marks, (d) show-through, and (e) JPEG image compression artifacts

Images can require a significant amount of digital memory to store. To compensate for this, various techniques have been developed that can reduce the memory requirements by taking advantage of redundancy in image content. If the compression technique allows reconstruction of the original image with no change in any of the pixel values, then it is called a lossless compression scheme, such as used in PNG, GIF, and TIF. However, some compression schemes, like JPEG, make small changes to the image that are designed to be minimally observable by humans. Some changes alter individual pixel values, and when color images are saved in JPEG format, the hue and chrominance in  $2 \times 2$  squares of pixels are averaged and one value instead of four is stored for each. These minimal changes are more noticeable in high-contrast images than in natural scene images and can cause greater effects on automated processing (Fig. 2.14e).

## Effects of Document Quality on Analysis and Recognition Results

The goal in optical character recognition is to accurately recognize all the characters in the document image. Character recognition and document analysis algorithms will perform best if the image is cleanly printed on quality paper with dark, non-spreading ink. Documents with these attributes make up only a small portion of the documents being processed. Research has been done to automatically determine if the image quality is “good.” One of the most common metrics is OCR error rate. Early definitions of document image quality determined whether a page was likely to have few enough OCR errors to make it cheaper to run OCR applications and correct, versus to hand type a document in its entirety. If the image has low quality, then the OCR accuracy is likely to be lower. Low OCR accuracy leads to the need to apply more algorithms to improve the image or post processing algorithms to fix or compensate for the errors, and if the quality is too low, then manually entering the document content may be more efficient.

Some common OCR errors are the letters  $r$  and  $n$  being merged and interpreted as an  $m$ , the letter  $m$  being broken and interpreted as an  $r$  followed by an  $n$ , or an  $e$  losing its horizontal bar or having its loop filled in so it is recognized as a  $c$  (Fig. 2.15). Rice et al. [3] looked at large-scale studies of common OCR errors and summarized the sources of these errors, providing both samples of





**Fig. 2.15** (a) Original characters and (b) characters after degradation. Through touching and broken artifacts, recognition algorithms can easily misclassify the characters

**Table 2.4** Quality measures used to quantify document image degradations

Quality measure	Description	Method	Sources
Small speckle factor (SSF)	The amount of black background speckle in the image	Number of the black connected components in an image that contain a range of pixel counts	Inkjet, laser print, sensor noise, JPEG artifacts
White speckle factor (WSF)	The degree to which fattened character strokes have shrunken existing holes in or gaps between characters causing several small white islands to form	The ratio of the number of white connected components less than $3 \times 3$ pixels in size relative to the number of white connected components in the total image	Ink dropout
Touching character factor (TCF)	The degree to which neighboring characters touch	Number of long and low connected components	Ink bleeding
Broken character factor (BCF)	If characters are broken, there will be many connected components that are thinner than if the characters were not broken	The number of thin connected components	Ink dropout

the degraded text and the common OCR results. One category of errors was due to imaging defect, which included heavy print, light print, stray marks, and curved baselines. Some image features have been identified that are correlated with OCR error rates. Documents with low recognition rates often have touching characters from stroke thickening and broken characters from stroke thinning. These degradations lead to certain characteristics that can be arbitrarily measured from a document page. Quality measures (Table 2.4) look at quantifiable features common in characters [4] and [5].

Document images are subjected to image processing algorithms (►Chap. 4 (Imaging Techniques in Document Analysis Processes)) to improve the quality of the image before being passed to the recognition stage. The effectiveness of these algorithms is often evaluated based on the relative OCR accuracy achievable with different techniques. While it is the image processing algorithm being adjusted, and thus affecting the recognition results, compatibility of the image processing and the recognition algorithm also determines the net results.

