2

Surface Roughness Analysis and Measurement Techniques

Bharat Bhushan
The Ohio State University

2.1 The Nature of Surfaces

A solid surface, or more exactly a solid–gas or solid–liquid interface, has a complex structure and complex properties depending on the nature of the solids, the method of surface preparation, and the interaction between the surface and the environment. Properties of solid surfaces are crucial to surface interaction because surface properties affect real area of contact, friction, wear, and lubrication. In addition to tribological functions, surface properties are important in other applications, such as optical, electrical and thermal performance, painting, and appearance.

Solid surfaces, irrespective of their method of formation, contain irregularities or deviations from the prescribed geometrical form (Whitehouse, 1994; Bhushan, 1996, 1999a,b; Thomas, 1999). The surfaces contain irregularities of various orders ranging from shape deviations to irregularities of the order of interatomic distances. No machining method, however precise, can produce a molecularly flat surface on conventional materials. Even the smoothest surfaces, such as those obtained by cleavage of some crystals, contain irregularities, the heights of which exceed the interatomic distances. For technological applications, both macro- and micro/nanotopography of the surfaces (surface texture) are important (Bhushan, 1999a,b).

In addition to surface deviations, the solid surface itself consists of several zones having physico-chemical properties peculiar to the bulk material itself (Figure 2.1) (Gatos, 1968; Haltner, 1969; Buckley, 1981). As a result of the forming process in metals and alloys, there is a zone of work-hardened or

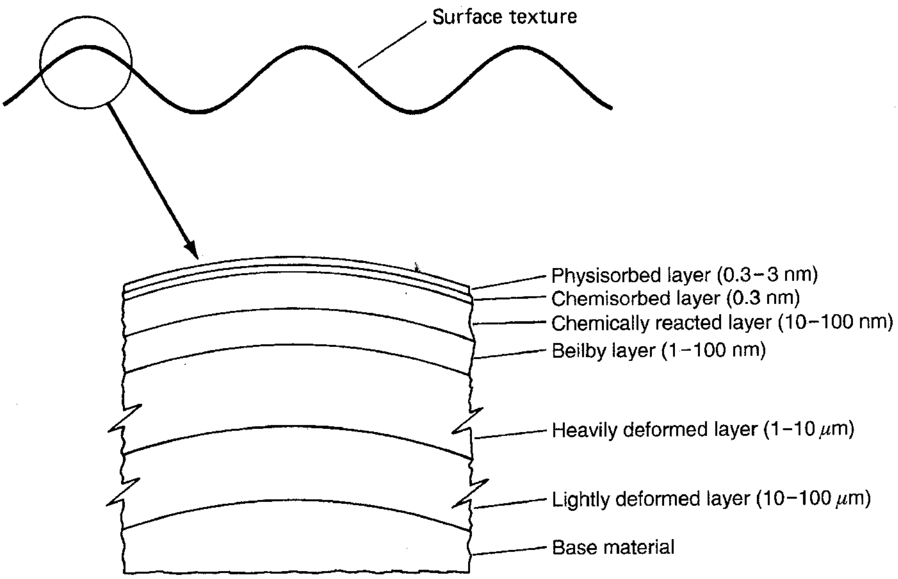
Surface texture
Physisorbed layer (0.3–3 nm)
Chemisorbed layer (0.3 nm)
Chemically reacted layer (10–100 nm)
Beilby layer (1–100 nm)
Heavily deformed layer (1–10 μm)
Lightly deformed layer (10–100 μm)
Base material

FIGURE 2.1 Solid surface details: surface texture (vertical axis magnified) and typical surface layers.

deformed material on top of which is a region of microcrystalline or amorphous structure that is called the Beilby layer. Deformed layers would also be present in ceramics and polymers. These layers are extremely important because their properties, from a surface chemistry point of view, can be entirely different from the annealed bulk material. Likewise, their mechanical behavior is influenced by the amount and depth of deformation of the surface layers.

Many of the surfaces are chemically reactive. With the exception of noble metals, all metals and alloys and many nonmetals form surface oxide layers in air, and in other environments they are likely to form other layers (for example, nitrides, sulfides, and chlorides). Besides the chemical corrosion film, there are also adsorbed films that are produced either by physisorption or chemisorption of oxygen, water vapor, and hydrocarbons, from the environment. Occasionally, there will be a greasy or oily film derived from the environment. These films are found on metallic and nonmetallic surfaces.

The presence of surface films affects friction and wear. The effect of adsorbed films, even a fraction of a monolayer, is significant on the surface interaction. Sometimes, the films wear out in the initial running period and subsequently have no effect. The effect of greasy or soapy film, if present, is more marked; it reduces the severity of surface interaction often by one or more orders of magnitude.

This chapter covers the details on the analysis and measurement of surface roughness.

2.2 Analysis of Surface Roughness

Surface texture is the repetitive or random deviation from the nominal surface that forms the three-dimensional topography of the surface. Surface texture includes (1) roughness (nano- and microroughness), (2) waviness (macroroughness), (3) lay, and (4) flaws. Figure 2.2 is a pictorial display of surface texture with unidirectional lay (Anonymous, 1985).

Nano- and microroughness are formed by fluctuations in the surface of short wavelengths, characterized by hills (asperities) (local maxima) and valleys (local minima) of varying amplitudes and spacings. These are large compared to molecular dimensions. Asperities are referred to as peaks in a profile (two dimensions) and summits in a surface map (three dimensions). Nano- and microroughness include those features intrinsic to the production process. These are considered to include traverse feed marks and other irregularities within the limits of the roughness sampling length. Waviness is the surface irregularity

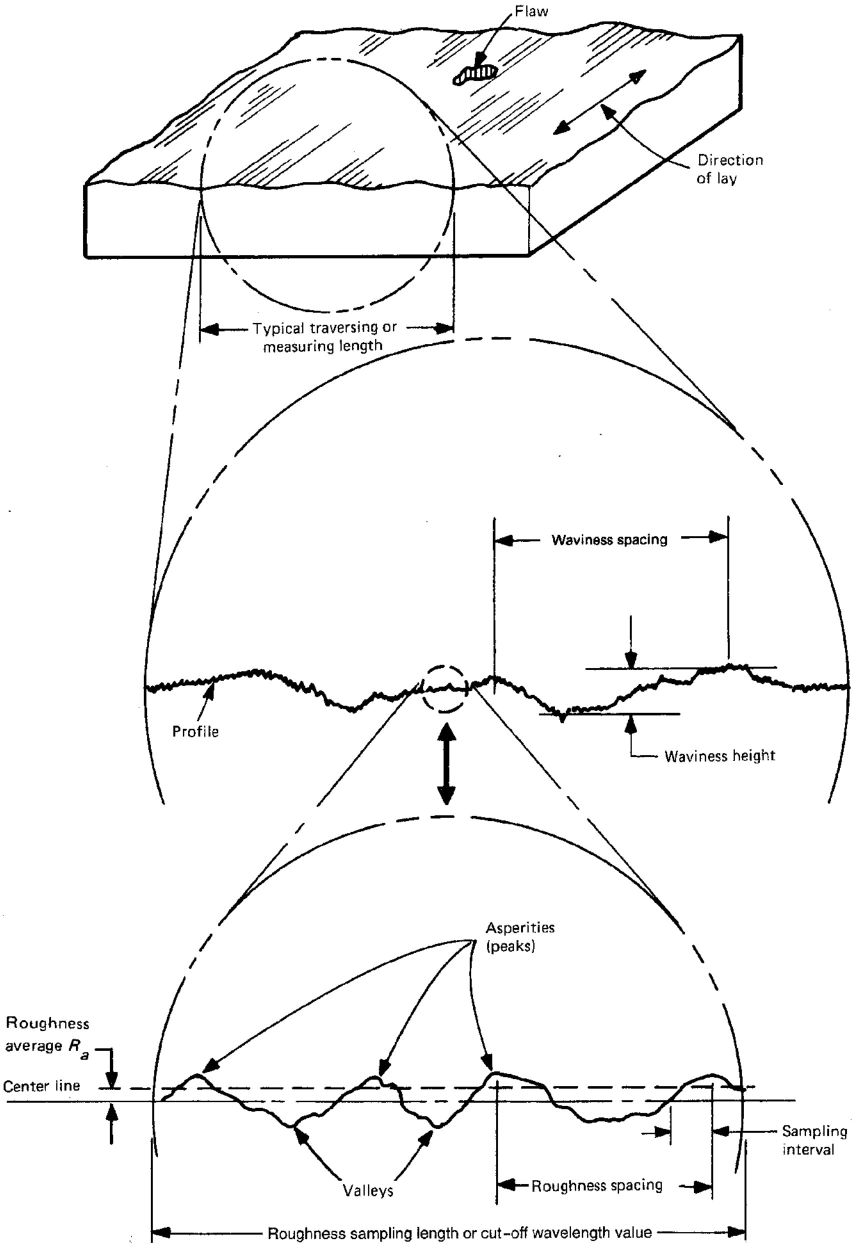

FIGURE 2.2 Pictorial display of surface texture. (From Anonymous (1985), Surface Texture (Surface Roughness, Waviness and Lay), ANSI/ASME B46.1, ASME, New York. With permission.)

of longer wavelengths and is referred to as macroroughness. Waviness may result from such factors as machine or workpiece deflections, vibration, chatter, heat treatment, or warping strains. Waviness includes all irregularities whose spacing is greater than the roughness sampling length and less than the waviness sampling length. Lay is the principal direction of the predominant surface pattern, ordinarily determined by the production method. Flaws are unintentional, unexpected, and unwanted interruptions in the texture. In addition, the surface may contain gross deviations from nominal shape of very long wavelength, which is known as errors of form. They are not normally considered part of the surface

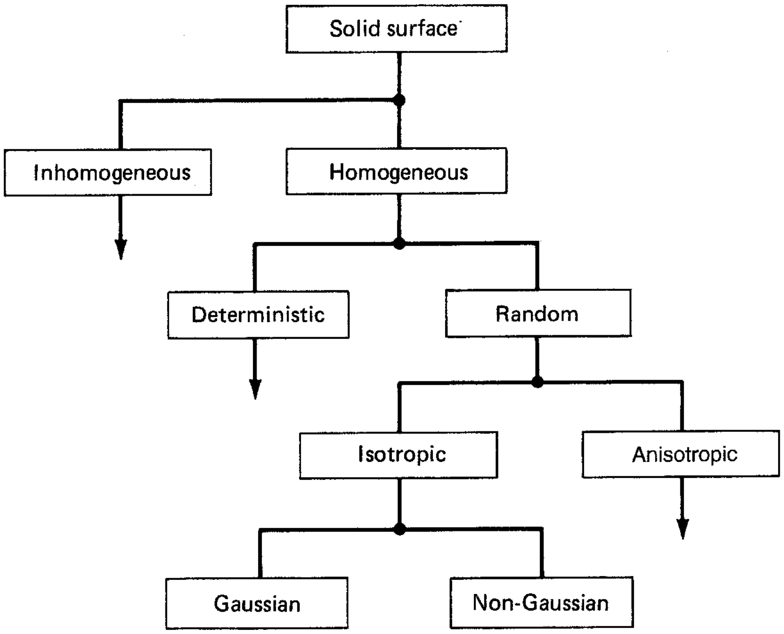

FIGURE 2.3 General typology of surfaces.

texture. A question often asked is whether various geometrical features should be assessed together or separately. What features are included together depends on the applications. It is generally not possible to measure all features at the same time.

A very general typology of a solid surface is seen in Figure 2.3. Surface textures that are deterministic may be studied by relatively simple analytical and empirical methods; their detailed characterization is straightforward. However, the textures of most engineering surfaces are random, either isotropic or anisotropic, and either Gaussian or non-Gaussian. Whether the surface height distribution is isotropic or anisotropic and Gaussian or non-Gaussian depends upon the nature of the processing method. Surfaces that are formed by cumulative processes (such as peening, electropolishing, and lapping), in which the final shape of each region is the cumulative result of a large number of random discrete local events and irrespective of the distribution governing each individual event, will produce a cumulative effect that is governed by the Gaussian form. It is a direct consequence of the central limit theorem of statistical theory. Single-point processes (such as turning and shaping) and extreme-value processes (such as grinding and milling) generally lead to anisotropic and non-Gaussian surfaces. The Gaussian (normal) distribution has become one of the mainstays of surface classification.

In this section, we first define average roughness parameters, followed by statistical analyses and fractal characterization of surface roughness that are important in contact problems. Emphasis is placed on random, isotropic surfaces that follow Gaussian distribution.

2.2.1 Average Roughness Parameters

2.2.1.1 Amplitude Parameters

Surface roughness most commonly refers to the variations in the height of the surface relative to a reference plane. It is measured either along a single line profile or along a set of parallel line profiles (surface maps). It is usually characterized by one of the two statistical height descriptors advocated by the American National Standards Institute (ANSI) and the International Standardization Organization (ISO) (Anonymous, 1975, 1985). These are (1) R_a, CLA (center-line average), or AA (arithmetic average) and (2) the standard deviation or variance (σ), R_q or root mean square (RMS). Two other statistical

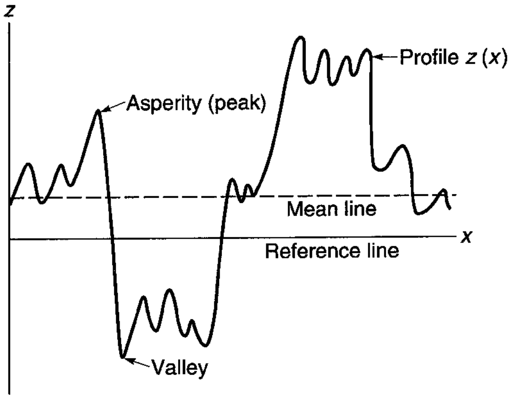

FIGURE 2.4 Schematic of a surface profile z(x).

height descriptors are skewness (Sk) and kurtosis (K); these are rarely used. Another measure of surface roughness is an extreme-value height descriptor (Anonymous, 1975, 1985) R_t (or R_y, R_{max}, or maximum peak-to-valley height or simply P–V distance). Four other extreme-value height descriptors in limited use, are: R_p (maximum peak height, maximum peak-to-mean height or simply P–M distance), R_v (maximum valley depth or mean-to-lowest valley height), R_z (average peak-to-valley height), and R_{pm} (average peak-to-mean height).

We consider a profile, z(x), in which profile heights are measured from a reference line Figure 2.4. We define a center line or mean line such that the area between the profile and the mean line above the line is equal to that below the mean line. R_a, CLA, or AA is the arithmetic mean of the absolute values of vertical deviation from the mean line through the profile. The standard deviation σ is the square root of the arithmetic mean of the square of the vertical deviation from the mean line.

In mathematical form, we write

$$R_a = CLA = AA = \frac{1}{L}\int_0^L |z - m|\, dx \tag{2.1a}$$

and

$$m = \frac{1}{L}\int_0^L z\, dx \tag{2.1b}$$

where L is the sampling length of the profile (profile length).

The variance is given as

$$\sigma^2 = \frac{1}{L}\int_0^L (z - m)^2\, dx \tag{2.2a}$$

$$= R_q^2 - m^2 \tag{2.2b}$$

where, σ is the standard deviation and R_q is the square root of the arithmetic mean of the square of the vertical deviation from a reference line, or

$$R_q^2 = RMS^2 = \frac{1}{L}\int_0^L \left(z^2\right) dx \tag{2.3a}$$

TABLE 2.1 Center-Line Average and Roughness Grades

R_a Values up to a Value in μm	Roughness Grade Number
0.025	N1
0.05	N2
0.1	N3
0.2	N4
0.4	N5
0.8	N6
1.6	N7
3.2	N8
6.3	N9
12.5	N10
25.0	N11

For the special case where *m* is equal to zero,

$$R_q = \sigma \tag{2.3b}$$

In many cases, the R_a and σ are interchangeable, and for Gaussian surfaces,

$$\sigma \sim \sqrt{\frac{\pi}{2}}\, R_a \sim 1.25\, R_a \tag{2.4}$$

The value of R_a is an official standard in most industrialized countries. Table 2.1 gives internationally adopted R_a values together with the alternative roughness grade number. The σ is most commonly used in statistical analyses.

The skewness and kurtosis in the normalized form are given as

$$Sk = \frac{1}{\sigma^3 L} \int_0^L (z - m)^3\, dx \tag{2.5}$$

and

$$K = \frac{1}{\sigma^4 L} \int_0^L (z - m)^4\, dx \tag{2.6}$$

More discussion of these two descriptors will be presented later.

Five extreme-value height descriptors are defined as follows: R_t is the distance between the highest asperity (peak or summit) and the lowest valley; R_p is defined as the distance between the highest asperity and the mean line; R_v is defined as the distance between the mean line and the lowest valley; R_z is defined as the distance between the averages of five highest asperities and the five lowest valleys; and R_{pm} is defined as the distance between the averages of the five highest asperities and the mean line. The reason for taking an average value of asperities and valleys is to minimize the effect of unrepresentative asperities or valleys which occasionally occur and can give an erroneous value if taken singly. R_z and R_{pm} are more reproducible and are advocated by ISO. In many tribological applications, height of the highest asperities above the mean line is an important parameter because damage may be done to the interface by the few high asperities present on one of the two surfaces; on the other hand, valleys may affect lubrication retention and flow.

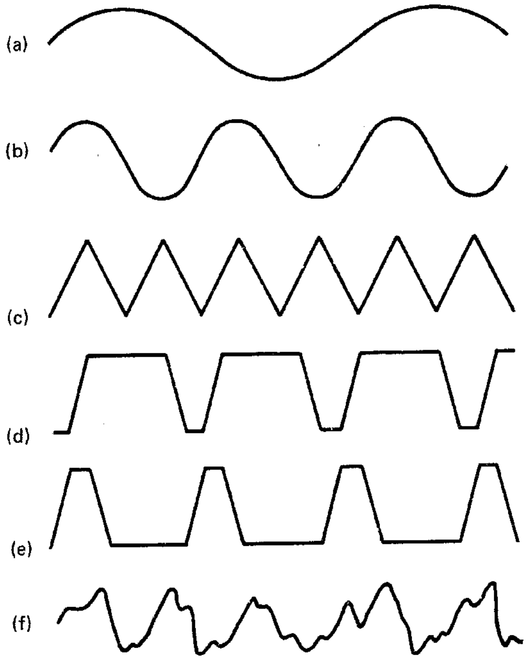

FIGURE 2.5 Various surface profiles having the same R_a value.

The height parameters R_a (or σ in some cases) are R_t (or R_p in some cases) are most commonly specified for machine components. For the complete characterization of a profile or a surface, none of the parameters discussed earlier are sufficient. These parameters are seen to be primarily concerned with the relative departure of the profile in the vertical direction only; they do not provide any information about the slopes, shapes, and sizes of the asperities or about the frequency and regularity of their occurrence. It is possible, for surfaces of widely differing profiles with different frequencies and different shapes, to give the same R_a or σ (R_q) values (Figure 2.5). These single numerical parameters are useful mainly for classifying surfaces of the same type that are produced by the same method.

Average roughness parameters for surface maps are calculated using the same mathematical approach as that for a profile presented here.

2.2.1.2 Spacing (or Spatial) Parameters

One way to supplement the amplitude (height) information is to provide some index of crest spacing or wavelength (which corresponds to lateral or spatial distribution) on the surface. Two parameters occasionally used are the peak (or summit) density, N_p (η), and zero crossings density, N_0. N_p is the density of peaks (local maxima) of the profile in number per unit length, and η is the density of summits on the surface in number per unit area. N_p and η are just measures of maxima irrespective of height. This parameter is in some use. N_0 is the zero crossings density defined as the number of times the profile crosses the mean line per unit length. From Longuet–Higgins (1957a), the number of surface zero crossings per unit length is given by the total length of the contour where the autocorrelation function (to be described later) is zero (or 0.1) divided by the area enclosed by the contour. This count N_0 is rarely used.

A third parameter — mean peak spacing (A_R) — is the average distance between measured peaks. This parameter is merely equal to ($1/N_p$). Other spacial parameters rarely used are the mean slope and mean curvature, which are the first and second derivatives of the profile/surface, respectively.

2.2.2 Statistical Analyses

2.2.2.1 Amplitude Probability Distribution and Density Functions

The cumulative probability distribution function, or simply cumulative distribution function (CDF), P(h) associated with the random variable z(x), which can take any value between –∞ and ∞ or z_{min} and z_{max}, is defined as the probability of the event z(x) ≤ h, and is written as (McGillem and Cooper, 1984; Bendat and Piersol, 1986)

$$P(h) = \text{Prob}(z \le h) \tag{2.7}$$

with P (–∞) = 0 and P (∞) = 1.

It is common to describe the probability structure of random data in terms of the slope of the distribution function given by the derivative

$$p(z) = \frac{dP(z)}{dz} \tag{2.8a}$$

where the resulting function p(z) is called the probability density function (PDF). Obviously, the cumulative distribution function is the integral of the probability density function p(z), that is,

$$P(z \le h) = \int_{-\infty}^{h} p(z)dz = P(h) \tag{2.8b}$$

and

$$P(h_1 \le z \le h_2) = \int_{h_1}^{h_2} p(z)dz = P(h_2) - P(h_1) \tag{2.8c}$$

Furthermore, the total area under the probability density function must be unity; that is, it is certain that the value of z at any x must fall somewhere between plus and minus infinity or z_{max} and z_{min}.

The data representing a wide collection of random physical phenomenon in practice tend to have a Gaussian or normal probability density function,

$$p(z) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{(z-m)^2}{2\sigma^2}\right] \tag{2.9a}$$

where σ is the standard deviation and m is the mean.