## Models of Document Degradations

Document creation was described in the section “Writing and Printing Processes” of this chapter. Documents can be created by handwriting, hand printing, or by

**Table 2.5** Summary of degradation models

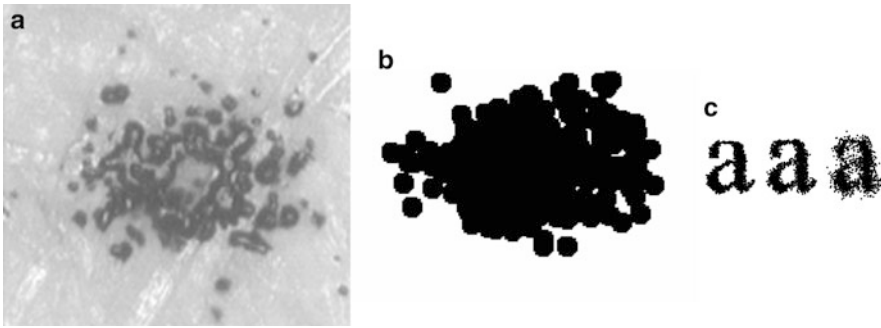
Model	Type of Process Modeled
<b>Franke and Rose</b>	Handwriting ink deposition
<b>Norris and Barney Smith</b>	Electrophotographic printing
<b>Baird</b>	Machine print printing and acquisition
<b>Obafemi-Ajayi and Agam</b>	Typewriter printing
<b>Kanungo</b>	General machine print degradation
<b>Show-through</b>	Appearance of ink from recto on verso
<b>Book binding</b>	Curvature of page

machine printing. As has been described, document images can contain various degradations, and they almost always decrease the image processing or recognition algorithm performance. Many of these degradations have been modeled mathematically (Table 2.5). The models aid development and analysis of image processing and recognition algorithms. At times they also allow researchers to generate a greater supply of samples for algorithm evaluation. Some are physics based, some aim to mimic a specific degradation, and others aim to produce degradations in general, not specifically related to any physical cause such as a problem with a particular paper or ink. Some models describe a specific document creation process, like pen handwriting, laser printing, or typewriters. Others describe the image acquisition process.

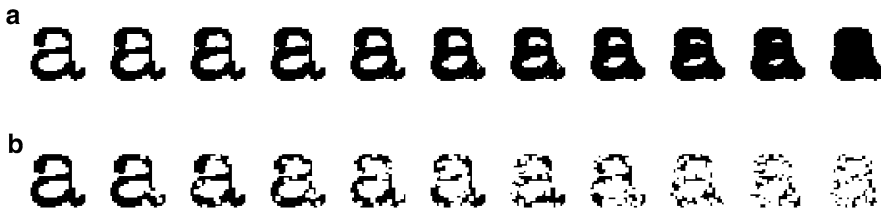
Physical movements of the pen relative to the paper affect the ink deposited on the paper in hand-printed or handwritten documents determining a relationship between the pen angle and pen force and the resulting ink intensity. The type of pen, its corresponding ink, and the type of paper influence the resulting ink intensity for a given pen force. The ink deposition model by Franke and Rose [6] relates the intensity to the applied force for solid, viscous, and liquid (or fluid) inks. This is a useful model for both handwriting recognition (►Chap. 11 (Handprinted Character and Word Recognition)) and document forensic analysis.

Nonimpact electrophotographic printing (i.e., laser printing) has been modeled by Norris and Barney Smith [7]. This provides ways to generate and analyze printed samples including the degradations that occur from toner spatter (Fig. 2.16). When the photoconductor drum is discharged by a laser with a Gaussian intensity profile, the charge change on the photoconductor will be proportional to the light intensity. Toner will stick to the photoconductor proportional to the amount of charge. That toner is then transferred to the paper and fused. The model depends on the radius of the toner pieces and the toner spatial density. The amount of paper actually covered by toner can also be modeled. Higher toner density will appear darker to the eye, scanner, or camera. The toner density and the coverage can be spatially varying functions and can provide a description for a whole page.

Typewritten characters have characteristic degradations of uneven ink intensity due to uneven typewriter key pressure, faded ink, or filled-in characters due to gummed type (Fig. 2.5). Degradations seen in characters produced through impact printing from a typewriter were modeled by Obafemi-Ajayi and Agam [8], Fig. 2.17.



**Fig. 2.16** Examples of (a) real and (b) simulated print samples. (c) Sample characters made with the Norris printer model



**Fig. 2.17** Examples of characters produced by the Obafemi-Ajayi model. Filled and broken samples degraded at multiple levels

Broken characters are created by randomly selecting windows of varying size and flipping the foreground pixels from black to white. Filled or thickened characters can be created from repeated morphological closing of the original character with a large kernel until the image is entirely closed. The closed image is subtracted from the original image to get regions defined as holes. The holes are eroded iteratively with a small kernel until it is completely eroded. The filled characters are created from turning background pixels at the boundary of the holes and the character from white to black. The level of broken degradation is measured by the percentage of foreground pixels in the degraded broken image relative to the number of foreground pixels in the original image. The filled degradation level is the percentage of foreground pixels added relative to the total number of pixels that can be filled.

Baird [9] proposed a parameterized model that looks at the whole document creation process, including both typesetting and scanning. Some degradations are global for the page, and some are local to the character or even the pixel. This model has a comprehensive set of deformation variables. Parameters that relate to the document formatting or typesetting include size, skew, xscale, yscale, xoffset, and yoffset (Table 2.6). To consider the defects introduced in acquired images that are related to the scanning process, as discussed in the section “[Acquisition Methods](#)” (Fig. 2.11), the parameters resolution, blur, threshold, sensitivity, and

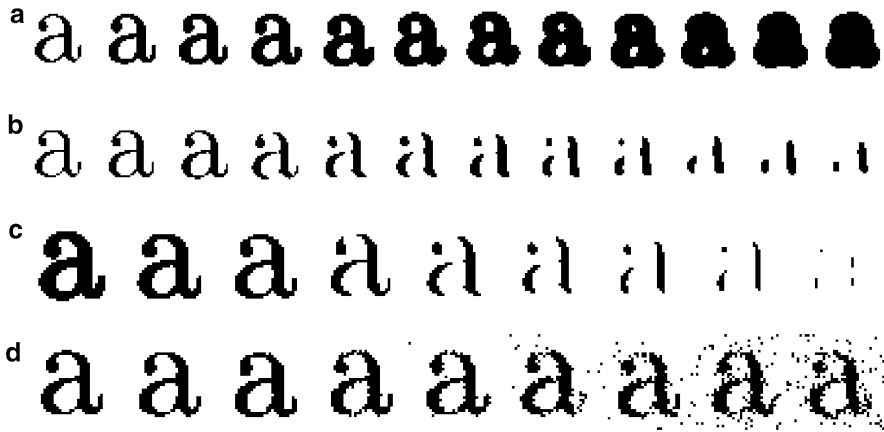
**Table 2.6** Summary of Baird degradation model parameters

	Parameter	Description
<b>Creation</b>	Size	Text size in points
	Skew	The amount of rotation from horizontal of each character
	xscale and yscale	Multiplicative scaling factors to adjust the size of the character from its nominal value
	xoffset and yoffset	Translations of the individual characters with respect to the baseline and the sampling grid. These could also be considered sampling phases in acquisition
<b>Acquisition</b>	Resolution	A combination of the character size in points and the scanning resolution
	Blur	The size of the PSF, which Baird models as the standard deviation of a Gaussian filter
	Threshold	The intensity level which determines whether the pixel will be black or white
	Sensitivity	Noise amount added to the level at each sensor before thresholding
	Jitter	A random offset of the pixel centers from a square grid

jitter are included. Baird included a likely range and distribution for each parameter associated with each degradation in the model which he used to generate characters for OCR experiments. Blur, threshold, and sensitivity control the appearance of the individual character (Fig. 2.18). These parameters were determined by Ho and Baird [10] to be the most related to the accuracy of an OCR system. One approach to getting the “best quality” image acquisition for best OCR accuracy could be to configure the acquisition system to a PSF width and binarization threshold that yield lower OCR error rates.

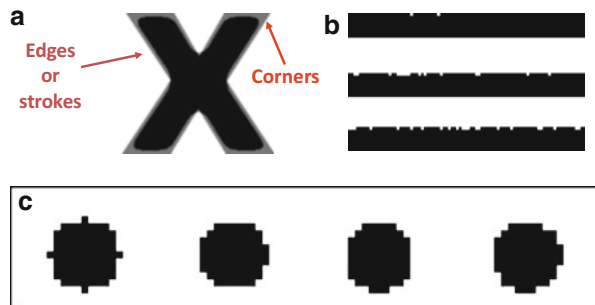
Based on the physical degradation model of blurring with a PSF of width  $w$ , and thresholding at a global threshold of  $\Theta$ , plus additive Gaussian noise with a standard deviation  $\sigma_{\text{noise}}$ , models to describe and quantify three resulting degradations have been developed by Barney Smith et al. [11, 12]. Edges between black and white such as character borders will relocate on the page as a function of the width of the PSF and the binarization threshold (Fig. 2.19a). If the edge spread is positive, strokes will be thicker, and when it is negative strokes will become thinner. If it is too negative, strokes will vanish, and if it is too positive, holes can fill and neighboring characters can touch, Fig. 2.15. Corners will be rounded by the scanning and binarization process. Because of the corner erosion and the possibilities of strokes or spaces totally vanishing, inverting the degradation to return the characters to their original states is not possible.