For convenience, the Gaussian function is plotted in terms of a normalized variable,

$$z^* = (z - m)/\sigma \tag{2.9b}$$

which has zero mean and unity standard deviation. With this transformation of variables, Equation 2.9 becomes

$$p(z^*) = \frac{1}{(2\pi)^{1/2}} \exp\left[\frac{-(z^*)^2}{2}\right] \tag{2.9c}$$

which is called the standardized Gaussian or normal probability density function. To obtain P(h) from p(z*) of Equation 2.9**c**, the integral cannot be performed in terms of the common functions, and the integral is often listed in terms of the "error function" and its values are listed in most statistical textbooks. The error function is defined as

$$\operatorname{erf}(h) = \frac{1}{(2\pi)^{1/2}} \int_0^h \exp\left[\frac{-(z^*)^2}{2}\right] dz^* \tag{2.10}$$

An example of a random variable z*(x) with its Gaussian probability density and corresponding cumulative distribution functions is shown in Figure 2.6. Examples of P(h) and P(z* = h) are also shown. The probability density function is bell shaped, and the cumulative distribution function is S-shaped.

We further note that for a Gaussian function

$$P(-1 \le z^* \le 1) = 0.682$$

$$P(-2 \le z^* \le 2) = 0.954$$

$$P(-3 \le z^* \le 3) = 0.999$$

and

$$P(-\infty \le z^* \le \infty) = 1$$

which implies that the probabilities that some number that follows a Gaussian distribution is within the limits of ±1σ, ±2σ, and ±3σ are 68.2, 95.4, and 99.9%, respectively.

A convenient method for testing for Gaussian distribution is to plot the cumulative distribution function on probability graph paper to show the percentage of the numbers below a given number; this is scaled such that a straight line is produced when the distribution is Gaussian (typical data to be

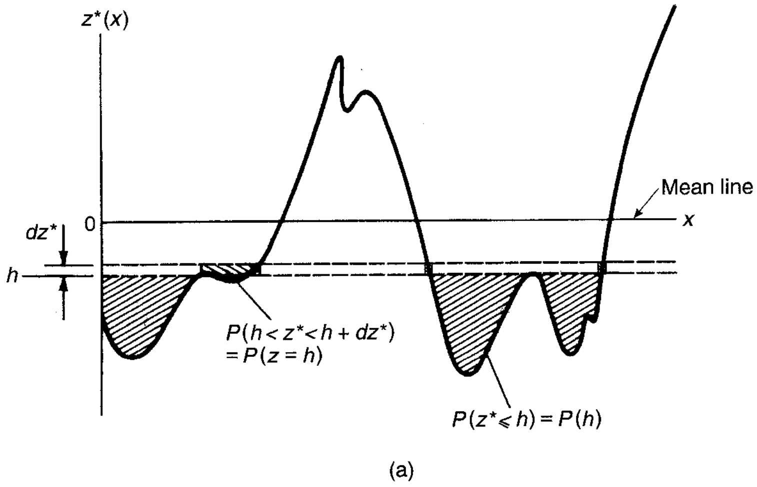

FIGURE 2.6 (a) Random function z*(x), which follows Gaussian probability functions.

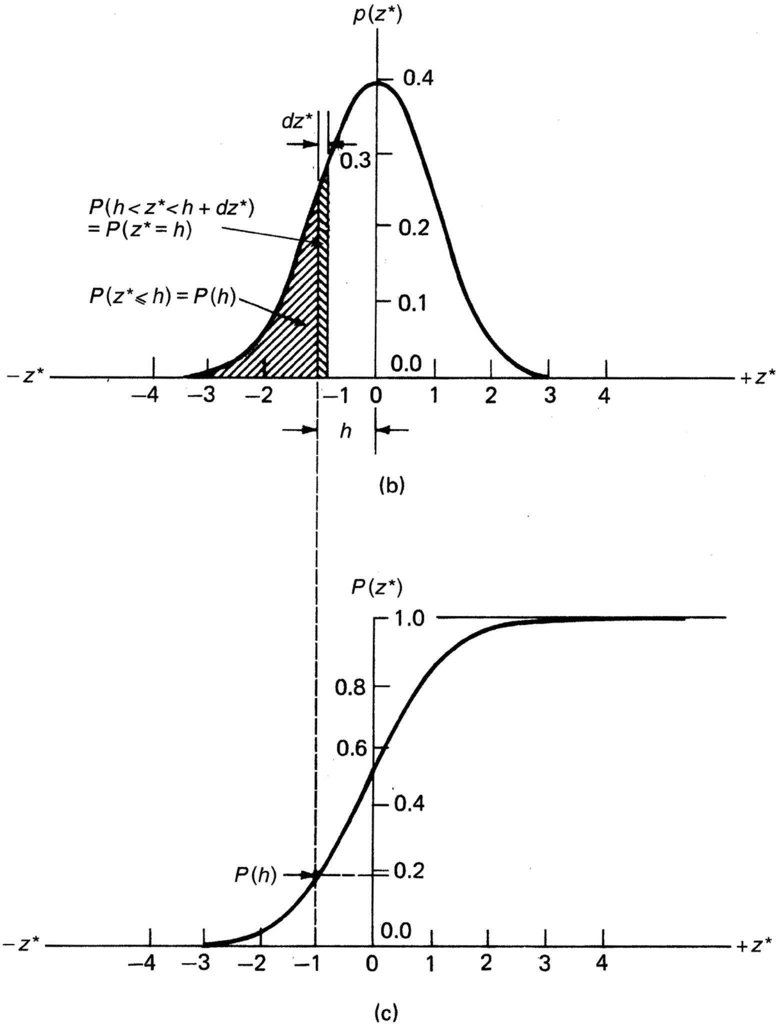

FIGURE 2.6 (b) Gaussian probability density function p(z*), and (c) Gaussian probability distribution function P(z*).

presented later). To test for Gaussian distribution, a straight line corresponding to a Gaussian distribution is drawn on the plot. The slope of the straight line portion is determined by σ, and the position of the line for 50% probability is set at the mean value (which is typically zero for surface height data).

The most practical method for the goodness of the fit between the given distribution and the Gaussian distribution is to use the Kolmogorov–Smirnov test (Smirnov, 1948; Massey, 1951; Siegel, 1956). In the Kolmogorov–Smirnov test, the maximum departure between the percentage of the numbers above a given number for the data and the percentage of the numbers that would be above a given number if the given distribution were a Gaussian distribution is first calculated. Then, a calculation is made to determine if the distribution is indeed Gaussian. The level of significance, P, is calculated; this gives the probability of mistakenly or falsely rejecting the hypothesis that the distribution is a Gaussian distribution. Common minimum values for P for accepting the hypothesis are 0.01 to 0.05 (Siegel, 1956). The chi-square test (Siegel, 1956) can also be used to determine how well the given distribution matches a Gaussian distribution. However, the chi-square test is not very useful because the goodness of fit calculated depends too much on how many bins or discrete cells the surface height data are divided into (Wyant et al., 1986).

For the sake of mathematical simplicity in some analyses, sometimes an exponential distribution is used instead of the Gaussian distribution. The exponential distribution is given as

$$p(z) = \frac{1}{\sigma}\exp\left[-\frac{(z-m)}{\sigma}\right] \tag{2.11a}$$

or

$$p(z^*) = \exp(-z^*) \tag{2.11b}$$

2.2.2.2 Moments of Amplitude Probability Functions

The shape of the probability density function offers useful information on the behavior of the process. This shape can be expressed in terms of moments of the function,

$$m_n = \int_{-\infty}^{\infty} z^n p(z)\, dz \tag{2.12}$$

m_n is called the *n*th moment. Moments about the mean are referred to as central moments,

$$m_n^c = \int_{-\infty}^{\infty} (z-m)^n p(z)\, dz \tag{2.13}$$

The zeroth moment (n = 0) is equal to 1. The first moment is equal to m, mean value of the function z(x), whereas the first central moment is equal to zero. For completeness, we note that

$$R_a = \int_{-\infty}^{\infty} |z-m|\, p(z)\, dz \tag{2.14}$$

The second moments are,

$$m_2 = \int_{-\infty}^{\infty} z^2 p(z)\, dz = R_q^2 \tag{2.15}$$

and

$$m_2^c = \int_{-\infty}^{\infty} (z-m)^2 p(z)\, dz = \sigma^2 \tag{2.16a}$$

$$= R_q^2 - m^2 \tag{2.16b}$$

The third moment m_3^c is the skewness (Sk). A useful parameter in defining variables with an asymmetric spread, it represents the degree of symmetry of the distribution function (Figure 2.7). It is usual to normalize the third central moment as,

$$Sk = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} (z-m)^3 p(z)\, dz \tag{2.17}$$

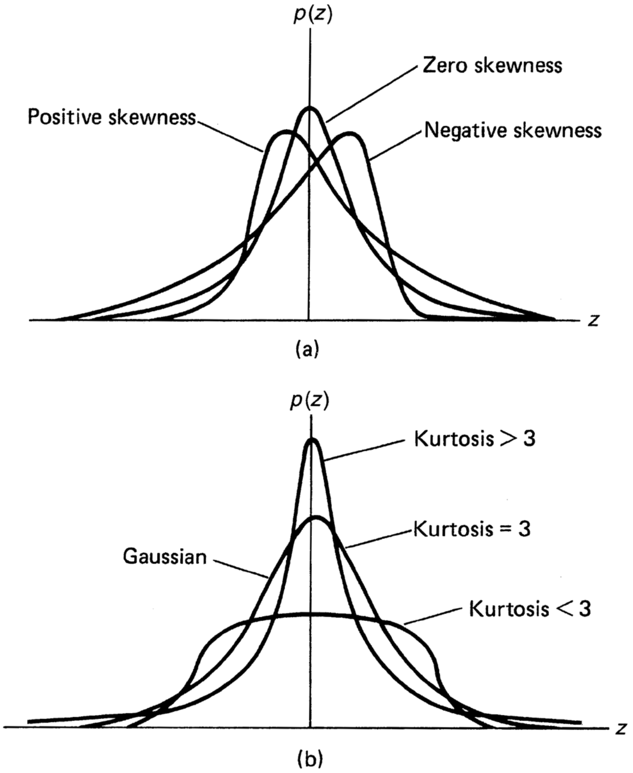

FIGURE 2.7 (a) Probability density functions for random distributions with different skewness, and for (b) symmetrical distributions (zero skewness) with different kurtosis.

Symmetrical distribution functions, including Gaussian, have zero skewness.

The fourth moment m_4^c is the kurtosis (K). It represents the peakedness of the distribution and is a measure of the degree of pointedness or bluntness of a distribution function (Figure 2.7). Again, it is usual to normalize the fourth central moment as,

$$K = \frac{1}{\sigma^4} \int_{-\infty}^{\infty} (z - m)^4 \, p(z) \, dz \qquad (2.18)$$

Note that the symmetric Gaussian distribution has a kurtosis of 3. Distributions with K > 3 are called leptokurtic, and those with K < 3 are called platykurtic.

Kotwal and Bhushan (1996) developed an analytical method to generate probability density functions for non-Gaussian distributions using the so-called Pearson system of frequency curves based on the methods of moments (Elderton and Johnson, 1969). (For a method of generating non-Gaussian distributions on the computer, see Chilamakuri and Bhushan [1998].) The probability density functions generated by this method for selected skewness and kurtosis values are presented in Table 2.2. These functions are plotted in Figure 2.8. From this figure, it can be seen that a Gaussian distribution with zero skewness and a kurtosis of three has an equal number of local maxima and minima at a certain height above and below the mean line. A surface with a high negative skewness has a larger number of local maxima above the mean as compared to a Gaussian distribution; for a positive skewness the converse is true, Figure 2.9**.** Similarly, a surface with a low kurtosis has a larger number of local maxima above the mean as compared to that of a Gaussian distribution; again, for a high kurtosis the converse is true (Figure 2.9).

TABLE 2.2 Probability Density Functions for Surfaces with Various Skewness and Kurtosis Values Based on the Pearson's System of Frequency Curves

Non-Gaussian Parameters		Number	
Sk	K	of Type	Probability Density Function, $p(z^*)$
–0.8	3	I	$p(z^*) = 0.33(1 + z^*/3.86)^{2.14}(1 - z^*/1.36)^{0.11}$
–0.5	3	I	$p(z^*) = 0.38(1 + z^*/6.36)^{9.21}(1 - z^*/2.36)^{2.79}$
–0.3	3	I	$p(z^*) = 0.39(1 + z^*/10.72)^{29.80}(1 - z^*/4.05)^{10.64}$
0.0	3	Normal	$p(z^*) = 0.3989\exp\left(-0.5(z^*)^2\right)$
0.3	3	I	$p(z^*) = 0.39(1 + z^*/4.05)^{10.64}(1 - z^*/10.72)^{29.80}$
0.5	3	I	$p(z^*) = 0.38(1 + z^*/2.36)^{2.79}(1 - z^*/6.36)^{9.21}$
0.8	3	I	$p(z^*) = 0.33(1 + z^*/1.36)^{0.11}(1 - z^*/3.86)^{2.14}$
0.0	2	II	$p(z^*) = 0.32\left(1 - (z^*)^2/16\right)^{0.5}$
0.0	3	Normal	$p(z^*) = 0.3989\exp\left(-0.5(z^*)^2\right)$
0.0	5	VII	$p(z^*) = 0.46\left(1 + (z^*)^2/25\right)^{-4}$
0.0	10	VII	$p(z^*) = 0.49\left(1 + (z^*)^2/8.20\right)^{-2.92}$
0.0	20	VII	$p(z^*) = 0.51\left(1 + (z^*)^2/5.52\right)^{-2.68}$

$z^* = z/\sigma$

In practice, many engineering surfaces have symmetrical Gaussian height distribution. Experience with most engineering surfaces shows that the height distribution is Gaussian at the high end, but at the lower end, the bottom 1 to 5% of the distribution is generally found to be non-Gaussian (Williamson, 1968). Many of the common machining processes produce surfaces with non-Gaussian distribution, Figure 2.10. Turning, shaping, and electrodischarge machining (EDM) processes produce surfaces with positive skewness. Grinding, honing, milling, and abrasion processes produce grooved surfaces with negative skewness but high kurtosis values. Laser polishing produces surfaces with high kurtosis.

2.2.2.3 Surface Height Distribution Functions

If the surface or profile heights are considered as random variables, then their statistical representation in terms of the probability density function p(z) is known as the height distribution, or a histogram. The height distribution can also be represented as cumulative distribution function P(z). For a digitized profile, the histogram is constructed by plotting the number or fraction of surface heights lying between two specific heights as a function of height (Figure 2.11)**.** The interval between two such heights is termed the class interval and is shown as dz in Figure 2.11. It is generally recommended to use 15 to 50 class intervals for general random data, but the choice is usually a trade-off between accuracy and resolution. Similarly, from the surface or profile height distribution, the cumulative distribution function is derived. It is constructed by plotting the cumulative number or proportion of the surface height lying at or below a specific height as a function of that height (Figure 2.11)**.** An example of a profile and corresponding histogram and cumulative height distribution on a probability paper for a lapped nickel–zinc ferrite is given in Figure 2.12.

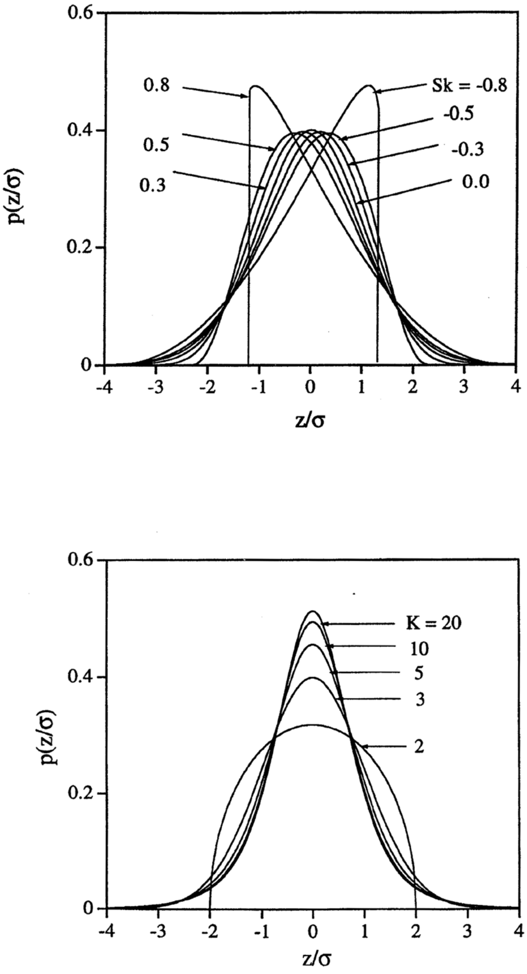

FIGURE 2.8 Probability density function for random distributions with selected skewness and kurtosis values.

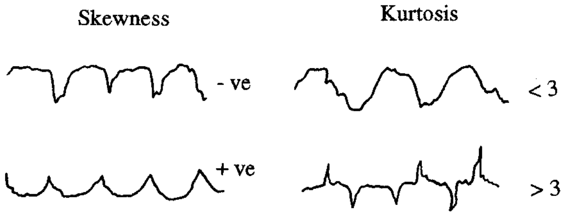

FIGURE 2.9 Schematic illustration for random functions with various skewness and kurtosis values.

Probability density and distribution curves can also be obtained for the slope and curvature of the surface or the profile. If the surface, or profile height, follows a Gaussian distribution, then its slope and curvature distribution also follows a Gaussian distribution. Because it is known that if two functions follow a Gaussian distribution, their sum and difference also follow a Gaussian distribution. Slope and curvatures are derived by taking the difference in a height distribution, and therefore slope and curvatures of a Gaussian height distribution would be Gaussian.

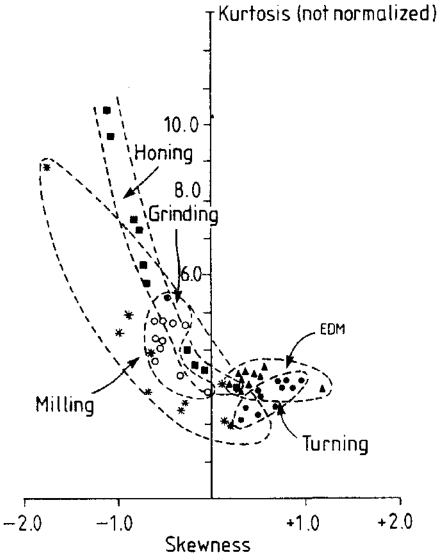

FIGURE 2.10 Typical skewness and kurtosis envelopes for various manufacturing processes (From Whitehouse, D.I. (1994), *Handbook of Surface Metrology,* Institute of Physics Publishing, Bristol, U.K. With permission.)

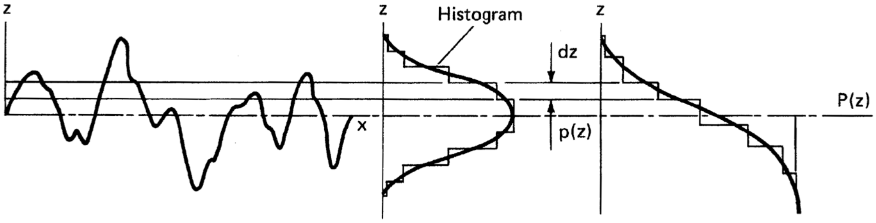

FIGURE 2.11 Method of deriving the histogram and cumulative distribution function from a surface height distribution.

For a digitized profile of length L with heights z_i, i = 1 to N, at a sampling interval $\Delta x = L/(N - 1)$, where N represents the number of measurements, average height parameters are given as

$$R_a = \frac{1}{N}\sum_{i=1}^{N}\left|z_i - m\right| \tag{2.19a}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(z_i - m\right)^2 \tag{2.19b}$$

$$Sk = \frac{1}{\sigma^3 N}\sum_{i=1}^{N}\left(z_i - m\right)^3 \tag{2.19c}$$

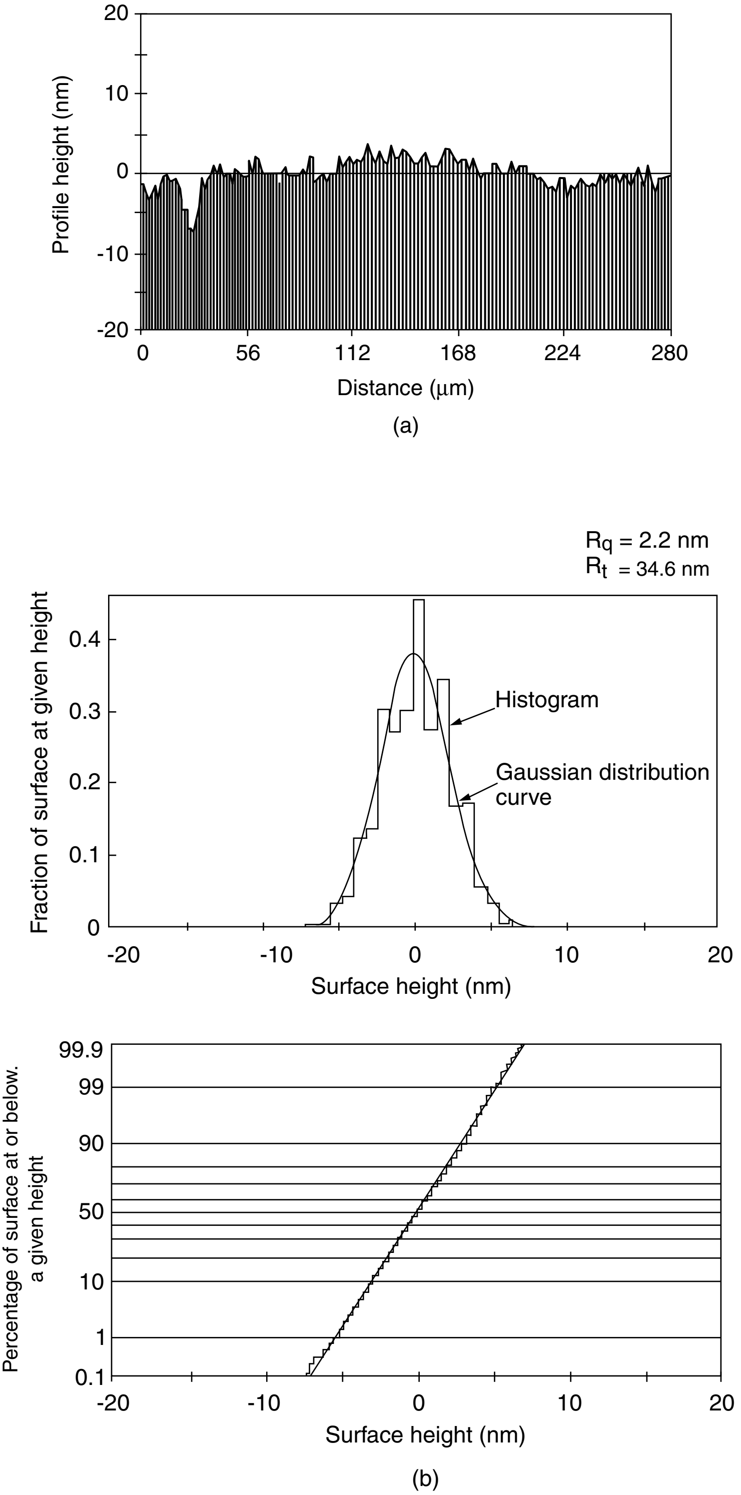

FIGURE 2.12 (a) Profile and (b) corresponding histogram and distribution of profile heights of lapped nickel–zinc ferrite.

$$K = \frac{1}{\sigma^4 N} \sum_{i=1}^{N} \left(z_i - m\right)^4 \tag{2.19d}$$

and

$$m = \frac{1}{N} \sum_{i=1}^{N} z_i \tag{2.19e}$$

Two average spacing parameters, mean of profile slope $\left(\frac{\partial z}{\partial x}\right)$ and profile curvature $\left(-\frac{\partial^2 z}{\partial x^2}\right)$ of a digitized profile, are given as

$$\text{mean slope} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(\frac{z_{i+1} - z_i}{\Delta x}\right) \tag{2.20a}$$

$$\text{and mean curvature} = \frac{1}{N-2} \sum_{i=2}^{N-1} \left(\frac{2z_i - z_{i-1} - z_{i+1}}{\Delta x^2}\right) \tag{2.20b}$$

The surface slope at any point on a surface is obtained by finding the square roots of the sum of the squares of the slopes in two orthogonal (x and y) axes. The curvature at any point on the surface is obtained by finding the average of the curvatures in two orthogonal (x and y) axes (Nayak, 1971).