The noise that is present in gray-level images can be characterized by the standard deviation of the additive noise. If the image is thresholded after noise is added, the standard deviation of the noise does not do a good job of characterizing its effect on the binarized image (Fig. 2.19b). The metric *noise spread* describes the noise present in a quantitative measure that also qualitatively describes the noise effect. Noise spread takes into account the blur and threshold.



**Fig. 2.18** Examples of characters produced by the Baird degradation model. (a) Varying PSF width with low then (b) medium threshold, (c) varying threshold, and (d) varying additive noise (sensitivity)

**Fig. 2.19** (a) Common degradations from scanning without noise include edge spread and corner erosion; (b) edge noise is also common. It can be measured better through noise spread ( $NS$ ) which is different for these samples than through noise variance  $\sigma_{noise}$  which is the same for these samples. (c) One image sampled at four different phases



Photocopying and facsimile involve steps of image digitization and image printing, both steps introduce additional distortions, Fig. 2.9. Some of the printing noise can be accounted for in the additive noise in the scanner model, so photocopying can be modeled by repeated application of the scanner model.

Since the Nyquist criterion does not hold for document images, sampling during acquisition plays a role in determining the resulting digital image and what information it will contain. The location of the character image relative to the sampling grid can cause significant differences in the bi-level outcome of the digitization process (Fig. 2.19c). As a result, the edge pixel values do not vary independently of each other. Sarkar [13] developed a model of the random-phase sampling aspect of the physical scanning process. He showed the effect that the random-phase sampling has on the character images produced after digitization.



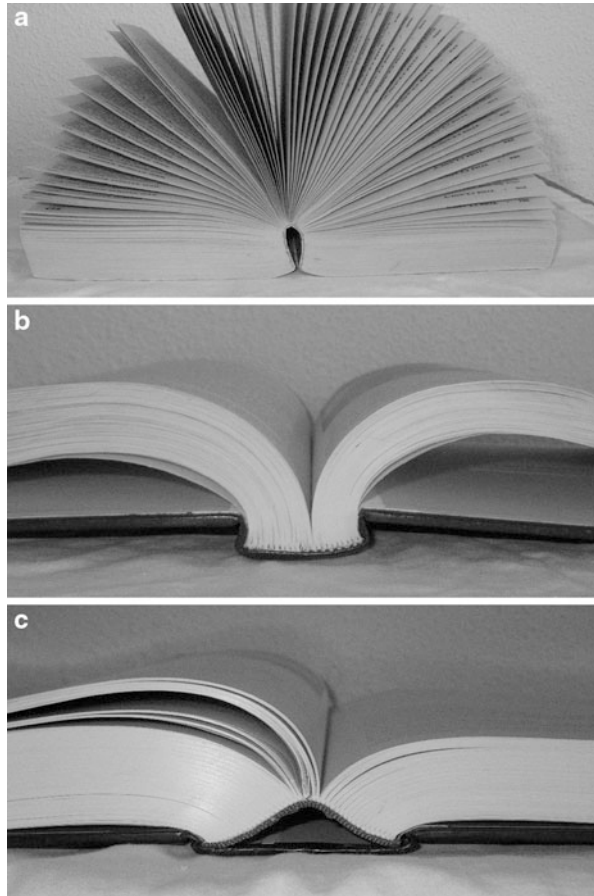
**Fig. 2.20** Examples of characters produced by the Kanungo degradation model. Effects of changing (a) foreground transition probability, (b) background transition probability, (c) base transition probability, and (d) diameter of morphological closing operator. Default parameters are  $[\eta, \alpha_0, \alpha, \beta_0, \beta, k] = [0, 1, 1.5, 1, 1.5, 5]$