Before calculation of roughness parameters, the height data are fitted in a least-square sense to determine the mean height, tilt, and curvature. The mean height is *always* subtracted, and the tilt is *usually* subtracted. In some cases, curvature needs to be removed as well. Spherical and cylindrical radii of curvature are removed for spherical and cylindrical surfaces, respectively (e.g., balls and cylinders), before roughness parameters are calculated.

2.2.2.4 Bearing Area Curves

The real area of contact (to be discussed in the next chapter) is known as the bearing area and may be approximately obtained from a surface profile or a surface map. The bearing area curve (BAC) first proposed by Abbott and Firestone (1933) is also called the Abbott–Firestone curve or simply the Abbott curve. It gives the ratio of air to material at any level, starting at the highest peak, called the *bearing ratio* or *material ratio*, as a function of level.

To produce a BAC from a surface profile, a parallel line (bearing line) is drawn some distance from a reference (or mean) line. The length of each material intercept (land) along the line is measured and these lengths are summed. The proportion of this sum to the total length, the bearing length ratio (t_p), is calculated. This procedure is repeated along a number of bearing lines, starting with the highest peak to the lowest valley, and the fractional land length (bearing length ratio) as a function of the height of each slice from the highest peak (cutting depth) is plotted (Figure 2.13). For a Gaussian surface, the BAC has an S-shaped appearance. In the case of a surface map, bearing planes are drawn, and the area of each material intercept is measured. For a random surface, the bearing length and bearing area fractions are numerically identical.

The BAC is related to the CDF. The fraction of heights lying above a given height z (i.e., the bearing ratio at height h) is given by

$$\text{Prob}\left(z \geq h\right) = \int_{h}^{\infty} p(z)\, dz \tag{2.21a}$$

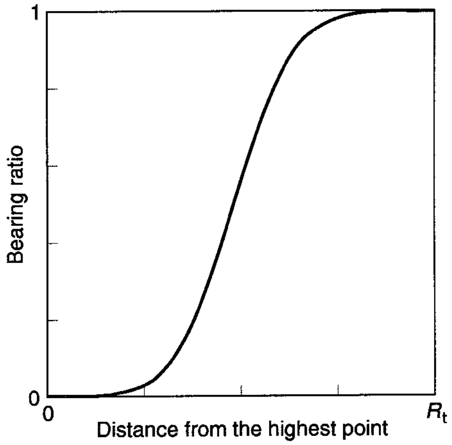

FIGURE 2.13 Schematic of bearing area curve.

which is 1 – P(h), where P(h) is the cumulative distribution function at z ≤ h (Figure 2.6). Therefore, the BAC can be obtained from the height distribution histogram. The bearing ratio histograph at height h is simply the progressive addition of all the values of p(z) starting at the highest point and working down to the height z = h, and this cumulative sum multiplied by the class interval Δz is

$$P(z \geq h) = \Delta z \sum_{z=h}^{\infty} p(z) \qquad (2.21b)$$

The relationship of bearing ratio to the fractional real area of contact is highly approximate as material is sliced off in the construction of BAC and the material deformation is not taken into account.

2.2.2.5 Spatial Functions

Consider two surfaces with sine wave distributions with the same amplitude but different frequencies. We have shown that these will have the same R_a and σ, but with different spatial arrangements of surface heights. Slope and curvature distributions are not, in general, sufficient to represent the surface, as they refer only to one particular spatial size of features. The spatial functions (McGillem and Cooper, 1984; Bendat and Piersol, 1986), namely the autocovariance (or autocorrelation) function (ACVF), structure function (SF), or power spectral (or autospectral) density function (PSDF), offer a means of representing the properties of all wavelengths, or spatial sizes of the feature; these are also known as surface texture descriptors.

ACVF has been the most popular way of representing spatial variation. The ACVF of a random function is most directly interpreted as a measure of how well future values of the function can be predicted based on past observations. SF contains no more information than the ACVF. The PSDF is interpreted as a measure of frequency distribution of the mean square value of the function, that is the rate of change of the mean square value with frequency. In this section, we will present the definitions for an isotropic and random profile z(x). Definitions for an isotropic surface z(x,y) can be found in a paper by Nayak (1971). Analysis of an anisotropic surface is considerably complicated by the number of parameters required to describe the surface. For example, profile measurements along three different directions are needed for complete surface characterization of selected anisotropic surfaces. For further details on anisotropic surfaces, see Longuet-Higgins (1957a), Nayak (1973), Bush et al. (1979), and Thomas (1982).

Autocovariance and Autocorrelation Functions

For a function z(x), the ACVF for a spatial separation of τ is an average value of the product of two measurements taken on the profile a distance τ apart, z(x) and z(x + τ). It is obtained by comparing the function z(x) with a replica of itself where the replica is shifted an amount τ (Figure 2.14),

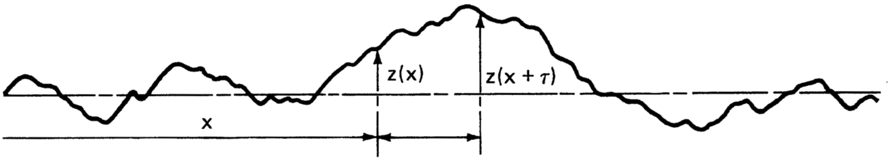

FIGURE 2.14 Construction of the autocovariance function.

$$R(\tau) = \lim_{L \to \infty} \frac{1}{L} \int_0^L z(x) z(x+\tau)\, dx \tag{2.22a}$$

where L is the sampling length of the profile. From its definition, ACVF is always an even function of τ, that is,

$$R(\tau) = R(-\tau) \tag{2.22b}$$

The values of ACVF at τ = 0 and ∞ are,

$$R(0) = R_q^2 = \sigma^2 + m^2 \tag{2.22c}$$

and

$$R(\infty) = m^2 \tag{2.22d}$$

The normalized form of the ACVF is called the autocorrelation function (ACF) and is given as

$$C(\tau) = \lim_{L \to \infty} \frac{1}{L\sigma^2} \left[z(x) - m\right]\left[z(x+\tau) - m\right] dx = \left[R(\tau) - m^2\right] \big/ \sigma^2 \tag{2.23}$$

For a random function, C(τ) would be maximum (= 1) at τ = 0. If the signal is periodic, C(τ) would peak whenever τ is a multiple of wavelength. Many engineering surfaces are found to have an exponential ACF,

$$C(\tau) = \exp(-\tau/\beta) \tag{2.24}$$

The measure of how quickly the random event decays is called the correlation length. The correlation length is the length over which the autocorrelation function drops to a small fraction of its value at the origin, typically 10% of its original value. The exponential form has a correlation length of β* [C(τ) = 0.1] equal to 2.3 β (Figure 2.15). Sometimes, correlation length is defined as the distance at which value of the autocorrelation function is 1/e, that is 37%, which is equal to β for exponential ACF. The correlation length can be taken as that at which two points on a function have just reached the condition where they can be regarded as being independent. This follows from the fact that when C(τ) is close to unity, two points on the function at a distance τ apart are strongly interdependent. However, when C(τ) attains values close to zero, two points on the function at a distance τ apart are weakly correlated. The correlation length, β* can be viewed as a measure of randomness. The degree of randomness of a surface increases with an increase in the magnitude of β*.

The directionality of a surface can be found from its autocorrelation function. By plotting the contours of equal autocorrelation values, one can obtain contours to reveal surface structure. The anisotropy of

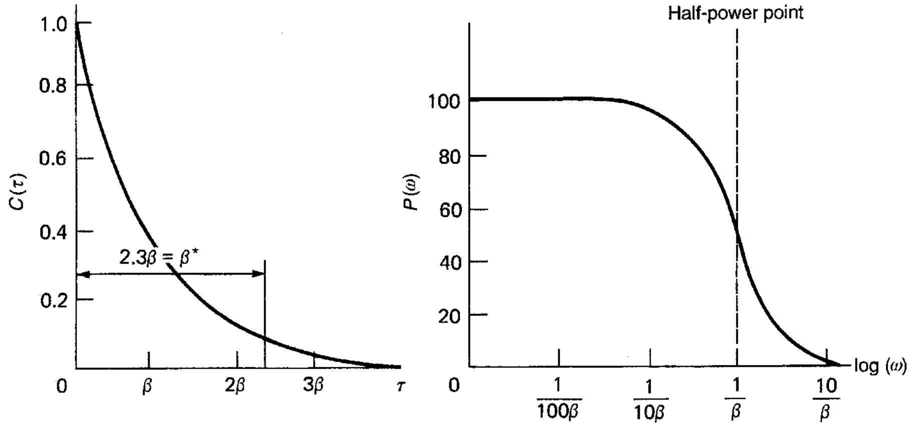

FIGURE 2.15 An exponential autocorrelation function and corresponding power spectral density function.

the surface structure is given as the ratio between the longer and shorter axes of the contour (Wyant et al., 1986; Bhushan, 1996). For a theoretically isotropic surface structure, the contour would have a constant radius; that is, it would be a circle.

The autocorrelation function can be calculated either by using the height distribution of the digitized profile or the fast Fourier transform (FFT) technique. In the FFT technique, the first PSDF (described later) is obtained by taking an FFT of the surface height and squaring the results; then an inverse FFT of the PSDF is taken to get ACVF.

Structure Function

The structure function (SF) or variance function (VF) in an integral form for a profile z(x) is,

$$S(\tau) = \lim_{L\to\infty} \frac{1}{L} \int_0^L \left[z(x) - z(x+\tau)\right]^2 dx \tag{2.25}$$

The function represents the mean square of the difference in height expected over any spatial distance τ. For stationary structures, it contains the same information as the ACVF. The two principal advantages of SF are that its construction is not limited to the stationary case, and it is independent of the mean plane.

Structure function is related to ACVF and ACF as

$$S(\tau) = 2\left[\sigma^2 + m^2 - R(\tau)\right] \tag{2.26a}$$

$$= 2\sigma^2\left[1 - C(\tau)\right] \tag{2.26b}$$

Power Spectral Density Function

The PSDF is another form of spatial representation and provides the same information as the ACVF or SF, but in a different form. The PSDF is the Fourier transform of the ACVF,

$$\begin{aligned} P(\omega) = P(-\omega) &= \int_{-\infty}^{\infty} R(\tau)\exp(-i\omega\tau)d\tau \\ &= \int_{-\infty}^{\infty} \sigma^2 C(\tau)\exp(-i\omega\tau)d\tau + m^2\delta(\omega) \end{aligned} \tag{2.27}$$

where ω is the angular frequency in length^{-1} (= 2πf or 2π/λ, f is frequency in cycles/length and λ is wavelength in length per cycle), and δ(ω) is the delta function. P(ω) is defined over all frequencies, both positive and negative, and is referred to as a two-sided spectrum. G(ω) is a spectrum defined over nonnegative frequencies only and is related to P(ω) for a random surface by

$$\begin{aligned} G(\omega) &= 2P(\omega),\ \omega \geq 0 \\ &= 0,\ \omega < 0. \end{aligned} \tag{2.28a}$$

Since the ACVF is an even function of τ, it follows that the PSDF is given by the real part of the Fourier transform in Equation 2.27. Therefore,

$$P(\omega) = \int_{-\infty}^{\infty} R(\tau)\cos(\omega\tau)d\tau = 2\int_{0}^{\infty} R(\tau)\cos(\omega\tau)d\tau \tag{2.28b}$$

Conversely, the ACVF is given by the inverse Fourier transform of the PSDF,

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)\exp(i\omega\tau)\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)\cos(\omega\tau)\,d\omega \tag{2.29}$$

For

$$\tau = 0,\ R(0) = R_q^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)\,d\omega \tag{2.30}$$

The equation shows that the total area under the PSDF curve (when frequency in cycles/length) is equal to R_q^2. The area under the curve between any frequency limits gives the mean square value of the data within that frequency range.

The PSDF can also be obtained directly in terms of the Fourier transform of the profile data z(x) by taking an FFT of the profile data and squaring the results, as follows:

$$P(\omega) = \lim_{L\to\infty}\frac{1}{L}\left[\int_{0}^{L} z(x)\exp(-i\omega x)\,dx\right]^2 \tag{2.31}$$

The PSDF can be evaluated from the data either via the ACVF using Equation 2.28 or the Fourier transform of the data (Equation 2.31). Note that the units of the one-dimensional PSDF are in terms of length to the third power, and for the two-dimensional case it is the length to the fourth power.

Figure 2.15 shows the PSDF for an exponential ACF previously presented in Equation 2.24. The magnitude of the P(ω) at ω = 1/β is known as the half-power point. For an exponential ACF, the PSDF is represented by white noise in the upper frequencies. The physical meaning of the model is that the main components of the function consist of a band covering the lower frequencies (longer wavelengths). Shorter wavelength components exist, but their magnitude declines with increasing frequency so that, in this range, the amplitude is proportional to wavelength. To cover large spatial range, it is often more convenient with surface data to represent ACF, SF, and PSDF on a log–log scale.

Figure 2.16a shows examples of selected profiles. Figures 2.16b and 2.16c show the corresponding ACVF and PSDF (Bendat and Piersol, 1986). (For calculation of ACVF and PSDF, profile length of multiple of wavelengths [a minimum of one wavelength] needs to be used.) The ACVF of a sine wave is a cosine wave. The envelope of the sine wave covariance function remains constant over all time delays,

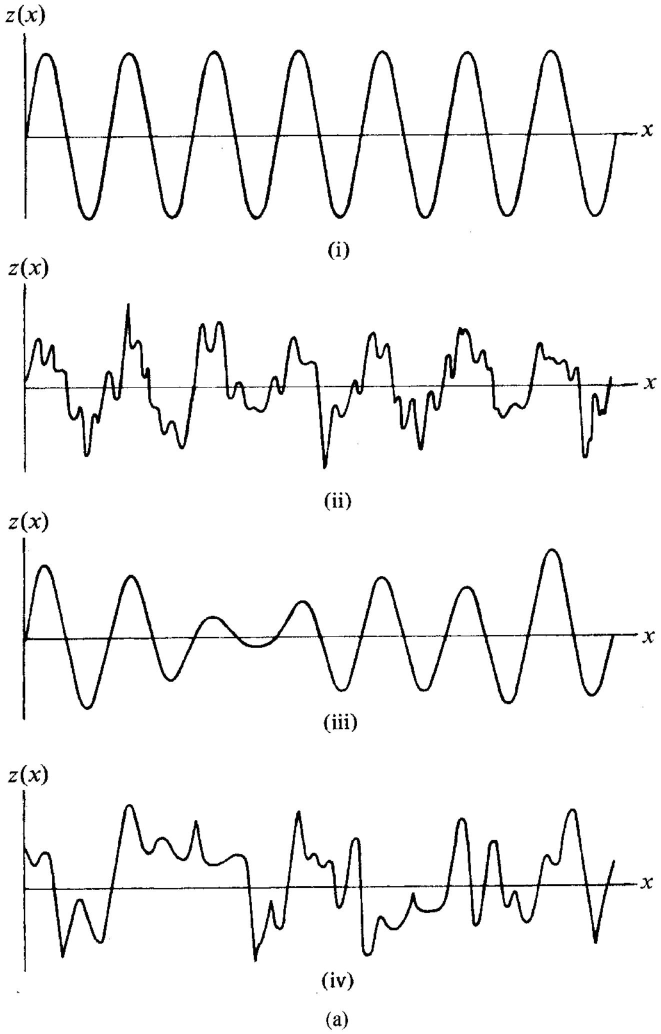

FIGURE 2.16 (a) Four special time histories: (i) sine wave, (ii) sine wave plus wide-band random noise, (iii) narrow-band random noise, and (iv) wide-band random noise. (b) corresponding idealized autocovariance functions, and (c) corresponding power spectral density functions (From Bendat, J.S. and Piersol, A.G. (1986), *Engineering Applications of Correlation and Spectral Analysis, 2nd edition,* Wiley, New York. With permission.)

suggesting that one can predict future values of the data precisely based on past observations. Looking at the PSDF of the sine wave, we note that the total mean square value of the sine wave is concentrated at the single frequency, ω_0. In all other cases, because of the erratic character of z(x) in Figure 2.16a, a past record does not significantly help one predict future values of the data beyond the very near future. To calculate the autocovariance function for (iii) to (iv) profiles, the power spectrum of the data is considered uniform over a wide bandwidth B. ACVF and PSDF of a sine wave plus wide-band random noise is simply the sum of the functions of the sine wave and wide-band random noise.

The moments of the PSDF are defined as

$$M_n = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[P(\omega) - m^2\delta(\omega)\right]\omega^n \, d\omega \tag{2.32}$$

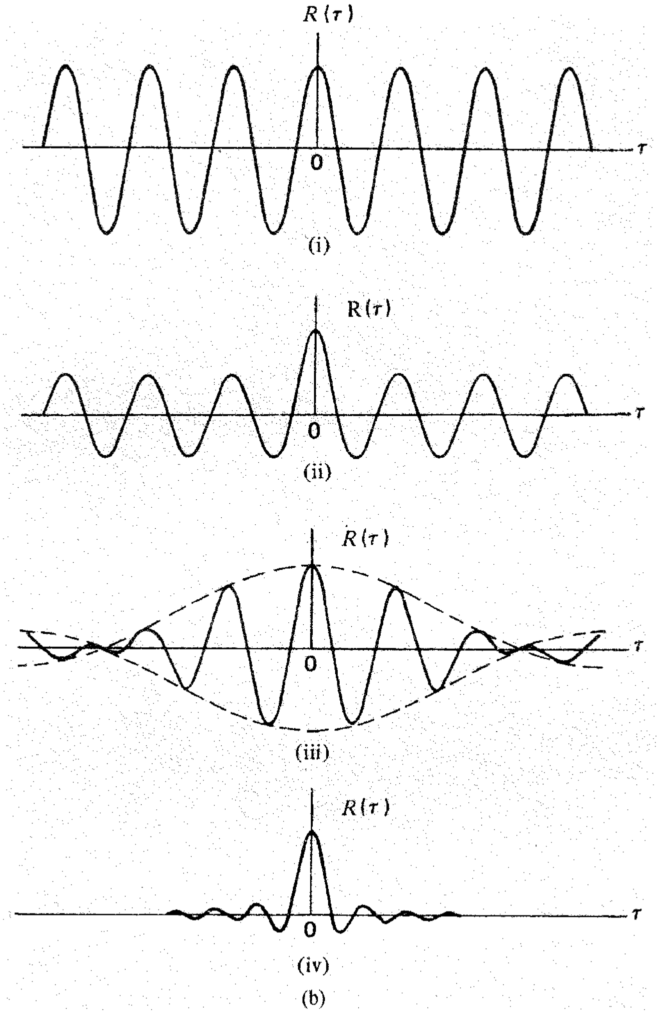

FIGURE 2.16 (**continued**)

where M_n are known as the spectral moments of the *n*th order. We note for a Gaussian function (Nayak, 1971),

$$M_0 = \sigma^2 = \frac{1}{L}\int_0^L (z-m)^2 \, dx \tag{2.33a}$$

$$M_2 = (\sigma')^2 = \frac{1}{L}\int_0^L (dz/dx)^2 \, dx \tag{2.33b}$$

and

$$M_4 = (\sigma'')^2 = \frac{1}{L}\int_0^L (d^2z/dx^2)^2 \, dx \tag{2.33c}$$

where σ′ and σ″ are the standard deviations of the first and second derivatives of the functions. For a surface/profile height, these are the surface/profile slope and curvature, respectively.