The degradations seen in digitized characters can also be mimicked without the model having a connection to the physical creation or acquisition process. Kanungo et al. [14] developed a morphological model for local distortion. The degradation model has six parameters:  $\eta$ ,  $\alpha_0$ ,  $\alpha$ ,  $\beta_0$ ,  $\beta$ , and  $k$ . These parameters are used to degrade an ideal binary image. In this noise model, the distance of a pixel from the boundary of the template character determines its probability of changing from black to white or from white to black. The background pixels are flipped with probabilities determined by the parameters  $\beta$  and  $\beta_0$ , with probability decreasing as the pixels get further from the edge of the character. Likewise, the foreground pixels are flipped with probabilities determined by the parameters  $\alpha$  and  $\alpha_0$ . After bit flipping, the character is processed by a morphological closing operation with a disk structuring element of diameter  $k$  to fill in some of the holes in the character and to account for the correlation present in the scanning process. This model can be used to produce degraded characters that have a wide range of parameterized degradations. Samples of characters generated through this model with a range of parameters are shown in Fig. 2.20.

### Show-Through

If the paper used for printing is thin, or some of the light source comes from the back side of the paper, the printing on the verso side can become visible on the recto side (Fig. 2.14d). Show-through models are used with images of both the observed verso and recto to estimate the corresponding original images. Nominally, a scaled version of the ideal verso side is added to the intensity of the image on the recto side [15, 16]. Variations on this basic idea have been developed to consider the optical

**Fig. 2.21** Examples of (a) flexible spine, (b) fast spine, and (c) hollow spine bindings



blurring or the optical diffusion the paper caused to the verso image before it is visible on the recto. Some models operate on the reflectivity of the verso image and some on the optical density of the ideal recto and verso sides. A model that utilizes more detailed physics of the process was developed by Khan and Hasan [17]. This includes the effect of the light of the scanner penetrating the paper and bouncing off the usually white reflecting surface of the scanner top.

### Book Binding

Finally, there are models addressing how book binding affects document image analysis during acquisition. There are three main types of bindings for books: *flexible spine*, *fast spine*, and *hollow spine* (Fig. 2.21). When imaged on a flatbed scanner, the paper near the spine will not sit as close to the glass where the focal point of the imaging system lies and will not receive as much light. This causes shadows and out of focus problems as well as varying curvature of the text lines (Fig. 2.22). Kanungo et al. [14] modeled this defect based on the underlying perspective geometry of the optical system. He assumed the curved part was circular



a

ured  $\tau_1$  value to the  $\tau_1$  value  
 placement curve will yield

$$(4.22)$$

$$(4.23)$$

to determine  $\Theta$ :

$$(4.24)$$

nd  $\Theta$ , for which the theoret-  
 $f_{r_{meas}}(\tau)$ , data. Best fit is

$$j)^2 \quad (4.25)$$

4.2.1. Then the parametric  
 regression. As described in  
 a form for which a unique  
 are 4.9. Within the shaded  
 describe  $f_r(\tau)$ . The equations  
 d star, there is a single scan-  
 be constant. In the test tar-  
 4.9, this corresponds to a

pend on the value of  $\Theta$  and  
 nctional forms depends on  
 h horizontal slice, the MSE



Image Processing and Analysis

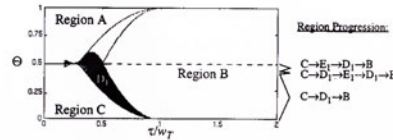


Figure 4.9: Fraction regions for triangular PSF.

the functional form of region  $E_1$ , then to the functional form of region  $D_1$  and finally to the functional form of region B. For  $\Theta$  values between these two ranges, the progression would be  $1 \rightarrow D_1 \rightarrow E_1 \rightarrow D_1 \rightarrow B$ . A similar set of progressions occur for  $0.5 < \Theta < 1.0$ .

To fit the parametric fraction to the data, assumptions about which of these six categories of progressions is valid must be made. Then the location of the boundaries between regions with respect to the data must be hypothesized. Each data region will be tested separately by regression to find the best parameters. The partition location whose estimate of  $(w_T, \Theta)$  gives the minimum MSE is considered correct and the corresponding  $(w_T, \Theta)$  estimates are designated to be the correct parameter values.