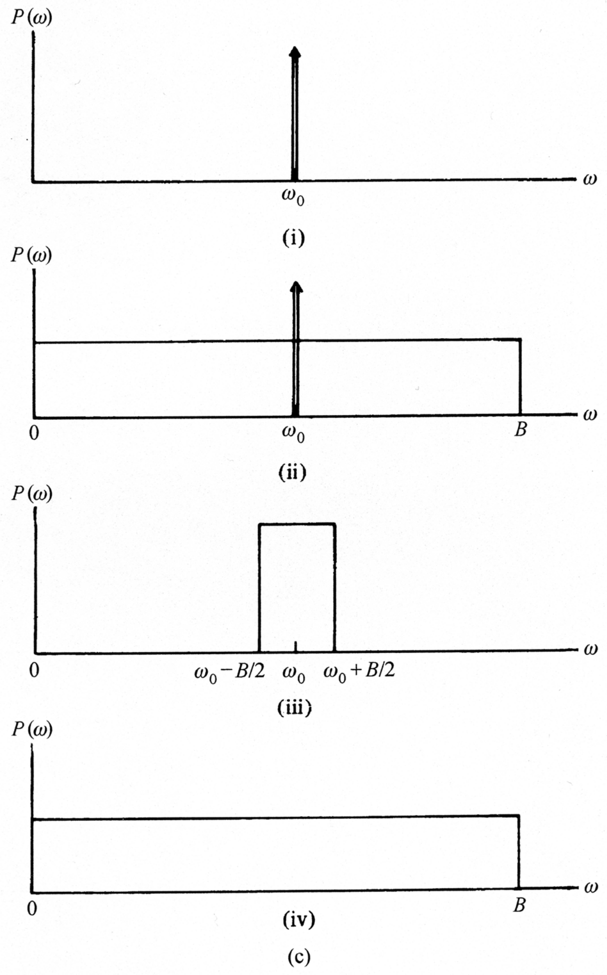

FIGURE 2.16 (**continued**)

According to Nayak (1971), a random and isotropic surface with a Gaussian height distribution can be adequately characterized by the three-zeroth (M_0), second (M_2) and fourth moments (M_4) of the power spectral density function. Based on the theory of random processes, a random and isotropic surface can be completely characterized in a statistical sense (rather than a deterministic sense) by two functions: the height distribution and the autocorrelation function. A random surface with Gaussian height distribution and exponential autocorrelation function can then simply be characterized by two parameters, two lengths: standard deviation of surface heights (σ) and the correlation distance (β^*) (Whitehouse and Archard, 1970). For characterization of a surface with a discrete, arbitrary autocorrelation function, three points C(0), C(h), and C(2h) for a profile, where h is an arbitrary distance and four or more points are needed on the C(τ), depending upon the type of the surface (Whitehouse and Phillips, 1978, 1982).

2.2.2.6 Probability Distribution and Statistics of the Asperities and Valleys

Surfaces consist of hills (asperities) of varying heights and spacing and valleys of varying depths and spacing. For a two-dimensional profile, the peak is defined as a point higher than its two adjacent points greater than a threshold value. For a three-dimensional surface map, the summit is defined as a point higher than its four adjacent points greater than a threshold value. A valley is defined in the same way but in reverse order. A threshold value is introduced to reduce the effect of noise in the measured data and ensure that every peak/summit identified is truly substantial. Based on analysis of roughness data of a variety of smooth samples, Poon and Bhushan (1995b) recommend a threshold value as one tenth of the σ roughness of smooth surfaces (with σ less than about 50 nm); it should be lower than 10% of the σ value for rougher surfaces.

Gaussian surfaces might be considered as comprising a certain number of hills (asperities) and an equal number of valleys. These features may be assessed and represented by their appropriate distribution curves, which can be described by the same sort of characteristics as were used previously for the surface height distributions. Similar to surface height distributions, the height distributions of peaks (or summits) and valleys often follow the Gaussian curve (Greenwood, 1984; Wyant et al., 1986; Bhushan, 1996). Distribution curves can also be obtained for the absolute values of slope and for the curvature of the peaks (or summits) and valleys. Distributions of peak (or summit) curvature follow a log normal distribution (Gupta and Cook 1972; Wyant et al., 1986; Bhushan, 1996). The mean of the peak curvature increases with the peak height for a given surface (Nayak, 1971).

The parameters of interest in some analytical contact models of two random rough surfaces to be discussed in the next chapter are the density of summits (η), the standard deviation of summit heights (σ_p), and the mean radius (R_p) (or curvature, κ_p) of the summit caps or η, σ, and β*. The former three roughness parameters (η, σ_p, R_p) can be related to other easily measurable roughness parameters using the theories of Longuet-Higgins (1957a,b), Nayak (1971), and Whitehouse and Phillips (1978, 1982).

For a random Gaussian profile, we note that $M_2(= \sigma'^2)$ and M_r $(= \sigma''^2)$ can be calculated from σ, N_0, and N_p using the following relationship (Longuet-Higgins, 1957a):

$$N_0 = \frac{\sigma'}{\pi\sigma} \tag{2.34a}$$

and

$$N_p = \frac{\sigma''}{2\pi\sigma'} \tag{2.34b}$$

Note that N_0 and N_p are frequency dependent, for example, if the profile has a high frequency riding on a low frequency, then N_p will be higher and N_0 will be lower.

We now define an auxiliary quantity, bandwidth parameter, α (Nayak, 1971),

$$\alpha = \left(\frac{2N_p}{N_0}\right)^2 = \left(\frac{\sigma\sigma''}{\sigma'^2}\right)^2 \tag{2.35}$$

which defines the width of the power spectrum of the random process forming the process from which the profile is taken. The distribution of peak heights and their expected tip curvature as a function of α are shown in Figure 2.17 (Nayak, 1971). z^*_p $(= z_p/\sigma)$ is the standardized cumulative summit/peak height distribution and κ^*_p $(= \kappa_p/\sigma'')$ is the standardized mean summit/peak curvature κ_p. It is observed that high peaks always have a longer expected mean curvature (i.e., a smaller mean radius) than lower peaks. If α = 1, the spectrum consists of a single frequency, where $2N_p$, total density of peaks and valleys (maxima and minima) equals N_0. If α = ∞, the spectrum extends over all frequencies (white noise spectrum). The peak heights have a Gaussian distribution and κ_p is nearly constant for peaks of all heights and is given by

$$\sigma_p \sim \sigma \tag{2.36a}$$

and

$$\kappa_p \sim 1.3\,\sigma'' \tag{2.36b}$$

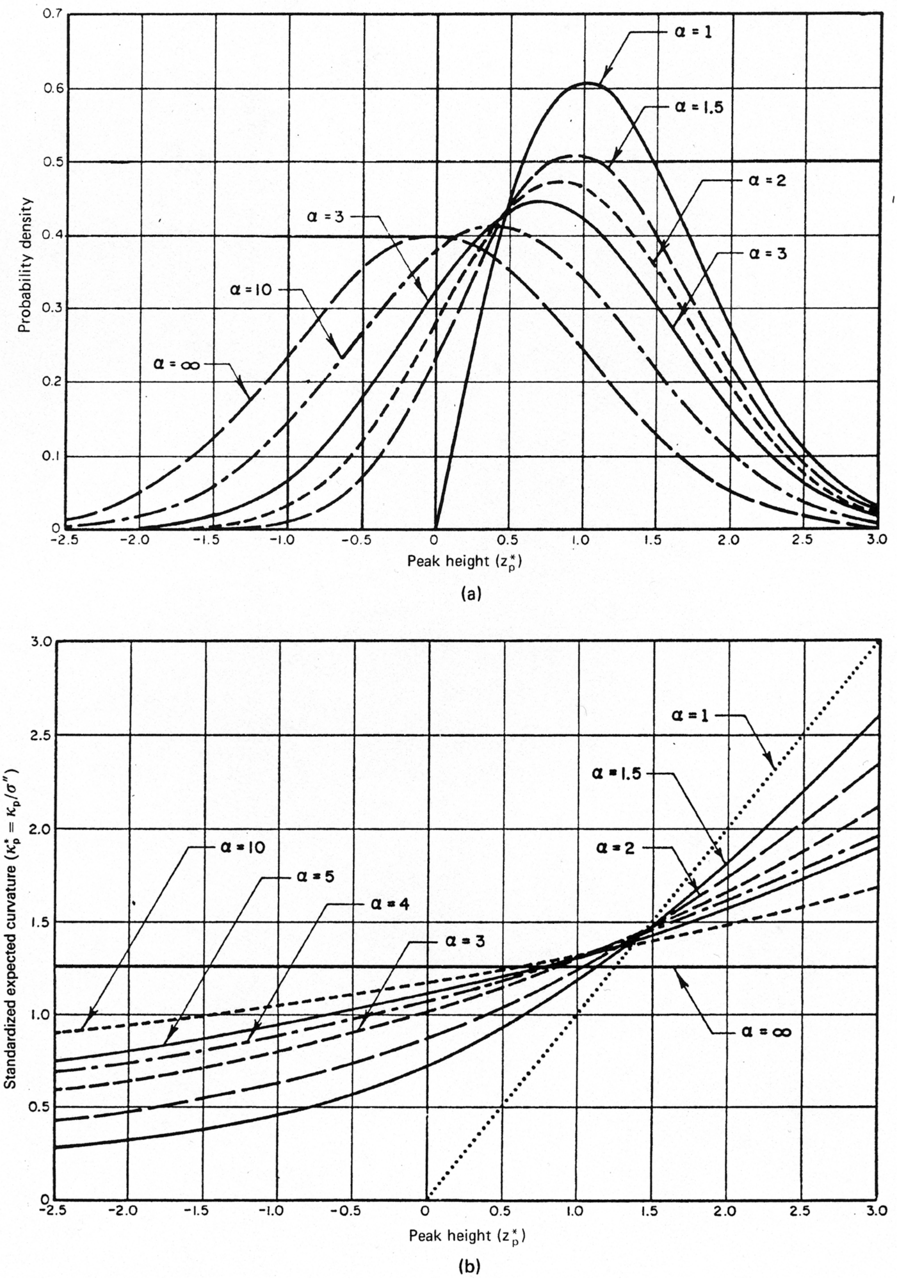

FIGURE 2.17 (a) Probability density of peak heights and (b) expected dimensionless curvature of peaks. (From Nayak, P.R. (1971), Random process model of rough surfaces, *ASME J. Lub. Tech.*, 93, 398-407. With permission.)

We also note that a profile with a very narrow spectrum has no peaks below a mean line, whereas a white noise profile with infinite spectral width has half of its peaks below this line.

From Bush et al. (1976), the standard deviation of the summit/peak height for any α is given by

$$\sigma_p \sim \left(1 - \frac{0.8968}{\alpha}\right)^{1/2} \sigma \tag{2.37}$$

From Nayak (1971), the density of summits per unit area can be related to the spectral moments and number of peaks per unit length by

$$\eta = 0.031\left(\frac{\sigma''}{\sigma'}\right)^2 \tag{2.38a}$$

$$\sim 1.2\,N_p^2 \tag{2.38b}$$

Using discrete random process analysis, Whitehouse and Phillips (1978, 1982) derived the relationship of tribological parameters of interest for a surface that has a Gaussian height distribution. For characterization of a profile, we need standard deviation and just two points on the measured ρ_1 and ρ_2 spaced h (sampling interval) and 2h from the origin of the normalized autocorrelation function. For characterization of a surface, we need between four and seven points on the ACF, depending upon the type of surface. Tribological parameters that can be predicted are the mean and standard deviation (σ_p) of the peak height; the mean (κ_p) and standard deviation (σ_p'') of the peak curvature; the average peak slope; the correlation coefficient between the peak height and its curvature; and the summit density.

2.2.2.7 Composite Roughness of Two Random Rough Surfaces

For two random rough surfaces in contact, the composite roughness of interest is defined as the sum of two roughness processes obtained by adding together the local heights (z), the local slope (θ), and local curvature (κ)

$$\begin{aligned} z &= z_1 + z_2 \\ \theta &= \theta_1 + \theta_2 \\ \kappa &= \kappa_1 + \kappa_2 \end{aligned} \tag{2.39}$$

For two random rough surfaces in contact, an equivalent rough surface can be described of which the values of σ, σ', σ'', $R(\tau)$, $P(\omega)$, and M_0, M_2, and M_4 are summed for the two rough surfaces, that is,

$$\begin{aligned} \sigma^2 &= \sigma_1^2 + \sigma_2^2 \\ \sigma'^2 &= \sigma_1'^2 + \sigma_2'^2 \\ \sigma''^2 &= \sigma_1''^2 + \sigma_2''^2 \\ R(\tau) &= R_1(\tau) + R_2(\tau) \\ P(\omega) &= P_1(\omega) + P_2(\omega) \\ \text{and} \quad M_i &= (M_i)_1 + (M_i)_2 \end{aligned} \tag{2.40a}$$

where *i* = 0, 2, 4. These equations state that variances, autocovariance function, and power spectra are simply additive. Since autocovariance functions of two functions are additive, simple geometry shows that correlation lengths of two exponential ACVFs are related as

$$\frac{1}{\beta^*} = \frac{1}{\beta_1^*} + \frac{1}{\beta_2^*} \tag{2.40b}$$

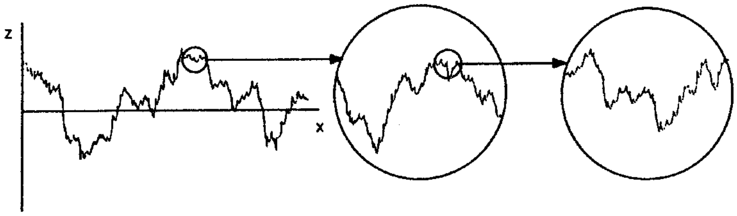

FIGURE 2.18 Qualitative description of statistical self-affinity for a surface profile.

2.2.3 Fractal Characterization

A surface is composed of a large number of length scales of roughness that are superimposed on each other. As stated earlier, surface roughness is generally characterized by the standard deviation of surface heights. However, due to the multiscale nature of the surface, it is known that the variances of surface height and its derivatives and other roughness parameters depend strongly on the resolution of the roughness measuring instrument or any other form of filter; hence they are not unique for a surface (Ganti and Bhushan, 1995; Poon and Bhushan, 1995a). Therefore rough surfaces should be characterized in such a way that the structural information of roughness at all scales is retained. It is necessary to quantify the multiscale nature of surface roughness.

A unique property of rough surfaces is that if a surface is repeatedly magnified, increasing details of roughness are observed right down to nanoscale. In addition, the roughnesses at all magnifications appear quite similar in structure, as is qualitatively shown in Figure 2.18. The statistical self-affinity is due to similarity in appearance of a profile under different magnifications. Such a behavior can be characterized by fractal geometry (Majumdar and Bhushan, 1990; Ganti and Bhushan, 1995; Bhushan, 1999b). The fractal approach has the ability to characterize surface roughness by scale-independent parameters and provides information on the roughness structure at all length scales that exhibit the fractal behavior. Surface characteristics can be predicted at all length scales within the fractal regime by making measurements at one scan length.

Structure function and power spectrum of a self-affine fractal surface follow a power law and can be written as (Ganti and Bhushan model)

$$S(\tau) = C\eta^{(2D-3)}\tau^{(4-2D)} \tag{2.41}$$

$$P(\omega) = \frac{c_1\eta^{(2D-3)}}{\omega^{(5-2D)}} \tag{2.42a}$$

and

$$c_1 = \frac{\Gamma(5-2D)\sin[\pi(2-D)]}{2\pi}C \tag{2.42b}$$

The fractal analysis allows the characterization of surface roughness by two parameters D and C which are instrument-independent and unique for each surface. The parameter D (ranging from 1 to 2 for surface profile) primarily relates to the relative power of the frequency contents, and C to the amplitude of all frequencies. η is the lateral resolution of the measuring instrument, τ is the size of the increment (distance), and ω is the frequency of the roughness. Note that if $S(\tau)$ or $P(\omega)$ are plotted as a function of ω or τ, respectively, on a log-log plot, then the power law behavior results in a straight line. The slope of the line is related to D, and the location of the spectrum along the power axis is related to C.

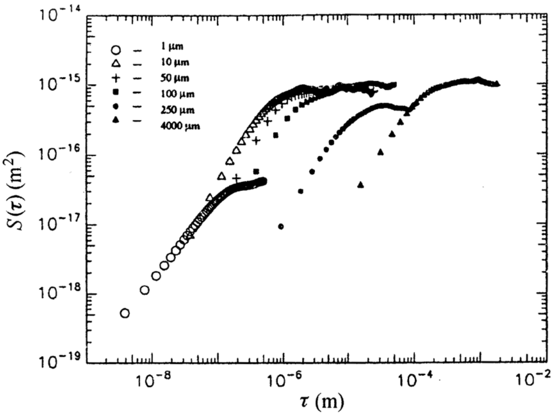

FIGURE 2.19 Structure functions for the roughness data measured using AFM and NOP, for a thin-film magnetic rigid disk. (From Ganti, S. and Bhushan, B. (1995), Generalized fractal analysis and its applications to engineering surfaces, *Wear,* 180, 17-34. With permission.)

TABLE 2.3 Surface Roughness Parameters for a Polished Thin-Film Rigid Disk

Scan Size (μm × μm)	σ(nm)	D	C(nm)
1(AFM)	0.7	1.33	9.8×10^{-4}
10(AFM)	2.1	1.31	7.6×10^{-3}
50(AFM)	4.8	1.26	1.7×10^{-2}
100(AFM)	5.6	1.30	1.4×10^{-2}
250(NOP)	2.4	1.32	2.7×10^{-4}
4000(NOP)	3.7	1.29	7.9×10^{-5}

AFM — Atomic force microscope
NOP — Noncontact optical profiler

Figure 2.19 presents the structure functions of a thin-film magnetic rigid disk measured using an atomic force microscope (AFM) and noncontact optical profiler (NOP). A horizontal shift in the structure functions from one scan to another arises from the change in the lateral resolution. The D and C values for various scan lengths are listed in Table 2.3. Note that fractal dimension of the various scans is fairly constant (1.26 to 1.33); however, C increases/decreases monotonically with σ for the AFM data. The error in estimation of η is believed to be responsible for the variation in C. These data show that the disk surface follows a fractal structure for three decades of length scales.

2.2.4 Practical Considerations in Measurement of Roughness Parameters

2.2.4.1 Short- and Long-Wavelength Filtering

Engineering surfaces cover a broad bandwidth of wavelengths, and samples, however large, often exhibit nonstationary properties (in which the roughness is dependent upon the sample size). Surface roughness is intrinsic; however, measured roughness is a function of the bandwidth of the measurement and thus is not an intrinsic property. Instruments using different sampling intervals measure features with different length scales. Roughness is found at scales ranging from millimeter to nanometer (atomic) scales. A surface is composed of a large number of length scales of roughness that are superimposed on each other. Therefore, on a surface, it is not that different asperities come in different sizes but that one asperity comes in different sizes. Distribution of size and shape of asperities is dependent on the short-wavelength limit or the sampling interval of the measuring instrument. When the sampling interval at which the surface is sampled is reduced, the number of asperities detected and their curvature appear to rise without

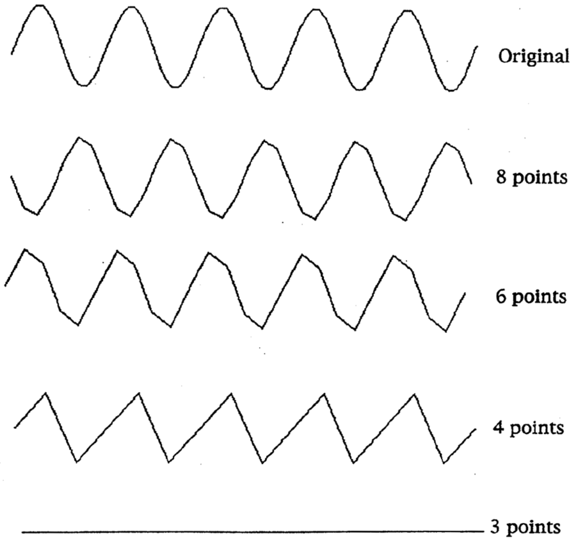

FIGURE 2.20 Sinusoidal profiles with different numbers of sampling points per wavelength.

limit down to atomic scales. This means that asperity is not a "definite object." Attempts are made to identify a correct sampling interval which yields the relevant number of asperities for a particular application. An asperity relevant for contact mechanics is defined as that which makes a contact in a particular application (contacting asperity) and carries some load.

The short-wavelength limit or the sampling interval affects asperity statistics. The choice of short-wavelength limit depends on the answer to the following question: what is the shortest wavelength that will affect the interaction? It is now known that it is the asperities on a nanoscale that first come into contact and plastically deform instantly, and subsequently, the load is supported by the deformation of larger-scale asperities (Bhushan and Blackman, 1991; Poon and Bhushan, 1996). Since plastic deformation in most applications is undesirable, asperities on a nanoscale need to be detected. Therefore, the short-wavelength limit should be as small as possible.

The effect of the short-wavelength limit on a roughness profile can be illustrated by a sinusoidal profile represented by different numbers of sampling points per wavelength as shown in Figure 2.20. The waveform of the sinusoidal profile is distorted when the number of sampling points decreases. The profile parameters do not change significantly with sampling points equal to 6 or greater per wavelength. Therefore, the minimum number of sampling points required to represent a wavelength structure may be set to 6, i.e., the optimum sampling interval is $\lambda/6$, where λ is the wavelength of the sinusoidal profile. By analogy, the suitable sampling interval should be related to the main wavelength structure of a random profile which is represented by β^*. However, β^* is a function of the bandwidth of the measurement and thus is not an intrinsic property. It is reasonable to select a sampling interval a fraction of β^* measured at the long wavelength limit, say 0.25 β^* to 0.5 β^* (Poon and Bhushan, 1995a).

Figure 2.21 demonstrates how the long wavelength limit, also called the cutoff wavelength or sampling length (size), can affect the measured roughness parameters (Anonymous, 1985). The top profile represents the actual movement of the stylus on a surface. The lower ones show the same profile using cutoff wavelength values of 0.8, 0.25, and 0.08 mm. A small cutoff value would isolate the waviness while a large cutoff value would include the waviness. Thomas (1999) has shown that the standard deviation of surface roughness, σ, will increase with an increase in the cutoff wavelength or sampling length L, as given by the following relation,

$$\sigma \propto L^{1/2} \tag{2.43}$$

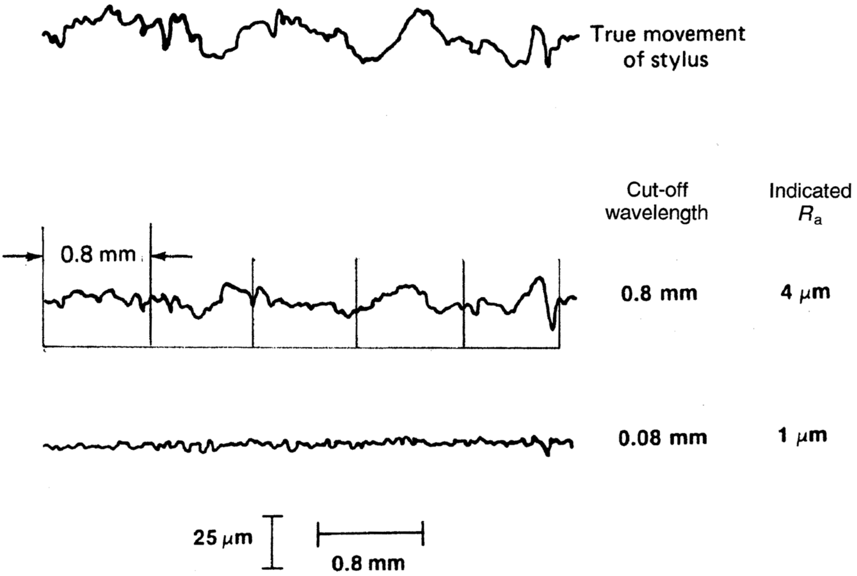

FIGURE 2.21 The effect of the cutoff wavelength is to remove all components of the total profile that have wavelengths greater than cutoff value.

Ganti and Bhushan (1995) and Poon and Bhushan (1995a) have reported that σ and other roughness parameters initially increase with L and then reach a constant value because engineering surfaces seem to have a long-wavelength limit. Thus, before the surface roughness can be effectively quantified, an application must be defined. Having a knowledge of the application enables a measurement to be planned and in particular for it to be decided to what bandwidth of surface features the information collected should refer. Features that appear as roughness in one application of a surface may well constitute waviness in another.

The long-wavelength limit (which is the same as scan size in many instruments) in contact problems is set by the dimensions of the nominal contact area (Figure 2.22). This is simply to say that a wavelength much longer than the nominal contact area will not affect what goes on inside it. In addition, the long-wavelength limit of the surface roughness in the nominal contact area, if it exists, should be obtained. The long-wavelength limit can be chosen to be twice the nominal contact size or the long-wavelength limit of the roughness structure in the nominal contact size, if it exists, whichever is smaller.

To provide a basis of instrumentation for roughness measurement, a series of cutoff wavelength values has been standardized in a British standard (BS1134-1972), an ANSI/ASME (B46.1-1985), and an ISO Recommendation (R468). The international standard cutoff values are 0.08, 0.25, and 0.8 mm. The preferred value of 0.8 mm is assumed unless a different value is specified. Note that waviness measurements are made without long-wavelength filtering.

Long- and short-wavelength filtering in measuring instruments is most commonly accomplished by digital filtering. For example, in a fast Fourier transform (FFT) technique, the FFT of the raw data is taken,

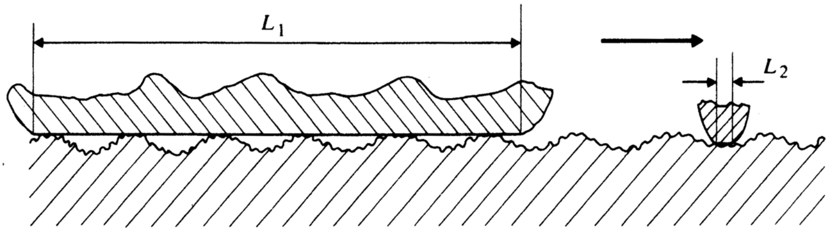

FIGURE 2.22 Contact size of two moving components of different lengths L_1 and L_2 on the same rough surface.

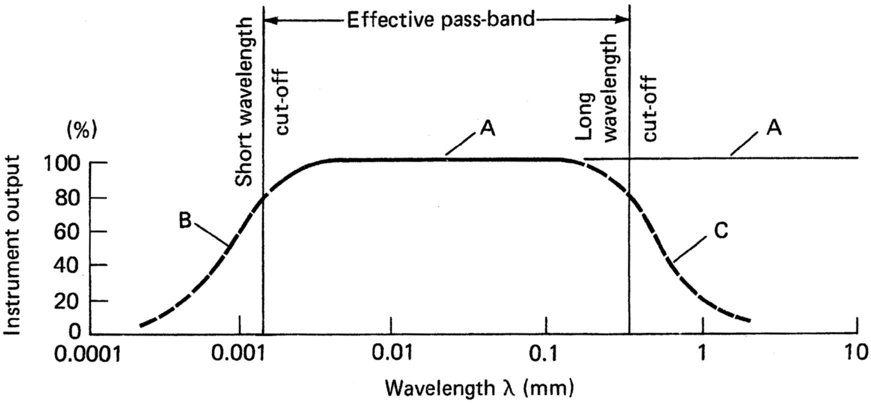

FIGURE 2.23 Transmission characteristics of a profiler with low bandpass and high bandpass filters.

the appropriate frequency contents are removed, and the inverse FFT is taken to obtain the filtered data. However, this technique is slow, and one method commercially used is the finite impulse response (FIR) technique. The FIR technique filters by convoluting the trace data with the impulse response of the filter specification. The impulse response is obtained by taking the inverse FFT of the filter specification.

Anonymous (1985) also describes the electronic filtering method for short- and long-wavelength filtering, which is accomplished by passing the alternating voltage representing the profile through an electrical wave filter, such as the standard RC filter. The electronic filtering is generally used to filter out short wavelength electronic noise (low band pass filtering).

In some profilers (Talysurf by Rank Taylor Hobson, Leicester, England), a skid on a pickup body is traversed along with the stylus arm (Bhushan, 1996). The skid provides a straight reference datum and provides a long-wavelength filtering which is a function of the size and shape of the skid.

Mechanical short-wavelength filtering also results from the design and construction of a measuring instrument. For example in the stylus instrument or the atomic force microscope, the stylus removes certain short wavelengths on the order of the stylus tip radius, which is referred to as lateral resolution of the instrument. The stylus is not able to enter the grooves. As the spacing between grooves increases, the stylus displacement will rise, but once it has become sufficient for the stylus to reach the bottom, there will be a full indication. In a digital optical profiler, lateral resolution is controlled by the physical size of the charge-coupled-device (CCD) image sensors at the microscope objective magnifications. A short wavelength limit, if selected, should be at least twice the lateral resolution of the instrument.

For the instrument in which a short-wavelength filter is introduced, the output will tend to fall off above a certain frequency, that is below a certain wavelength, for example, as shown by the dotted curve B in Figure 2.23, even though the stylus continues to rise and fall over the irregularities. Dotted curve C in Figure 2.23 also shows the fall off of instrument output at longer wavelength. Only within the range of wavelengths for which the curve is substantially level will the indication be a measure solely of the amplitude and be independent of wavelength curve A in Figure 2.23.

2.2.4.2 Scan Size

After the short-wavelength and long-wavelength limits are selected, the roughness measurement must be made on a length large enough to provide a statistically significant value for the chosen locality. The total length involved is called the measuring length, evaluation length, traversing length, or scan length. In some cases, a length of several individual scan lengths (say five) are chosen (Whitehouse, 1994). In most measurements, scan length is the same as the long-wavelength limit. For two-dimensional measurement, a certain area is measured rather than a length.

Wyant et al. (1984) and Bhushan et al. (1985) have suggested that in measurement of a random surface, a scan length equal to or greater than 200 β* should be used.

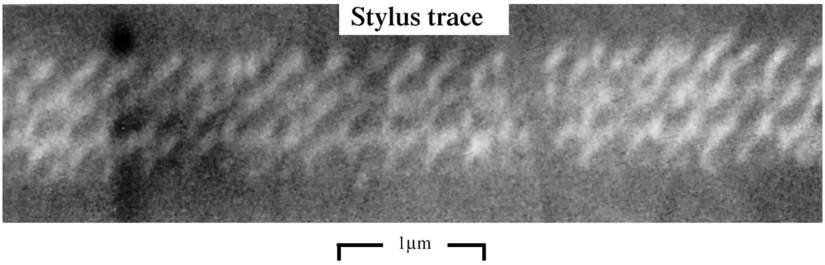

FIGURE 2.24 SEM micrograph of a trace made by a stylus instrument showing surface damage of electroless Ni–P coating (stylus material, diamond; stylus radius = 0.1 μm; and stylus load = 10 μN or 1 mg). (From Poon, C.Y. and Bhushan, B. (1995a), Comparison of surface roughness measurements by stylus profiler, AFM and non-contact optical profiler, *Wear,* 190, 76-88. With permission.)

2.3 Measurement of Surface Roughness

A distinction is made between methods of evaluating the nanoscale to atomic scale and microscale features of surface roughness. Physicists and physical chemists require fine-scale details of surfaces and often details of molecular roughness. These details are usually provided using methods such as low-energy electron diffraction, molecular-beam methods, field-emission and field-ion microscopy, scanning tunneling microscopy, and atomic force microscopy. On the other hand, for most engineering and manufacturing surfaces, microscopic methods suffice, and they are generally mechanical or optical methods. Some of these methods can also be used to measure geometrical parameters of surfaces (Bhushan, 1996, 1999).

Various instruments are available for the roughness measurement. The measurement technique can be divided into two broad categories: (a) a contact type in which during measurement a component of the measurement instrument actually contacts the surface to be measured; and (2) a noncontact type. A contact-type instrument may damage surfaces when used with a sharp stylus tip, particularly soft surfaces (Figure 2.24). For these measurements, the normal loads have to be low enough so that the contact stresses do not exceed the hardness of the surface to be measured.

The first practical stylus instruments were developed by Abbott and Firestone (1933). In 1939, Rank Taylor Hobson in Leicester, England, introduced the first commercial instrument called Talysurf. Today, contact-type stylus instruments using electronic amplification are the most popular. The stylus technique, recommended by the ISO, is generally used for reference purposes. In 1983, a noncontact optical profiler based on the principle of two-beam optical interferometry was developed and is now widely used in the electronics and optical industries to measure smooth surfaces. In 1985, an atomic force microscope was developed which is basically a nano-profiler operating at ultra-low loads. It can be used to measure surface roughness with lateral resolution ranging from microscopic to atomic scales. This instrument is commonly used in research to measure roughness with extremely high lateral resolution, particularly nanoscale roughness.

There exists a number of other techniques that have been either demonstrated in the laboratory and never commercially used or used in specialized applications. We will divide the different techniques into six categories based on the physical principle involved: mechanical stylus method, optical methods, scanning probe microscopy (SPM) methods, fluid methods, electrical method, and electron microscopy methods. Descriptions of these methods are presented, and the detailed descriptions of only three — stylus, optical (based on optical interferometry), and AFM techniques — are provided. We will conclude this section by comparing various measurement methods.

2.3.1 Mechanical Stylus Method

This method uses an instrument that amplifies and records the vertical motions of a stylus displaced at a constant speed by the surface to be measured. Following is a partial list of commercial profilers: Rank

Taylor Hobson (UK) Talysurf profilers, Tencor Instruments AlphaStep and P-series profilers, Veeco/Sloan Technology Dektak profilers, Gould Inc. Instruments Division profilers, and Kosaka Laboratory, Tokyo (Japan) profilers. The stylus is mechanically coupled mostly to a linear variable differential transformer (LVDT), an optical or a capacitance sensor. The stylus arm is loaded against the sample and either the stylus is scanned across the stationary sample surface using a traverse unit at a constant speed or the sample is transported across an optical flat reference. As the stylus or sample moves, the stylus rides over the sample surface detecting surface deviations by the transducer. It produces an analog signal corresponding to the vertical stylus movement. This signal is then amplified, conditioned, and digitized (Thomas, 1999; Bhushan, 1996).

In a profiler, as shown in Figure 2.25a, the instrument consists of a stylus measurement head with a stylus tip and a scan mechanism (Anonymous, 1996a). The measurement head houses a stylus arm with a stylus, sensor assembly, and the loading system. The stylus arm is coupled to the core of an LVDT to monitor vertical motions. The core of a force solenoid is coupled to the stylus arm and its coil is energized to load the stylus tip against the sample. A proximity probe (photo optical sensor) is used to provide a soft limit to the vertical location of the stylus with respect to the sample. The sample is scanned under the stylus at a constant speed.

In high precision, ultra-low load profilers, shown in Figures 2.25b and 2.25c, the vertical motion is sensed using a capacitance sensor, and a precision stage transports the sample during measurements (Anonymous, 1996b). The hardware consists of two main components: a stylus measurement head with stylus tip and a scan mechanism. The stylus measurement head houses a sensor assembly, which includes the stylus, the appropriate sensor electronics and integrated optics, and the loading system. The capacitance sensor exhibits a lower noise, has a lower mass, and scales well to smaller dimensions as compared to LVDTs. The capacitive sensor assembly consists of a stylus arm suspended by a flexure pivot, connected to a sensor vane that extends through the center of a highly sensitive capacitive sensor. Vertical movement of the stylus arm results in movement of the vane, which is registered by the capacitance sensor. The analog signal of the capacitance sensor output is digitized and displayed in a surface roughness map. The entire stylus assembly is mounted on a plate, which is driven by a motor for coarse vertical motion.

In order to track the stylus across the surface, force is applied to the stylus. The ability to accurately apply and control this force is critical to the profiler performance. The measurement head uses a wire coil to set a programmable stylus load as low as 0.05 mg. Attached above the stylus flexure pivot is an arm with a magnet mounted to the end. The magnet is held in close proximity to the wire coil, and the coil, when energized, produces a magnetic field that moves the magnet arm. This applied force pushes the stylus arm past its null position to a calibrated force displacement, where the horizontal position of the stylus arm represents zero applied force to the stylus. The force coil mechanism and a sophisticated digital signal processor are used to maintain a constant applied force to the stylus. The flexure pivot allows the stylus to move easily, but its tension affects the applied force to the stylus as the stylus arm is moved through its vertical range. To locally correct for the pivot tensions during roughness measurement for a constant stylus force, the force is calibrated by serving the drive current to the force coil to move the stylus several regularly spaced positions, with the stylus not in contact with a sample (zero stylus force). A table of stylus position vs. current settings is generated. A digital signal processor uses these data to dynamically change the force setting as the roughness measurements are made.

The scan mechanism shown in Figure 2.25c, holds the sensor assembly stationary while the sample stage is moved with a precision lead screw drive mechanism. This drive mechanism, called the X drive, uses a motor to drive the lead screw, which then moves the sample stage with guide wires along an optical flat via PTFE skids. The motion is monitored by an optical encoder and is accurate to 1 to 2 μm. The optical flat ensures a smooth and stable movement of the stage across the scan length, while a guide bar provides a straight, directional movement. This scanning of the sample limits the measurement noise from the instrument, by decoupling the stage motion from vertical motions of the stylus measured using the sensor. Surface topography measurements can be acquired with high sensitivity over a 205-mm scan. Three-dimensional images can be obtained by acquiring two-dimensional scans in the X direction while

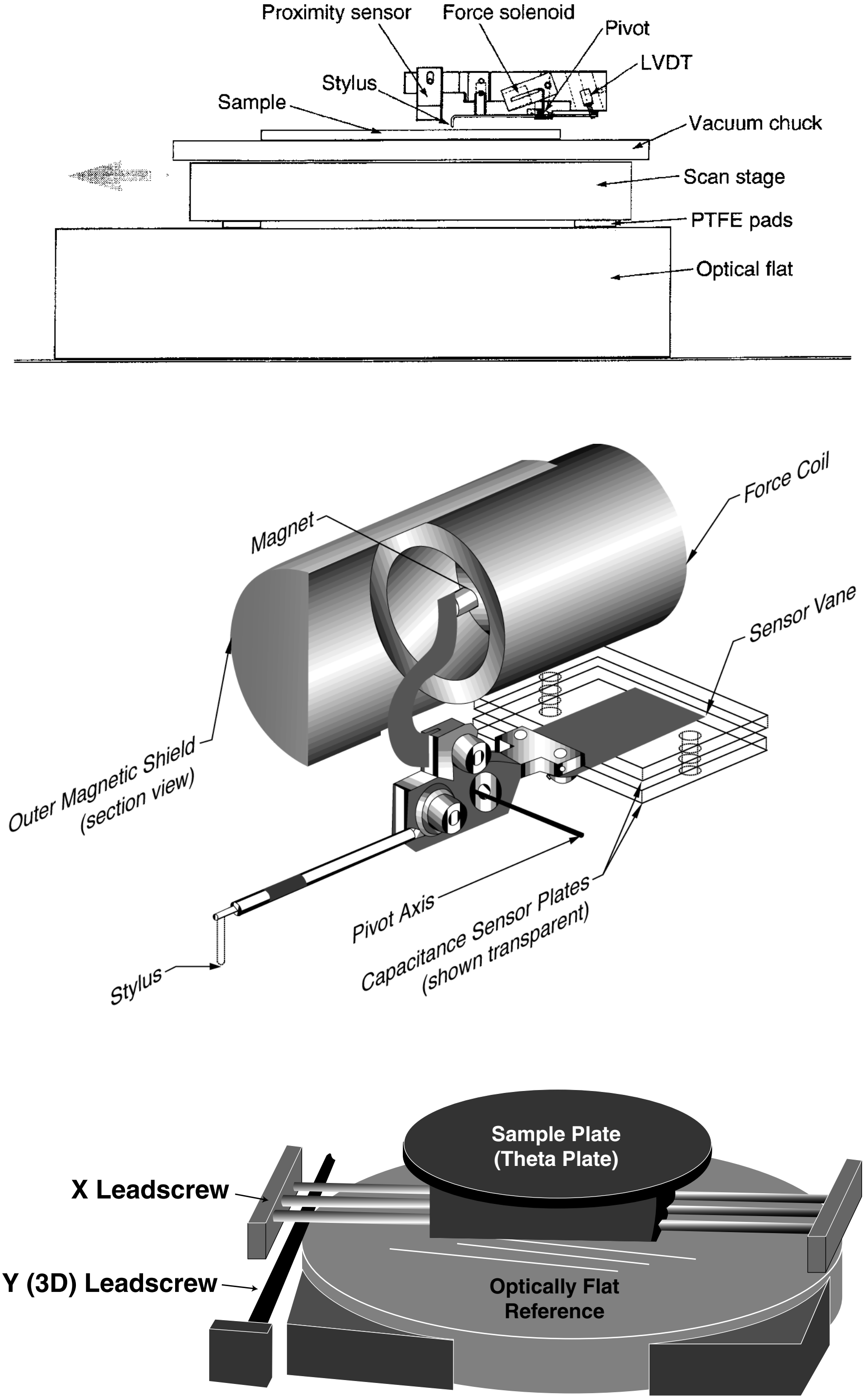

FIGURE 2.25 Schematics of (a) stylus measurement head with loading system and scan mechanism used in Veeco/Sloan Dektak profilers. (Courtesy of Veeco/Sloan Technology, Santa Barbara, CA.) (b) Stylus measurement head with loading system and (c) scan mechanism used in Tencor P-series profilers. (Courtesy of Tencor Instruments, Milpitas, CA.)

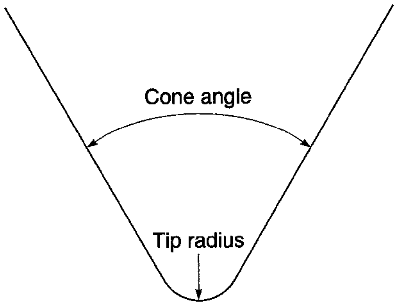

FIGURE 2.26 Schematic of a diamond conical stylus showing its cone angle and tip radius.

stepping in the Y direction by 5 μm with the Y lead screw used for precise sample positioning. When building a surface map by parallel traversing, it is essential to maintain a common origin for each profile. This can be achieved by a flattening procedure in which the mean of each profile is calculated separately and these results are spliced together to produce an accurate surface map. The surface maps are generally presented such that the vertical axis is magnified by three to four orders of magnitude as compared to the horizontal scan axis.

Measurements on circular surfaces with long scan lengths can be performed by a modified stylus profiler (such as Talyround) in which a cylindrical surface is rotated about an axis during measurement.

Styli are made of diamond. The shapes can vary from one manufacturer to another. Chisel-point styli with tips (e.g., 0.25 μm × 2.5 μm) may be used for detection of bumps or other special applications. Conical tips are almost exclusively used for microroughness measurements (Figure 2.26). According to the international standard (ISO 3274–1975), a stylus is a cone of a 60° to 90° included angle and a (spherical) tip radius of curvature of 2, 5, or 10 μm. The radius of a stylus ranges typically from 0.1 to 0.2 μm to 25 μm with the included angle ranging from 60° to 80°. The stylus is a diamond chip tip that is braised to a stainless steel rod mounted to a stylus arm. The diamond chip is cleaved, then ground and polished to a specific dimension. The radius of curvature for the submicrometer stylus tip, which is assumed to be spherical, is measured with an SEM, or against a standard. The portion of the stylus tip that is in contact with the sample surface, along with the known radius of curvature, determines the actual radius of the tip with regard to the feature size. The stylus cone angle is determined from the cleave and grind of the diamond chip and is checked optically or against a standard.