The regions F, G, etc. (regions where  $0 < f_r(\tau) < 1$  and  $\tau/w_T < 1/4$  have been eliminated in the implementation because they are very small and the effect of the star being in two dimensions is likely to obscure the data in these regions any ways. If necessary, they could be added to the implementation.

4.2.4 Implementation of Estimation from Wedges

The final estimation method uses isolated wedges. The location of the two straight edges and the apex of the blurred wedge must be found from the image and together they determine the PSF width and threshold value.

The pixels denoting the top and bottom edges of the wedge are identified by applying a MATLAB contour function to the image and these coordinates are fit to straight

b

ured  $\tau_1$  value to the  $\tau_1$  value  
 placement curve will yield

$$(4.22)$$

$$(4.23)$$

to determine  $\Theta$ :

$$(4.24)$$

nd  $\Theta$ , for which the theoret-  
 $f_{r_{meas}}(\tau)$ , data. Best fit is

$$j)^2 \quad (4.25)$$

4.2.1. Then the parametric  
 regression. As described in  
 a form for which a unique  
 are 4.9. Within the shaded  
 describe  $f_r(\tau)$ . The equations  
 d star, there is a single scan-  
 be constant. In the test tar-  
 4.9, this corresponds to a

pend on the value of  $\Theta$  and  
 nctional forms depends on  
 h horizontal slice, the MSE



Image Processing and Analysis

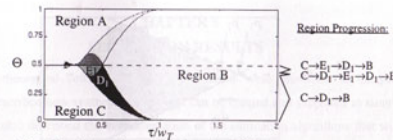


Figure 4.9: Fraction regions for triangular PSF.

the functional form of region  $E_1$ , then to the functional form of region  $D_1$  and finally to the functional form of region B. For  $\Theta$  values between these two ranges, the progression would be  $1 \rightarrow D_1 \rightarrow E_1 \rightarrow D_1 \rightarrow B$ . A similar set of progressions occur for  $0.5 < \Theta < 1.0$ .

To fit the parametric fraction to the data, assumptions about which of these six categories of progressions is valid must be made. Then the location of the boundaries between regions with respect to the data must be hypothesized. Each data region will be tested separately by regression to find the best parameters. The partition location whose estimate of  $(w_T, \Theta)$  gives the minimum MSE is considered correct and the corresponding  $(w_T, \Theta)$  estimates are designated to be the correct parameter values.

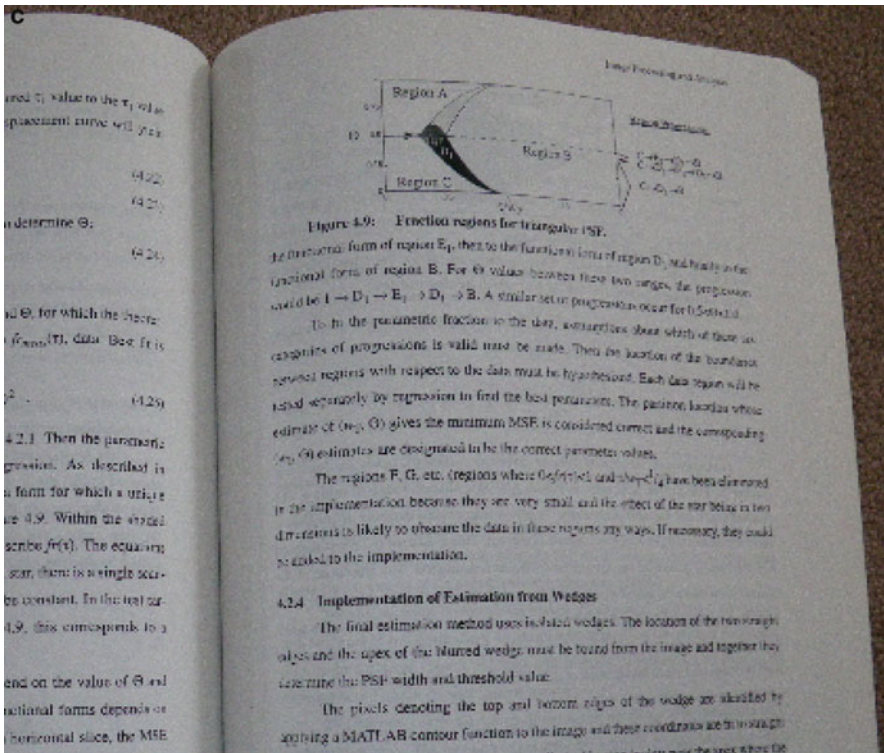
The regions F, G, etc. (regions where  $0 < f_r(\tau) < 1$  and  $\tau/w_T < 1/4$  have been eliminated in the implementation because they are very small and the effect of the star being in two dimensions is likely to obscure the data in these regions any ways. If necessary, they could be added to the implementation.

4.2.4 Implementation of Estimation from Wedges

The final estimation method uses isolated wedges. The location of the two straight edges and the apex of the blurred wedge must be found from the image and together they determine the PSF width and threshold value.

The pixels denoting the top and bottom edges of the wedge are identified by applying a MATLAB contour function to the image and these coordinates are fit to straight

Fig. 2.22 (continued)



**Fig. 2.22** Examples images of bound books taken with (a) a bi-level flatbed scanner, (b) a gray-level flatbed scanner, and (c) a camera

with a radius  $\rho$  and calculated the amount of bending of all parts of the image, which was followed by a perspective distortion, nonlinear illumination, and nonlinear blurring to account for out of focus. This can be used to correct the illumination and straighten out the lines of text.

## Conclusion

Over time, many methods and materials have been developed and used to create documents. Each offered an advantage based on current technology and available materials, and each produces a resulting document image that will have different characteristics. If the best combination of materials is used, good quality documents can be created. The recognition of the content of a document depends heavily on the processes of how a document was created. To begin document analysis, digital images of the physical document will be acquired. The acquisition process produces an image which can contain many image features from the document but can also introduce undesired artifacts. When those image characteristics and acquisition artifacts are undesired, they can decrease document analysis performance. Recognition

methods are developed based on what is expected from the document generation and acquisition stages. Models have been created to try and understand the degradations and to provide tools to increase performance of future document analysis algorithms.

---

## References

1. ISO/IEC 12233:2000, Photography – Electronic still-picture cameras – Resolution measurements (2000)
2. ISO/IEC 13660:2001 Information Technology – Office equipment – Measurement of image quality attributes for hardcopy output – Binary monochrome text and graphic images (2001)
3. Rice SV, Nagy G, Nartker T (1999) Optical character recognition: an illustrated guide to the frontier. Kluwer Academic, Boston
4. Cannon M, Hochberg J, Kelly P (1999) Quality assessment and restoration of typewritten document images. *Int J Doc Anal Recognit* 2:80–89
5. Souza A, Cheriet M, Naoi S, Suen CY (2003) Automatic filter selection using image quality assessment. In: Proceedings international conference on document analysis and recognition, Edinburgh, pp 508–511.
6. Franke K, Rose S (2004) Ink-deposition model: the relation of writing and ink deposition processes. In: Proceedings of the 9th international workshop on frontiers in handwriting recognition (IWFHR-9 2004), Kokubunji
7. Norris M, Barney Smith EH (2004) Printer modeling for document imaging. In: Proceedings of the international conference on imaging science, systems, and technology (CISST'04), Las Vegas, 21–24 June 2004, pp 14–20
8. Obafemi-Ajayi T, Agam G (2012) Character-based automated human perception quality assessment in document images. In: *IEEE Trans Syst Man Cybern A Syst Hum* 42(3):584–595
9. Baird HS (1990) Document image defect models. In: Proceedings of the IAPR workshop on syntactic and structural pattern recognition, Murray Hill, 13–15 June 1990
10. Ho TK, Baird HS (1997) Large-scale simulation studies in image pattern recognition. *IEEE Trans Pattern Anal Mach Intell* 19(10):1067–1079
11. Barney Smith EH (2005) Document scanning. McGraw-Hill 2005 yearbook of science and technology. McGraw-Hill, New York/London
12. Barney Smith EH (2009) A new metric describes the edge noise in bi-level images. *SPIE Newsroom*. doi:10.1117/2.1200910.1829
13. Sarkar P, Nagy G, Zhou J, Lopresti D (1998) Spatial sampling of printed patterns. *IEEE Trans Pattern Anal Mach Intell* 20(3):344–351
14. Kanungo T, Haralick RM, Phillips I (1994) Nonlinear local and global document degradation models. *Int. J Imaging Syst Technol* 5(4):220–30
15. Sharma G (2001) Show-through cancellation in scans of duplex printed documents. *IEEE Trans Image Process* 10(5):736–754
16. Tonazzini A, Salerno E, Bedini L (2007) Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. *IJDAR* 10(1):17–25
17. Khan MR, Hasan MK (2010) A novel model for show-through in scan of duplex printed documents. *Signal, Image Video Process*, 6(4):625–645