Maximum vertical and spatial (horizontal) magnifications that can be used are on the order of 100,000× and 100×, respectively. The vertical resolution is limited by sensor response, background mechanical vibrations, and thermal noise in the electronics. Resolution for smooth surfaces is as low as 0.1 nm and 1 nm for rough surfaces for large steps. Lateral resolution is on the order of the square root of the stylus radius. The step height repeatability is about 0.8 nm for a step height of 1 μm. The stylus load ranges typically from 0.05 to 100 mg. Long-wave cutoff wavelengths range typically from 4.5 μm to 25 mm. Short-wave cutoff wavelengths range typically from 0.25 μm to several millimeters. The scan lengths can be typically as high as 200 mm, and for three-dimensional imaging, the scan areas can be as large as 5 mm × 5 mm. The vertical range typically ranges from 2 to 250 μm. The scan speed ranges typically from 1 μm/s to 25 mm/s. The sampling rate ranges typically from 50 Hz to 1 kHz.

2.3.1.1 Relocation

There are many situations where it would be very useful to look at a particular section of a surface before and after some experiment, such as grinding or run-in, to see what changes in the surface roughness have occurred. This can be accomplished by the use of a relocation table (Thomas, 1999). The table is bolted to the bed of the stylus instrument, and the specimen stage is kinematically located against it at three points and held in position pneumatically. The stage can be lowered and removed, an experiment of some kind performed on the specimen, and the stage replaced on the table. The stylus is then relocated to within the width of the original profile.

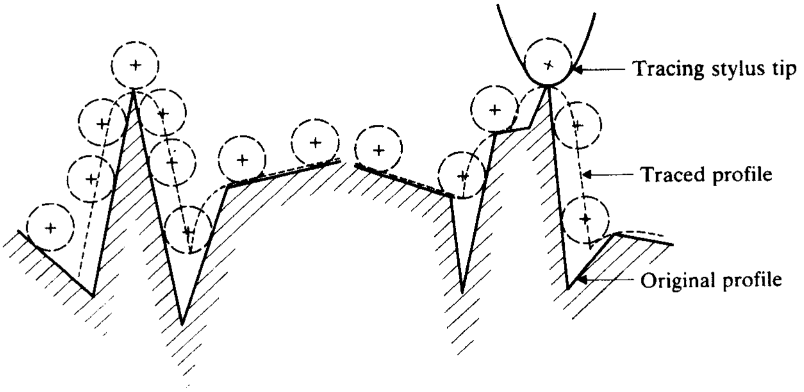

FIGURE 2.27 Distortion of profile due to finite dimensions of stylus tip (exaggerated).

2.3.1.2 Replication

Replication is used to obtain measurements on parts that are not easily accessible, such as internal surfaces or underwater surfaces. It is used in compliant surfaces because direct measurement would damage or misrepresent the surface (Thomas, 1999). The principle is simply to place the surface to be measured in contact with a liquid that will subsequently set to a solid, hopefully reproducing the detail of the original as faithfully as a mirror image or a negative. Materials such as plaster of paris, dental cement, or polymerizing liquids are used. The vital question is how closely the replica reproduces the features of the original. Lack of fidelity may arise from various causes.

2.3.1.3 Sources of Errors

A finite size of stylus tip distorts a surface profile to some degree (Radhakrishnan, 1970; McCool, 1984). Figure 2.27 illustrates how the finite size of the stylus distorts the surface profile. The radius of curvature of a peak may be exaggerated, and the valley may be represented as a cusp. A profile containing many peaks and valleys of radius of curvature of about 1 µm or less or many slopes steeper than 45° would probably be more or less misrepresented by a stylus instrument.

Another error source is due to stylus kinematics (McCool, 1984). A stylus of finite mass held in contact with a surface by a preloaded spring may, if traversing the surface at a high enough velocity, fail to maintain contact with the surface being traced. Where and whether this occurs depends on the local surface geometry, the spring constant to the mass ratio, and the tracing speed. It is clear that a trace for which stylus contact has not been maintained presents inaccurate information about the surface microroughness.

Stylus load also introduces error. A sharp stylus even under low loads results in an area of contact so small that the local pressure may be sufficiently high to cause significant local elastic deformation of the surface being measured. In some cases, the local pressure may exceed the hardness of the material, and plastic deformation of the surface may result. Styli generally make a visible scratch on softer surfaces, for example, some steels, silver, gold, lead, and elastomers (Poon and Bhushan, 1995a; Bhushan, 1996). The existence of scratches results in measurement errors and unacceptable damage. As shown in Figure 2.24 presented earlier, the stylus digs into the surface and the results do not truly represent the microroughness. It is important to select stylus loads low enough to minimize plastic deformation.

2.3.2 Optical Methods

When electromagnetic radiation (light wave) is incident on an engineering surface, it is reflected either specularly or diffusely or both (Figure 2.28). Reflection is totally specular when the angle of reflection is equal to the angle of incidence (Snell's law); this is true for perfectly smooth surfaces. Reflection is totally

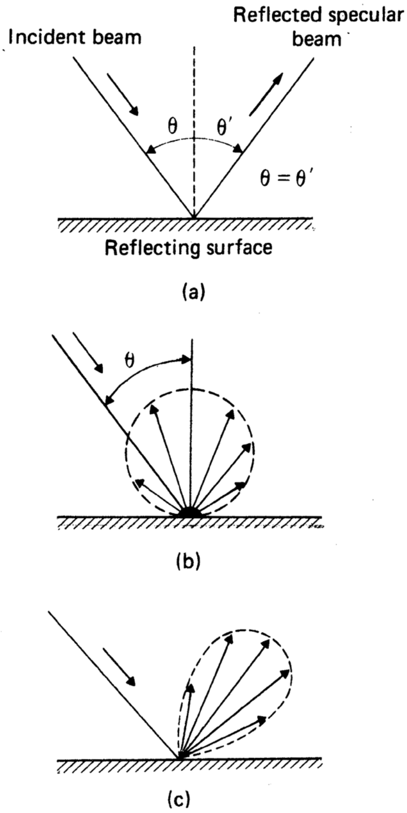

FIGURE 2.28 Modes of reflection of electromagnetic radiation from a solid surface, (a) specular only, (b) diffuse only, and (c) combined specular and diffuse. (From Thomas, T.R. (1999), *Rough Surfaces*, 2nd ed., Imperial College Press, London, U.K. With permission.)

diffuse or scattered when the energy in the incident beam is distributed as the cosine of the angle of reflection (Lambert's law). As roughness increases, the intensity of the specular beam decreases while the diffracted radiation increases in intensity and becomes more diffuse. In most real surfaces, reflections are neither completely specular nor completely diffuse. Clearly, the relationships between the wavelength of radiation and the surface roughness will affect the physics of reflection; thus, a surface that is smooth to radiation of one wavelength may behave as if it were rough to radiation of a different wavelength.

The reflected beams from two parallel plates placed normal to the incident beam interfere and result in the formation of the fringes (Figure 2.29). The fringe spacing is a function of the spacing of the two plates. If one of the plates is a reference plate and another is the engineering surface whose roughness is to be measured, fringe spacing can be related to the surface roughness. We have just described so-called two-beam optical interference. A number of other interference techniques are used for roughness measurement.

Numerous optical methods have been reported in the literature for measurement of surface roughness. Optical microscopy has been used for overall surveying, which only provides qualitative information. Optical methods may be divided into geometrical and physical methods (Thomas, 1999). Geometrical methods include taper sectioning and light sectioning methods. Physical methods include specular and diffuse reflections, speckle pattern, and optical interference.

2.3.2.1 Taper-Sectioning Method

In this technique, a section is cut through the surface to be examined at a shallow angle θ, thus effectively magnifying height variations by a factor cot θ, and is subsequently examined by an optical microscope.

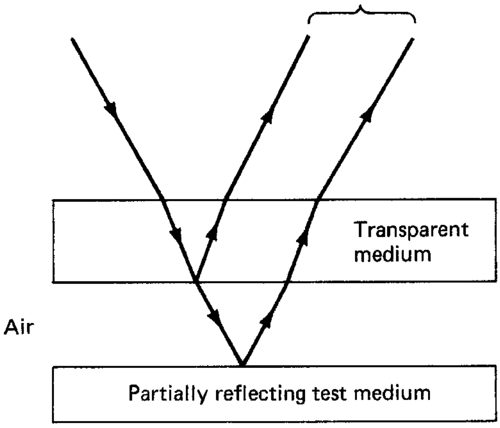

FIGURE 2.29 Optical schematic of two-beam interference.

The technique was first described by Nelson (1969). The surface to be sectioned has to be supported with an adherent coating that will prevent smearing of the contour during the sectioning operation. This coating must firmly adhere to the surface, must have a similar hardness, and should not diffuse into the surface. For steel surfaces, about 0.5-mm-thick electroplated nickel can be used. The specimen is then ground on a surface grinder at a typical taper angle between 1° and 6°. The taper section so produced is lapped, polished, and possibly lightly etched or heat tinted to provide good contrast for the optical examination.

The disadvantages of this technique include destruction of the test surface, tedious specimen preparation, and poor accuracy.

2.3.2.2 Light-Sectioning Method

The image of a slit (or a straight edge such as a razor blade) is thrown onto the surface at an incident angle of 45°. The reflected image will appear as a straight line if the surface is smooth, and as an undulating line if the surface is rough. The relative vertical magnification of the profile is the cosecant of the angle of incidence, in this case 1.4. Lateral resolution is about 0.5 µm. An automated system for three-dimensional measurement of surface roughness was described by Uchida et al. (1979). Their system consists of using the optical system to project the incident slit beam and then observing the image with an industrial television camera projected through a microscope; the table for the test surface is driven by a stepping motor.

2.3.2.3 Specular Reflection Methods

Gloss or specular reflectance (sometimes referred to as sheen or luster) is a surface property of the material, namely, the refractive index and surface roughness. Fresnel's equations provide a relationship between refractive index and reflectance. Surface roughness scatters the reflected light, thus affecting the specular reflectance. If the surface roughness σ is much smaller than the wavelength of the light (λ) and the surface has a Gaussian height distribution, the correlation between specular reflectance (*R*) and σ is described by (Beckmann and Spizzichino, 1963)

$$\frac{R}{R_0} = \exp\left[-\left(\frac{4\pi\sigma\cos\theta_i}{\lambda}\right)^2\right] \sim 1 - \left(\frac{4\pi\sigma\cos\theta_i}{\lambda}\right)^2 \tag{2.44}$$

where θ_i is the angle of incidence measured with respect to the sample normal, and R_0 is the total reflectance of the rough surface and is found by measuring the total light intensity scattered in all

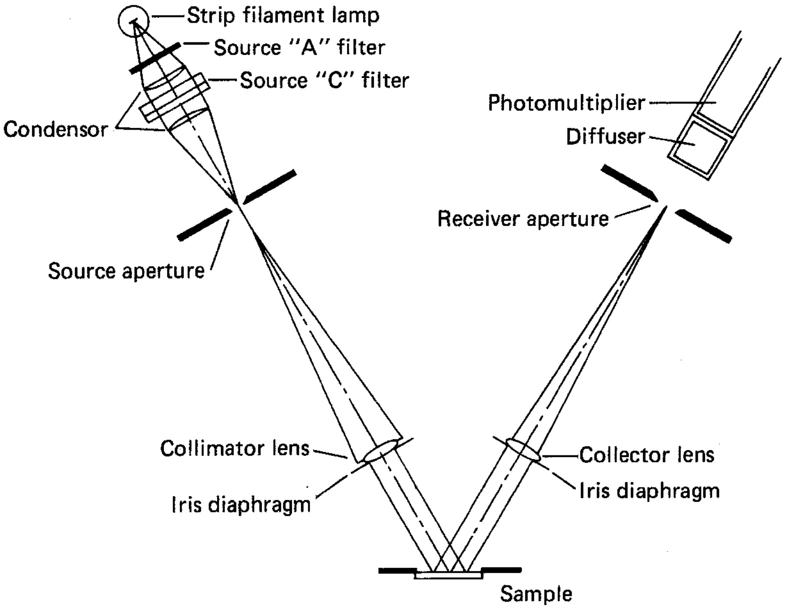

FIGURE 2.30 Schematic of a glossmeter. (From Budde, W. (1980), A reference instrument for 20°, 40°, and 85° gloss measurements, *Metrologia,* 16, 1-5. With permission.)

directions including the specular direction. If roughness-induced, light-absorption processes are negligible, R_0 is equal to the specular reflectance of a perfectly smooth surface of the same material. For rougher surfaces ($\sigma \geq \lambda/10$), the true specular beam effectively disappears, so R is no longer measurable. Commercial instruments following the general approach are sometimes called specular glossmeters (Figure 2.30). The first glossmeter was used in the 1920s. A glossmeter detects the specular reflectance (or gloss) of the test surface (of typical size of 50 mm × 50 mm), which is simply the fraction of the incident light reflected from a surface (Gardner and Sward, 1972). Measured specular reflectance is assigned a gloss number. The gloss number is defined as the degree to which the finish of the surface approaches that of the theoretical gloss standard, which is the perfect mirror, assigned a value of 1000.

The practical, primary standard is based on the black gloss (refractive index, $n = 1.567$) under angles of incidence of 20°, 60°, or 85°, according to ISO 2813 or American Society for Testing and Materials (ASTM) D523 standards. The specular reflectance of the black gloss at 60° for unpolarized radiation is 0.100 (Fresnel's equation, to be discussed later). By definition, the 60° gloss value of this standard is $1000 \times 0.10 = 100$. For 20 and 85°, Fresnel reflectances are 0.049 and 0.619, respectively, which are again by definition set to give a gloss value of 100. The glossmeter described by Budde (1980), operates over the wavelength range from 380 to 760 nm with a peak at 555 nm. There are five different angles of incidence that are commonly used — 20°, 45°, 60°, 75°, and 85°. Higher angles of incidence are used for rougher surfaces and vice versa.

Glossmeters are commonly used in the paint, varnish, and paper-coating industries (Gardner and Sward, 1972). These are also used in magnetic tapes at 45° or 60° incident angles, depending on the level of roughness (Bhushan, 1996). It is very convenient to measure the roughness of magnetic tape coatings during manufacturing by a glossmeter. The advantage of a glossmeter is its intrinsic simplicity, ease, and speed of analysis.

Other than accuracy and reproducibility, the major shortcoming of the gloss measurement is its dependence on the refractive index. Specular reflectance of a dielectric surface for unpolarized incident radiation increases with an increase in the refractive index according to Fresnel's equations (Hecht and Zajac, 1974),

$$R = \frac{\left(R_1 + R_2\right)}{2} \tag{2.45a}$$

where

$$R_1 = \left[\frac{\cos\theta_i - \left(n^2 - \sin^2 \theta_i\right)^{1/2}}{\cos\theta_i + \left(n^2 - \sin^2 \theta_i\right)^{1/2}}\right]^2 \tag{2.45b}$$

and

$$R_2 = \left[\frac{n^2 \cos\theta_i - \left(n^2 - \sin^2 \theta_i\right)^{1/2}}{n^2 \cos\theta_i + \left(n^2 - \sin^2 \theta_i\right)^{1/2}}\right]^2 \tag{2.45c}$$

where n is the refractive index of the dielectric material, θ_i is the angle of incidence with respect to surface normal, and R_1 and R_2 are the reflectance in the perpendicular and the parallel to the incident plane, respectively. For $\theta_i = 0$ (normal incidence), Equation 2.45 reduces to

$$R = \left(\frac{1-n}{1+n}\right)^2 \tag{2.46}$$

From Equations 2.45 and 2.46, we can see that a slight change of refractive index of the surface can change the gloss number. A change in the refractive index can come from a change in the supply of the raw material used in manufacturing the test surface (Fineman et al., 1981), a change in the composition of the surface (Wyant et al., 1984), or the aging of the surface (Alince and Lepoutre, 1980; Wyant et al., 1984). We therefore conclude that use of a glossmeter for roughness measurement is not very appropriate; however, for luster or general appearance, it may be acceptable.

2.3.2.4 Diffuse Reflection (Scattering) Methods

Vision depends on diffuse reflection or scattering. Texture, defects, and contamination causes scattering (Bennett and Mattson, 1989; Stover, 1995). It is difficult to obtain detailed roughness distribution data from the scattering measurements. Its spatial resolution is based on optical beam size, typically 0.1 to 1 mm in diameter. Because scatterometers measure light reflectance rather than the actual physical distance between the surface and the sensor, they are relatively insensitive to changes in temperature and mechanical or acoustical vibrations, making them extremely stable and robust. To measure large surface areas, traditional methods scan the roughness of several, relatively small areas (or sometimes just a single scan line) at a variety of locations on the surface. On the other hand, with scatterometers, the inspection spot is quickly and automatically rastered over a large surface. The scattering is sometimes employed to measure surface texture. This technique is particularly suitable for on-line roughness measurement during manufacture because it is continuous, fast, noncontacting, nondestructive, and relatively insensitive to the environment.

Three approaches have been used to measure defects and roughness by light scattering.

Total Integrated Scatter

The total integrated scatter (TIS) method is complementary to specular reflectance. Instead of measuring the intensity of the specularly reflected light, one measures the total intensity of the diffusely scattered

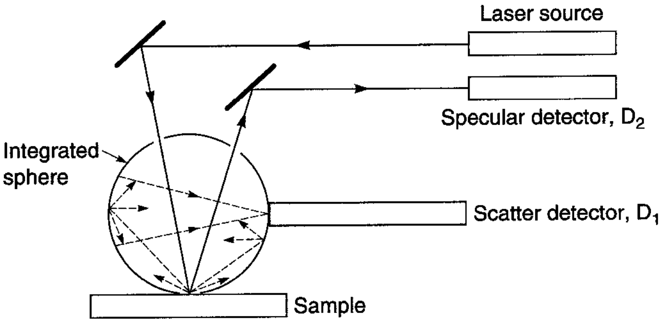

FIGURE 2.31 Schematic of the total integrated scatter apparatus with a diffuse integrated sphere. (From Stover, J.C., Bernt, M., and Schiff, T. (1996), TIS uniformity maps of wafers, disks and other samples, *Proc. Soc. Photo-Opt. Instrum. Eng.*, 2541, 21-25. With permission.)

light (Bennett, 1978; Stover, 1995). In the first TIS instrument, an aluminized, specular Coblentz sphere (90° integrating sphere) was used (Bennett and Porteus, 1961). Another method, shown in Figure 2.31 uses a high-reflectance diffuse integrated sphere. The incident laser beam travels through the integrated sphere and strikes the sample port at a few degrees off-normal. The specular reflection traverses the sphere again and leaves through the exit port where it is measured by the specular detector, D_2. The inside of the sphere is covered with a diffuse white coating that rescatters the gathered sample scatter throughout the interior of the sphere. The sphere takes on a uniform glow regardless of the orientation of the scatter pattern. The scatter signal is measured by sampling this uniform glow with a scatter detector, D_1, located on the right side of the sphere. The TIS is then the ratio of the total light scattered by the sample to the total intensity of scattered radiation (both specular and diffuse). If the surface has a Gaussian height distribution and its standard deviation σ is much smaller than the wavelength of light (λ), the TIS can be related to σ as given by Equation 2.41 (Bennett, 1978):

$$\mathrm{TIS} = \frac{R_0 - R}{R_0} = 1 - \exp\left[-\left(\frac{4\pi\sigma\cos\theta_i}{\lambda}\right)^2\right] \sim \left(\frac{4\pi\sigma\cos\theta_i}{\lambda}\right)^2 \tag{2.47a}$$

$$= \left(\frac{4\pi\sigma}{\lambda}\right)^2, \text{ if } \theta_i = 0 \tag{2.47b}$$

Samples of known specular reflectance are used to calibrate the reflected power (R_0) signals. The same samples, used to reflect the beam onto the sphere interior, can be used to calibrate the scattered power ($R_0 - R$) measurement signals (Stover et al., 1996).

Several commercial instruments, such as a Surfscan (Tencor Instruments, Mountain View, CA), Diskan (GCA Corp., Bedford, MA), and Dektak TMS-2000 (Veeco/Sloan Technology, Santa Barbara, CA) are built on this principle. In these instruments, to map a surface, either the sample moves or the light beam raster scans the sample. These instruments are generally used to generate maps of asperities, defects, or particles rather than microroughness distribution.

Diffuseness of Scattered Light

This approach relies on the observation that, over a large roughness range, the pattern of scattered radiation becomes more diffuse with increasing roughness. Hence, the goal here is to measure a parameter that characterizes the diffuseness of the scattered radiation pattern and to relate this parameter to the surface roughness. The ratio of the specular intensity to the intensity at one off-specular angle is measured.

Since this ratio generally decreases with increasing surface roughness, it provides a measure of the roughness itself.

Peters (1965) used this technique with the detector held 40° off-specular to determine the roughness of cylindrical parts while they were being ground. His results show a good correlation between the diffuseness and R_a over a range up to 0.3 µm. Using a pair of transmitting/receiving fiber-optic bundles set at different angles, the roughness measurements can be made with optics and instrumentation located remotely (North and Agarwal, 1983).

Angular Distributions

In principle, the entire angular distribution (AD) of the scattered radiation contains a great deal of information about the surface roughness. In addition to σ roughness, measurements of the angular distributions can yield other surface parameters, such as the average wavelength or the average slope. The angle of incidence is normally held constant and the AD is measured by an array of detectors or by a movable detector and is stored as a function of the angle of scattering.

The kind of surface information that may be obtained from the AD depends on the roughness regime. For $\sigma > \lambda$ and surface spatial wavelength $>\lambda$ (rough-surface limit), one is working in the geometrical optics regime, where the scattering may be described as scattering from a series of glints or surface facets oriented to reflect light from the incident beam into the scattering direction. This AD is therefore related to the surface slope distribution, and its width is a measure of the characteristic slope of the surface.