## Further Reading

Document Creation Materials: Paper, Pens, and Inks

- Nickell J (1990) Pen, ink & evidence: a study of writing and writing materials for the penman, collector and document detective. University Press of Kentucky
- Whalley JI (1975) Writing implements and accessories. Gale Research, Detroit

### Printing

- Kipphan H (2001) Handbook of print media. Springer, Heidelberg
- Williams EM (1993) The physics and technology of xerographic processes. Krieger
- Proudfoot WB (1972) The origin of stencil duplicating. Hutchinson, London
- Robert U (2000) A review of halftoning techniques. In: Color imaging: device-independent color, color hardcopy, and graphics arts V. Proceedings of SPIE, vol 3963, pp 378–391

### Imaging Methods

- Trussell HJ, Saber E, Vrhel M (2005) Color image processing. IEEE Signal Process Mag, Jan 2005, 14–22

### Displays

- Heikenfeld J (2010) Light, brite displays. IEEE Spectrum, Mar 2010, NA-28-33

### Image Quality

- ISO/IEC 12233:2000 (2000) Photography – electronic still-picture cameras – resolution measurements
- ISO/IEC 13660:2001 (2001) Information technology – office equipment – measurement of image quality attributes for hardcopy output – binary monochrome text and graphic images
- Ho TK, Baird HS (1997) Large-scale simulation studies in image pattern recognition. IEEE Trans Pattern Anal Mach Intell 19(10):1067–1079
- Rice SV, Nagy G, Nartker T (1999) Optical character recognition: an illustrated guide to the frontier. Kluwer Academic
- Cannon M, Hochberg J, Kelly P (1999) Quality assessment and restoration of typewritten document images. Int J Doc Anal Recognit 2:80–89
- Souza A, Cheriet M, Naoi S, Suen CY (2003) Automatic filter selection using image quality assessment. In: Proceedings of the international conference on document analysis and recognition, pp 508–511

### Models

- Franke K, Rose S (2004) Ink-deposition model: the relation of writing and ink deposition processes. In: Proceedings of the 9th international workshop on frontiers in handwriting recognition (IWFHR-9 2004)
- Norris M, Barney Smith EH (2004) Printer modeling for document imaging. In: Proceedings of the international conference on imaging science, systems, and technology: CISST'04, Las Vegas, 21–24 June 2004, pp 14–20
- Obafemi-Ajayi T, Agam G (2012) Character-based automated human perception quality assessment in document images. IEEE Trans Syst Man Cybern A-Syst Hum 42(3):584–595
- Baird HS (1990) Document image defect models. In: Proceedings of the IAPR workshop on syntactic and structural pattern recognition, Murray Hill, 13–15 June 1990
- Kanungo T, Haralick RM, Phillips I (1994) Nonlinear local and global document degradation models. Int J Imaging Syst Technol 5(4):220–230
- Sarkar P, Nagy G, Zhou J, Lopresti D (1998) Spatial sampling of printed patterns. IEEE Trans Pattern Anal Mach Intell 20(3):344–351
- Barney Smith EH (2005) Document scanning. McGraw-Hill 2005 yearbook of science and technology. McGraw-Hill
- Barney Smith EH (2009) A new metric describes the edge noise in bi-level images. SPIE Newsroom, Nov 2009. doi:10.1117/2.1200910.1829
- Sharma G (2001) Show-through cancellation in scans of duplex printed documents. IEEE Trans Image Process 10(5):736–754
- Tonazzini A, Salerno E, Bedini L (2007) Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. IJDAR 10:17–25
- Khan MR, Hasan MK (2010) A novel model for show-through in scan of duplex printed documents. Signal Image Video Process, 1–21