As σ and surface spatial wavelength decrease, the distribution becomes a much more complicated function of both surface slopes and heights and is difficult to interpret. For $\sigma < \lambda$ (smooth-surface limit), the scattering arises from the diffraction of light by the residual surface roughness viewed as a set of sinusoidal diffraction grating with different amplitudes, wavelengths, and directions across the surface. The intensity of the scattered light is determined by the vertical scale of the roughness and its angular width by the transverse scale; both scales are measured in units of the radiation wavelengths. It can be shown theoretically that the AD should directly map the power spectral density function of the surface roughness (Hildebrand et al., 1974; Stover, 1975; Church, 1979).

Figure 2.32 shows a sketch of various texture classes (in the smooth surface limit) on the left and their scattering signatures or power spectral densities on the right. The final sketch in the figure is the sum of the preceding ones, representing a real diamond-turned surface. Such signatures can be easily seen by reflecting a beam laser light from the surface onto a distant screen in a dark room (Church, 1979). Figure 2.33 shows the measured scattered intensity distribution from a diamond-turned gold surface using He–Ne laser light. He used a nonconventional method of scanning the scattering angle and the incident angle simultaneously by holding both the source and the detector fixed and rotating the specimen. The upper curve in Figure 2.33 shows the AD in the plane of incidence and perpendicular to the predominant lay of the surface. The sharp peak in the center is the specular reflection; a broad scattering distribution is due to the random component of the roughness; and a series of discrete lines are due to a periodic component roughness caused by the feed rate of the diamond tool. The lower curve in Figure 2.33 is an AD measured parallel to the lay direction, and it shows another broad distribution characteristic of the random roughness pattern in this direction. In principle, then, one can distinguish between effects due to periodic and random roughness components and can detect the directional property of surfaces.

A number of experimental systems have been developed. Clarke and Thomas (1979) developed a laser scanning analyzer system to measure rough surfaces; and scattered light was empirically related to the roughness. In their technique, a laser beam is reflected from a polygonal mirror rotating at high speed onto a surface where it is reflected into a fixed photodetector receiver masked to a narrow slit. The angular reflectance function is produced as the spot scans the strip. At a given moment in any scan, the fixed detector receives light scattered from the single point on the strip which happens at that instant to be illuminated by the deflected beam. The spot diameter can be set from 200 µm upward at a scan width of 623 mm, and the scanning speed is 5 kHz maximum.

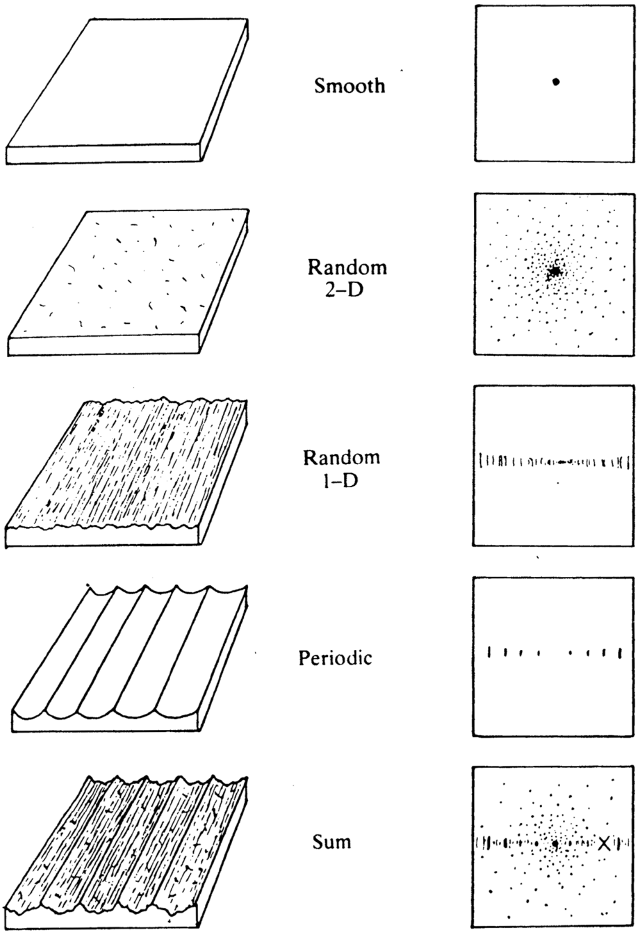

FIGURE 2.32 Texture classes and their scattering signatures when illuminated by coherent light. The final sketch is the sum of the preceding ones representing a real diamond-turned surface. (From Church, E.L. (1979), The measurement of surface texture and topography by differential light scattering, *Wear*, 57, 93-105. With permission.)

In the measurements reported by Clarke and Thomas (1979), all reflection curves were symmetrical and roughly the same shape irrespective of finish and resembled a Gaussian error curve (similar to the upper curve as shown in Figure 2.33). The surface roughness was found to be related to the width of the curve at half the maximum amplitude. Half-width tends to increase fairly linearly with the arithmetic average roughness and varies as about the fourth power of the mean absolute profile slope (Figure 2.34).

Vorburger et al. (1984) developed an AD instrument shown in Figure 2.35 in which a beam from an He–Ne laser illuminates the surfaces at an angle of incidence that may be varied. The scattered light distribution is detected by an array of 87 fiber-optic sensors positioned in a semicircular yoke that can be rotated about its axis so that the scattered radiation may be sampled over an entire hemisphere. They compared the angular scattering data with theoretical angular scattering distributions computed from digitized roughness profiles measured by a stylus instrument and found a reasonable correlation.

The three scattering methods described so far are generally limited by available theories to studies of surfaces whose σ are much less than λ. With an He–Ne laser as the light source, the preceding constraint means that these techniques have been used mainly on optical quality surfaces where σ < 0.1 µm. Within that limited regime, they can provide high-speed, quantitative measurements of the roughness of both isotropic surfaces and those with a pronounced lay. With rougher surfaces, AD may be useful as a comparator for monitoring both amplitude and wavelength surface properties. The ultimate vertical

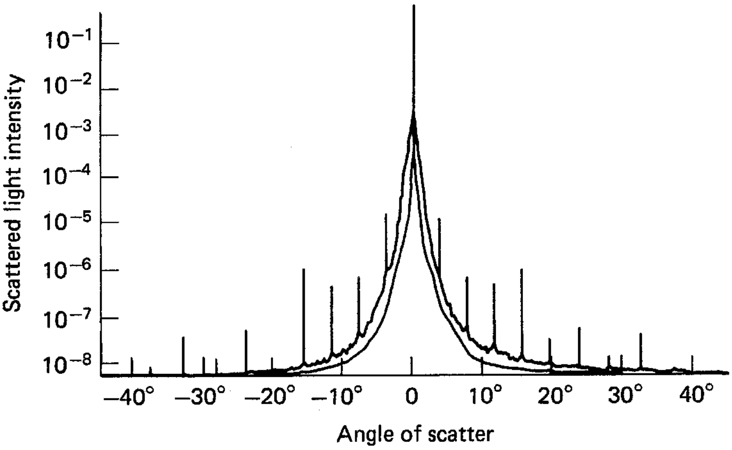

FIGURE 2.33 Experimental AD scattering spectrum of a diamond-turned surface. The upper curve shows a scan mode perpendicular to the lay direction. The lower curve shows a scan parallel to the lay direction. (From Church, E.L. (1979), The measurement of surface texture and topography by differential light scattering, *Wear,* 57, 93-105. With permission.)

resolution is 1 nm or better, but the horizontal range is limited to fairly short surface wavelengths. Both the vertical and horizontal ranges can be increased by using long wavelength (infrared) radiation, but there is an accompanying loss of vertical and horizontal resolution.

2.3.2.5 Speckle Pattern Method

When a rough surface is illuminated with partially coherent light, the reflected beam consists, in part, of random patterns of bright and dark regions known as speckle. Speckle is the local intensity variation between neighboring points in the overall AD discussed earlier. One means of clarifying the distinction between speckle and the AD is to note that speckle is the intensity noise that is usually averaged out to obtain the AD.

The technique used to relate speckle and surface roughness is the speckle pattern correlation measurement. Here, two speckle patterns are obtained from the test surface by illuminating it with different angles of incidence or different wavelengths of light. Correlation properties of the speckle patterns are then studied by recording the patterns. Goodman (1963) and others have shown that the degree of correlation between speckle patterns depends strongly on the surface roughness (σ) (see e.g., Ruffing and Fleischer, 1985).

2.3.2.6 Optical Interference Methods

Optical interferometry is a valuable technique for measuring surface shape, on both a macroscopic and microscopic scale (Tolansky, 1973). The traditional technique involves looking at the interference fringes and determining how much they depart from being straight and equally spaced. With suitable computer analysis, these can be used to completely characterize a surface. Bennett (1976) developed an interferometric system employing multiple-beam fringes of equal chromatic order (FECO). FECO are formed when a collimated beam of white light undergoes multiple reflections between two partially silvered surfaces, one of which is the surface whose profile is being measured and the other is a super-smooth reference surface. Based upon a television camera for the detection of the positional displacement of the fringes, this technique has yielded accuracies of σ on the order of 0.80 nm for the measurement of surface profiles. Lateral resolution of this system has been reported to be between 2 and 4 μm, over a 1 mm profile length.

Both the differential interference contrast (DIC) and the Nomarski polarization interferometer techniques (Francon, 1966; Francon and Mallick, 1971) are commonly used for qualitative assessment of surface roughness. While those interferometers are very easy to operate, and they are essentially insensitive to vibration, they have the disadvantage that they measure what is essentially the slope of the surface

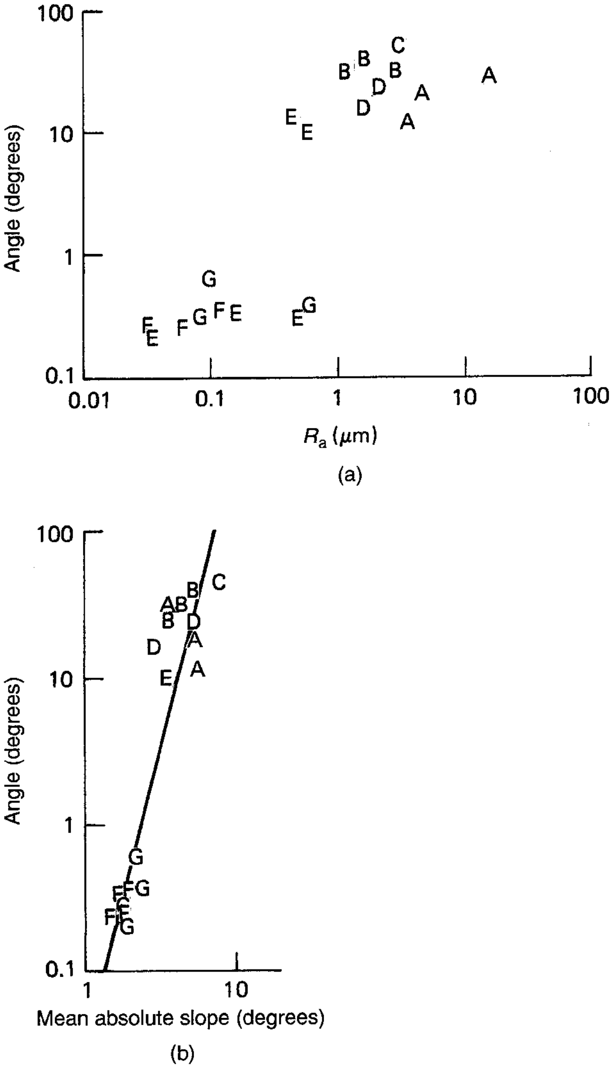

FIGURE 2.34 Variation of half-width with (a) average roughness and (b) mean absolute slope: A, milled: B, turned; C, spark eroded; D, shaped; E, ground; F, criss-cross lapped; G, parallel lapped. (From Clarke, G.M. and Thomas, T.R. (1979), Roughness measurement with a laser scanning analyzer, *Wear,* 57, 107-116. With permission.)

errors, rather than the surface errors themselves. A commercial Nomarski type profiler based on the linearly polarized laser beam is made by Chapman Instruments, Rochester, New York.

The Tolansky or multiple-beam interferometer is another common interferometer used with a microscope. The surface being examined must have a high reflectivity and must be in near contact with the interferometer reference surface, which can scratch the surface under test.

One of the most common optical methods for the quantitative measurement of surface roughness is to use a two-beam interferometer. The actual sample can be measured directly without applying a high-reflectivity coating. The surface-height profile itself is measured. The option of changing the magnification can be used to obtain different values of lateral resolution and different fields of view. Short-wavelength visible-light interferometry and computerized phase-shifting techniques can measure surface-height variations with resolutions better than 1/100 of a wavelength of light. The short wavelength of visible light is a disadvantage, however, when measuring large surface-height variations and slopes. If a single wavelength is used to make a measurement and the surface-height difference between adjacent measurement

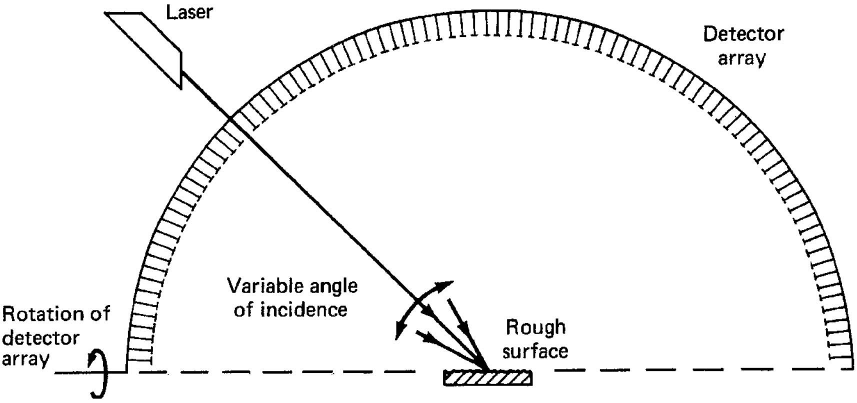

FIGURE 2.35 Schematic of angular-distribution scatter apparatus. (From Vorburger, T.V., Teague, E.C., Scire, F.E., McLay, M.J., and Gilsinn, D.E. (1984), Surface roughness studies with DALLAS array for light angular scattering, *J. Res. of NBS,* 89, 3-16. With permission.)

points is greater than one-quarter wavelength, height errors of multiple half-wavelengths may be introduced. The use of white light, or at least a few different wavelengths, for the light source can solve this height ambiguity problem. Two techniques can extend the range of measurement of surface microstructure where the surface slopes are large. One technique, measuring surface heights at two or more visible wavelengths, creates a much longer nonvisible synthetic wavelength, which increases the dynamic range of the measurement by the ratio of the synthetic wavelength to the visible wavelength. Increases in the dynamic range by factors of 50 to 100 are possible. Another more powerful method uses a white-light scanning interferometer, which involves measuring the degree of fringe modulation or coherence, instead of the phase of the interference fringes. Surface heights are measured by changing the path length of the sample arm of the interferometer to determine the location of the sample for which the white-light fringe with the best contrast is obtained. Vertical position at each location gives the surface height map. Various commercial instruments based on optical phase-shifting and vertical scanning interferometry are available (Wyko Corp., Tucson, AZ; Zygo Corp., Middlefield, CT; and Phase Shift Technology, Tucson, AZ).

Next, we describe the principles of operation following by a description of a typical commercial optical profiler.

Phase Shifting Interferometry

Several phase-measurement techniques (Wyant, 1975; Bruning, 1978; Wyant and Koliopoulos, 1981; Creath, 1988) can be used in an optical profiler to give more accurate height measurements than are possible by simply using the traditional technique of looking at the interference fringes and determining how much they depart from going straight and being equally spaced. One mode of operation used in commercial profilers is the so-called integrated bucket phase-shifting technique (Wyant et al., 1984, 1986; Bhushan et al., 1985).

For this technique, the phase difference between the two interfering beams is changed at a constant rate as the detector is read out. Each time the detector array is read out, the time variable phase α(t), has changed by 90° for each pixel. The basic equation for the irradiance of a two-beam interference pattern is given by

$$I = I_1 + I_2 \cos\left[\phi\left(x, y\right) + \alpha\left(t\right)\right] \tag{2.48}$$

where the first term is the average irradiance, the second is the interference term, and $\phi(x, y)$ is the phase distribution being measured. If the irradiance is integrated while $\alpha(t)$ varies from 0 to $\pi/2$, $\pi/2$ to π, and π to $3\pi/2$, the resulting signals at each detected point are given by

$$
\begin{aligned}
A(x, y) &= I_1' + I_2'\left[\cos \phi(x, y) - \sin \phi(x, y)\right] \\
B(x, y) &= I_1' + I_2'\left[-\cos \phi(x, y) - \sin \phi(x, y)\right] \\
C(x, y) &= I_1' + I_2'\left[-\cos \phi(x, y) + \sin \phi(x, y)\right]
\end{aligned}
\tag{2.49}
$$

From the values of A, B, and C, the phase can be calculated as

$$
\phi(x, y) = \tan^{-1}\left[\left(C(x, y) - B(x, y)\right) \big/ \left(A(x, y) - B(x, y)\right)\right] \tag{2.50}
$$

The subtraction and division cancel out the effects of fixed-pattern noise and gain variations across the detector, as long as the effects are not so large that they make the dynamic range of the detector too small to be used.

Four frames of intensity data are measured. The phase $\phi(x, y)$ is first calculated, by means of Equation 2.50, using the first three of the four frames. It is then similarly calculated using the last three of the four frames. These two calculated phase values are then averaged to increase the accuracy of the measurement.

Because Equation 2.50 gives the phase modulo 2π, there may be discontinuities of 2π present in the calculated phase. These discontinuities can be removed as long as the slopes on the sample being measured are limited so that the actual phase difference between adjacent pixels is less than π. This is done by adding or subtracting a multiple of 2π to a pixel until the difference between it and its adjacent pixel is less than π.

Once the phase $\phi(x, y)$ is determined across the interference field, the corresponding height distribution $h(x, y)$ is determined by the equation

$$
h(x, y) = \left(\frac{\lambda}{4\pi}\right)\phi(x, y) \tag{2.51}
$$

Phase shifting interferometry using a single wavelength has limited dynamic range. The height difference between two consecutive data points must be less than $\lambda/4$, where λ is the wavelength of the light used. If the slope is greater than $\lambda/4$ per detector pixel, then height ambiguities of multiples of half-wavelengths exist. One technique that has been very successful in overcoming these slope limitations is to perform the measurement using two or more wavelengths λ_1 and λ_2, and then to subtract the two measurements. This results in the limitation in height difference between two adjacent detector points of one quarter of a synthesized equivalent wavelength λ_{eq},

$$
\lambda_{eq} = \frac{\lambda_1 \lambda_2}{\left|\lambda_1 - \lambda_2\right|} \tag{2.52}
$$

Thus, by carefully selecting the two wavelengths it is possible to greatly increase the dynamic range of the measurement over what can be obtained using a single wavelength (Cheng and Wyant, 1985).

While using two wavelength phase-shifting interferometry works very well with step heights, it does not work especially well with rough surfaces. A much better approach is to use a broad range of wavelengths and the fringe modulation or coherence peak sensing approach, whose description follows.

Vertical Scanning Coherence Peak Sensing

In the vertical scanning coherence peak sensing mode of operation, a broad spectral white light source is used. Due to the large spectral bandwidth of the source, the coherence length of the source is short,

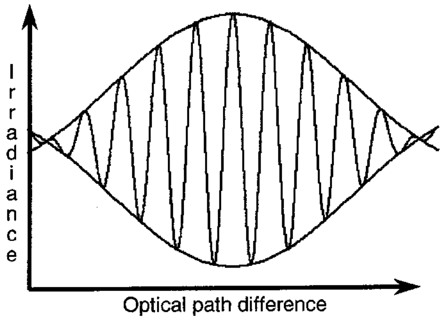

FIGURE 2.36 Irradiance at a single sample point as the sample is translated through focus. (From Caber, P. (1993), An interferometric profiler for rough surfaces, *Appl. Opt.*, 32, 3438-3441. With permission.)

and good contrast fringes will be obtained only when the two paths of the interferometer are closely matched in length. Thus, if in the interference microscope the path length of the sample arm of the interferometer is varied, the height variations across the sample can be determined by looking at the sample position for which the fringe contrast is a maximum. In this measurement there are no height ambiguities and, since in a properly adjusted interferometer the sample is in focus when the maximum fringe contrast is obtained, there are no focus errors in the measurement of surface texture (Davidson et al., 1987). Figure 2.36 shows the irradiance at a single sample point as the sample is translated through focus. It should be noted that this signal looks a lot like an amplitude modulated (AM) communication signal.

The major drawback of this type of scanning interferometer measurement is that only a single surface height is being measured at a time and a large number of measurements and calculations are required to determine a large range of surface height values. One method for processing the data that gives both fast and accurate measurement results is to use conventional communication theory and digital signal processing (DSP) hardware to demodulate the envelope of the fringe signal to determine the peak of the fringe contrast (Caber, 1993). This type of measurement system produces fast, noncontact, true three-dimensional area measurements for both large steps and rough surfaces to nanometer precision.

A Commercial Digital Optical Profiler

Figure 2.37 shows a schematic of a commercial phase shifting/vertical sensing interference microscope (Wyant, 1995). For smooth surfaces, the phase shifting mode is used since it gives subnanometer height resolution capability. For rough surfaces and large steps, up to 500-μm surface height variations, the

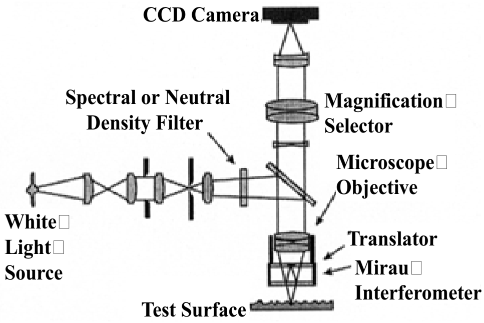

FIGURE 2.37 Optical schematic of the three-dimensional digital optical profiler based on phase-shifting/vertical sensing interferometer, Wyko HD-2000. (From Wyant, J.C. (1995), Computerized interferometric measurement of surface microstructure, *Proc. Soc. Photo-Opt. Instrum. Eng.*, 2576, 122-130. With permission.)

vertical scanning coherence sensing technique is used, which gives an approximately 3-nm height resolution. The instrument operates with one of several interchangeable magnification objectives. Each objective contains an interferometer, consisting of a reference mirror and beams splitter, which produces interference fringes when light reflected off the reference mirror recombines with light reflected off the sample. Determination of surface height using phase-shifting interferometry typically involves the sequential shifting of the phase of one beam of the interferometer relative to the other beam by known amounts, and measuring the resulting interference pattern irradiance. Using a minimum of three frames of intensity data, the phase is calculated and is then used to calculate the surface height variations over a surface. In vertical scanning interferometry when short coherence white light is used, these interference fringes are present only over a very shallow depth on the surface. The surface is profiled vertically so that each point on the surface produces an interference signal and the exact vertical position where each signal reaches its maximum amplitude can be located. To obtain the location of the peak, and hence the surface height information, this irradiance signal is detected using a CCD array. The instrument starts the measurement sequence by focusing above the top of the surface being profiled and quickly scanning downward. The signal is sampled at fixed intervals, such as every 50 to 100 nm, as the sample path is varied. The motion can be accomplished using a piezoelectric transducer. Low-frequency and DC signal components are removed from the signal by digital high bandpass filtering. The signal is next rectified by square-law detection and digitally lowpass filtered. The peak of the lowpass filter output is located and the vertical position corresponding to the peak is noted. Frames of interference data imaged by a video camera are captured and processed by high-speed digital signal processing hardware. As the system scans downward, an interference signal for each point on the surface is formed. A series of advanced algorithms are used to precisely locate the peak of the interference signal for each point on the surface. Each point is processed in parallel and a three-dimensional map is obtained.

The configuration shown in Figure 2.37 utilizes a two-beam Mirau interferometer at the microscope objective. Typically the Mirau interferometer is used for magnifications between 10 and 50×, a Michelson interferometer is used for low magnifications (between 1.5 and 5×), and the Linnik interferometer is used for high magnifications (between 100 and 200x) (Figure 2.38). A separate magnification selector is placed between the microscope objective and the CCD camera to provide additional image magnifications. High magnifications are used for roughness measurement (typically 40×), and low magnifications (typically 1.5×), are used for geometrical parameters. A tungsten halogen lamp is used as the light source. In the phase shifting mode of operation a spectral filter of 40-nm bandwidth centered at 650 nm is used to increase the coherence length. For the vertical scanning mode of operation the spectral filter is not used. Light reflected from the test surface interferes with light reflected from the reference. The resulting interference pattern is imaged onto the CCD array, with a size of about 736 × 480 and pixel spacing of about 8 μm. The output of the CCD array can be viewed on the TV monitor. Also, output from the CCD array is digitized and read by the computer. The Mirau interferometer is mounted on either a piezoelectric transducer (PZT) or a motorized stage so that it can be moved at constant velocity. During this movement, the distance from the lens to the reference surface remains fixed. Thus, a phase shift is introduced into one arm of the interferometer. By introducing a phase shift into only one arm while recording the interference pattern that is produced, it is possible to perform either phase-shifting interferometry or vertical scanning coherence peak sensing interferometry.

Major advantages of this technique are that its noncontact and three-dimensional measurements can be made rapidly without moving the sample or the measurement tool. One of the limitations of these instruments is that they can only be used for surfaces with similar optical properties. When dealing with thin films, incident light may penetrate the film and can be reflected from the film-substrate interface. This reflected light wave would have a different phase from that reflected from the film surface.

The smooth surfaces using the phase measuring mode can be measured with a vertical resolution as low as 0.1 nm. The vertical scanning mode provides a measurement range to about 500 μm. The field of view depends on the magnification, up to 10 mm × 10 mm. The lateral sampling interval is given by the detector spacing divided by the magnification; it is about 0.15 μm at 50× magnification. The optical

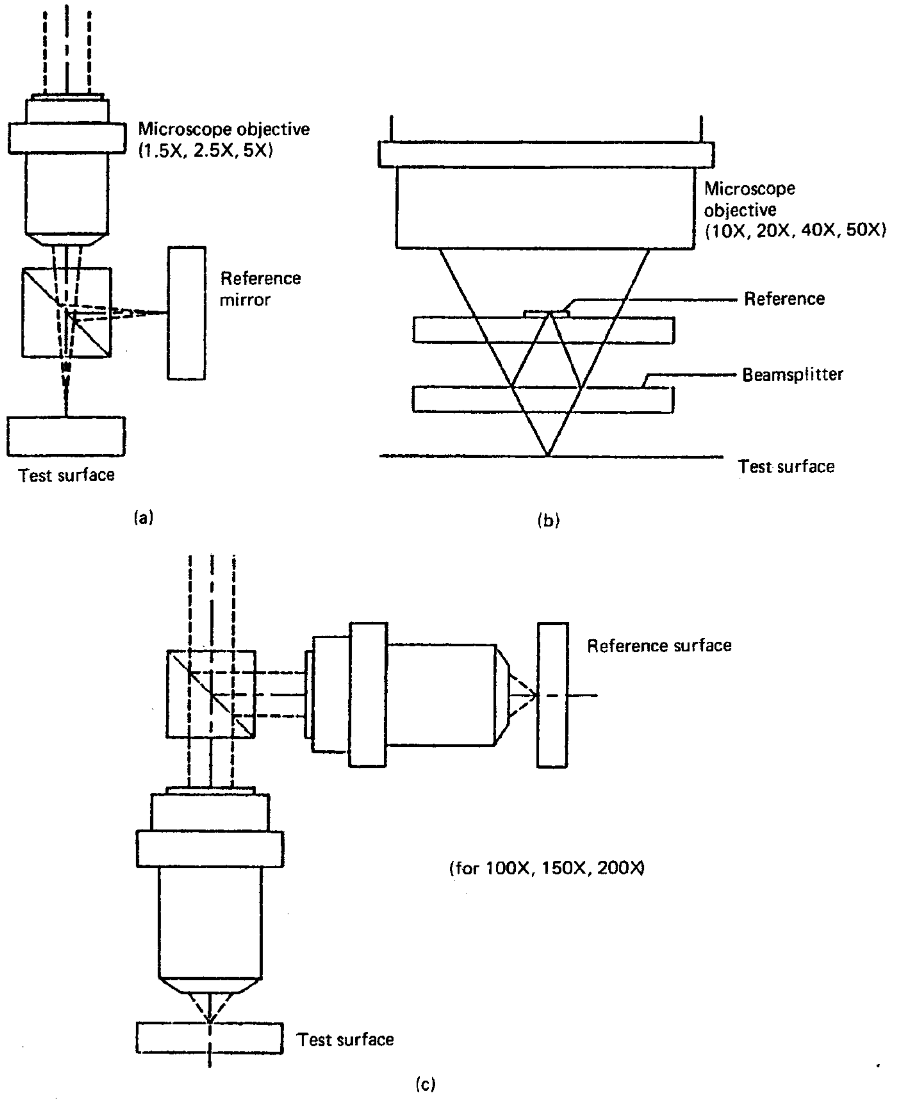

FIGURE 2.38 Optical schematics of (a) Michelson interferometer, (b) Mirau interferometer, and (c) Linnik interferometer.

resolution which can be thought of as the closest distance between two features on the surface such that they remain distinguishable, is given by 0.61 λ/(NA), where λ is the wavelength of the light source and NA is the numerical aperture of the objective (typically ranging from 0.036 for 1.5× to 0.5 for 40×). In practice, because of aberrations in the optical system, the actual resolution is slightly worse than the optical resolution. The best optical resolution for a lens is on the order of 0.5 μm. The scan speed is typically up to about 7 μm/s. The working distance, which is the distance between the last element in the objective and the sample, is simply a characteristic of the particular objective used.

Church et al. (1985) measured a set of precision-machined smooth optical surfaces by a mechanical-stylus profiler and an optical profiler in phase measuring mode. They reported an excellent quantitative agreement between the two profilers. Boudreau et al. (1995) measured a set of machined (ground, milled, and turned) steel surfaces by a mechanical stylus profiler and an optical profiler in the vertical scanning mode. Again, they reported an excellent quantitative agreement between the two profilers.

Typical roughness data using a digital optical profiler can be found in Wyant et al. (1984, 1986), Bhushan et al. (1985, 1988), Lange and Bhushan (1988), Caber (1993), and Wyant (1995).

2.3.3 Scanning Probe Microscopy (SPM) Methods

The family of instruments based on scanning tunneling microscopy (STM) and atomic force microscopy (AFM) are called scanning probe microscopies (SPM).

2.3.3.1 Scanning Tunneling Microscopy (STM)

The principle of electron tunneling was proposed by Giaever (1960). He envisioned that if a potential difference is applied to two metals separated by a thin insulating film, a current will flow because of the ability of electrons to penetrate a potential barrier. To be able to measure a tunneling current, the two metals must be spaced no more than 10 nm apart. In 1981, Dr. Gerd Binnig, Heinrich Rohrer, and their colleagues introduced vacuum tunneling combined with lateral scanning (Binnig et al., 1982; Binnig and Rohrer, 1983). Their instrument is called the scanning tunneling microscope (STM). The vacuum provides the ideal barrier for tunneling. The lateral scanning allows one to image surfaces with exquisite resolution, laterally less than 1 nm and vertically less than 0.1 nm, sufficient to define the position of single atoms. The very high vertical resolution of the STM is obtained because the tunnel current varies exponentially with the distance between the two electrodes, that is, the metal tip and the scanned surface. Very high lateral resolution depends upon the sharp tips. Binnig et al. overcame two key obstacles for damping external vibrations and for moving the tunneling probe in close proximity to the sample. An excellent review of this subject is presented by Bhushan (1999b). STM is the first instrument capable of directly obtaining three-dimensional images of solid surfaces with atomic resolution.

The principle of STM is straightforward. A sharp metal tip (one electrode of the tunnel junction) is brought close enough (0.3 to 1 nm) to the surface to be investigated (second electrode) that, at a convenient operating voltage (10 mV to 2 V), the tunneling current varies from 0.2 to 10 nA, which is measurable. The tip is scanned over a surface at a distance of 0.3 to 1 nm, while the tunnel current between it and the surface is sensed.

Figure 2.39 shows a schematic of one of Binnig and Rohrer's designs. The metal tip was fixed to rectangular piezodrives P_x, P_y, and P_z made out of commercial piezoceramic material for scanning. The sample was mounted on either a superconducting magnetic levitation or two-stage spring system to achieve the stability of a gap width of about 0.02 nm. The tunnel current J_T is a sensitive function of the

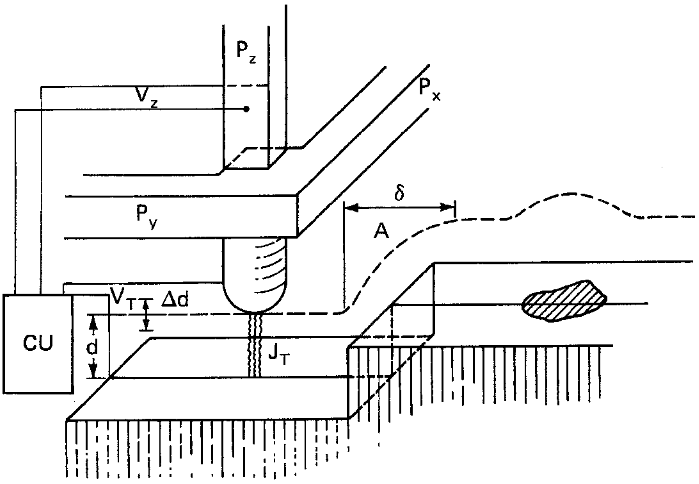

FIGURE 2.39 Principle of the operation of the scanning tunneling microscope. (From Binnig, G. and Rohrer, H. (1983), Scanning tunnelling microscopy, *Surf. Sci.*, 126, 236-244. With permission.)

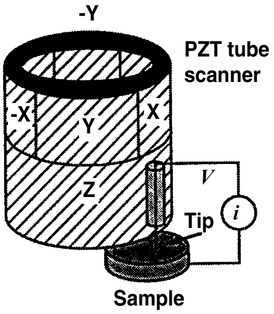

FIGURE 2.40 Principle of operation of a commercial STM; a sharp tip attached to a piezoelectric tube scanner is scanned on a sample.

gap width d, that is, $J_T \propto V_T \exp(-A\phi^{1/2}d)$, where V_T is the bias voltage, ϕ is the average barrier height (work function), and A ~ 1 if ϕ is measured in eV and d in Å. With a work function of a few eV, J_T changes by an order of magnitude for every angstrom change of h. If the current is kept constant to within, for example, 2%, then the gap h remains constant to within 1 pm. For operation in the constant current mode, the control unit (CU) applies a voltage V_z to the piezo P_z such that J_T remains constant when scanning the tip with P_y and P_x over the surface. At the constant work function ϕ, $V_z(V_x, V_y)$ yields the roughness of the surface z(x,y) directly, as illustrated at a surface step at A. Smearing of the step, δ (lateral resolution) is on the order of $(R)^{1/2}$, where *R* is the radius of the curvature of the tip. Thus, a lateral resolution of about 2 nm requires tip radii of the order of 10 nm. A 1-mm-diameter solid rod ground at one end at roughly 90° yields overall tip radii of only a few hundred nanometers, but with closest protrusion of rather sharp microtips on the relatively dull end yields a lateral resolution of about 2 nm. *In situ* sharpening of the tips by gently touching the surface brings the resolution down to the 1-nm range; by applying high fields (on the order of 10^8 V/cm) during, for example, half an hour, resolutions considerably below 1 nm can be reached.

There are a number of commercial STMs available on the market. Digital Instruments Inc. introduced the first commercial STM, the Nanoscope I, in 1987. In the Nanoscope III STM for operation in ambient air, the sample is held in position while a piezoelectric crystal in the form of a cylindrical tube scans the sharp metallic probe over the surface in a raster pattern while sensing and outputting the tunneling current to the control station (Figure 2.40) (Anonymous, 1992a). The digital signal processor (DSP) calculates the desired separation of the tip from the sample by sensing the tunneling current flowing between the sample and the tip. The bias voltage applied between the sample and the tip encourages the tunneling current to flow. The DSP completes the digital feedback loop by outputting the desired voltage to the piezoelectric tube. The STM operates in both the "constant height" and "constant current" modes depending on a parameter selection in the control panel. In the constant current mode, the feedback gains are set high, the tunneling tip closely tracks the sample surface, and the variation in the tip height required to maintain constant tunneling current is measured by the change in the voltage applied to the piezo tube (Figure 2.41). In the constant height mode, the feedback gains are set low, the tip remains at a nearly constant height as it sweeps over the sample surface, and the tunneling current is imaged (Figure 2.41). A current mode is generally used for atomic-scale images. This mode is not practical for rough surfaces. A three-dimensional picture [z(x,y)] of a surface consists of multiple scans [z(x)] displayed laterally from each other in the y direction. Note that if atomic species are present in a sample, the different atomic species within a sample may produce different tunneling currents for a given bias voltage. Thus the height data may not be a direct representation of the texture of the surface of the sample.

Samples to be imaged with STM must be conductive enough to allow a few nanometers of current to flow from the bias voltage source to the area to be scanned. In many cases, nonconductive samples can be coated with a thin layer of a conductive material to facilitate imaging. The bias voltage and the tunneling

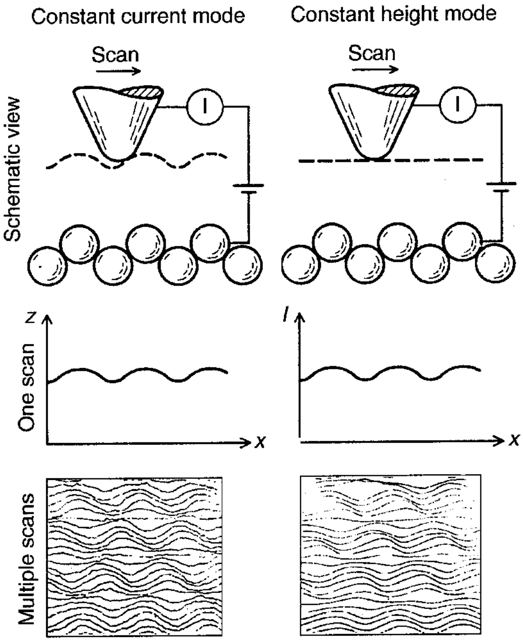

FIGURE 2.41 Scanning tunneling microscope can be operated in either the constant current or the constant height mode. The images are of graphite in air.

current depend on the sample. The scan size ranges from a fraction of a nanometer each way to about 125 μm × 125 μm. A maximum scan rate of 122 Hz can be used. Typically, 256 × 256 data formats are used. The lateral resolution at larger scans is approximately equal to scan length divided by 256.

The standalone STMs are available to scan large samples which rest directly on the sample.

The STM cantilever should have a sharp metal tip with a low aspect ratio (tip length/tip shank) to minimize flexural vibrations. Ideally, the tip should be atomically sharp, but in practice, most tip preparation methods produce a tip that is rather ragged and consists of several asperities with the one closest to the surface responsible for tunneling. STM cantilevers with sharp tips are typically fabricated from metal wires of tungsten (W), platinum-iridium (Pt–Ir), or gold (Au) and sharpened by grinding, cutting with a wire cutter or razor blade, field emission/evaporator, ion milling, fracture, or electrochemical polishing/etching (Ibe et al., 1990). The two most commonly used tips are made from either a Pt–Ir (80/20) alloy or tungsten wire. Iridium is used to provide stiffness. The Pt–Ir tips are generally mechanically formed and are readily available. The tungsten tips are etched from tungsten wire with an electrochemical process. The wire diameter used for the cantilever is typically 250 μm with the radius of curvature ranging from 20 to 100 nm and a cone angle ranging from 10 to 60° (Figure 2.42a). For calculations of normal spring constant and natural frequency of round cantilevers, see Sarid and Elings (1991).

Controlled geometry (CG) Pt/Ir probes are commercially available (Figure 2.42b). These probes are electrochemically etched from Pt/Ir (80/20) wire and polished to a specific shape which is consistent from tip to tip. Probes have a full cone angle of approximately 15° and a tip radius of less than 50 nm. For imaging of deep trenches (>0.25 μm) and nanofeatures, focused ion beam (FIB) milled CG milled probes with an extremely sharp tip radius (<5 nm) are used. For electrochemistry, Pt/Ir probes are coated with a nonconducting film (not shown in the figure).

2.3.3.2 Atomic Force Microscopy (AFM)

STM requires that the surface to be measured is electrically conductive. In 1985, Gerd Binnig and his colleagues developed an instrument called the atomic force microscope, capable of investigating surfaces

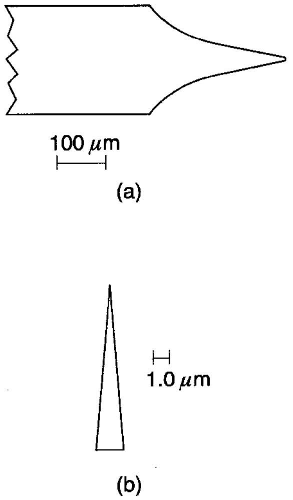

FIGURE 2.42 Schematics of (a) a typical tungsten cantilever with a sharp tip produced by electrochemical etching, and (b) CG Pt/Ir.

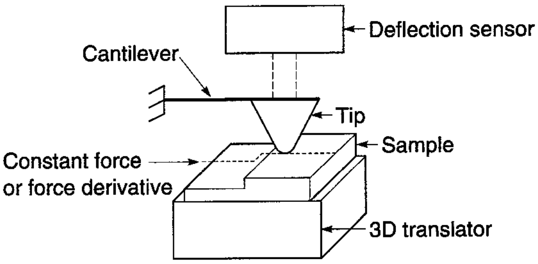

FIGURE 2.43 Principle of operation of the atomic force microscope.

of both conductors and insulators on an atomic scale (Binnig et al., 1986). Like the STM, the AFM relies on a scanning technique to produce very high resolution, three-dimensional images of sample surfaces. AFM measures ultrasmall forces (less than 1 nN) present between the AFM tip surface and a sample surface. These small forces are measured by measuring the motion of a very flexible cantilever beam having an ultrasmall mass. In the operation of high-resolution AFM, the sample is generally scanned instead of the tip as an STM, because AFM measures the relative displacement between the cantilever surface and reference surface, and any cantilever movement would add vibrations. However, AFMs are now available where the tip is scanned and the sample is stationary. As long as the AFM is operated in the so-called contact mode, little if any vibration is introduced.

The AFM combines the principles of the STM and the stylus profiler (Figure 2.43). In the AFM, the force between the sample and tip is detected rather than the tunneling current to sense the proximity of the tip to the sample. A sharp tip at the end of a cantilever is brought into contact with a sample surface by moving the sample with piezoelectric scanners. During initial contact, the atoms at the end of the tip experience a very weak repulsive force due to electronic orbital overlap with the atoms in the sample surface. The force acting on the tip causes a lever deflection which is measured by tunneling, capacitive, or optical detectors such as laser interferometry. The deflection can be measured to within ±0.02 nm, so for a typical lever force constant at 10 N/m a force as low as 0.2 nN (corresponding normal pressure ~200 MPa for an Si_3N_4 tip with a radius of about 50 nm against single-crystal silicon) could be detected. This operational mode is referred to as “repulsive mode” or “contact mode” (Binnig et al., 1986). An alternative is to use “attractive force imaging” or “noncontact imaging,” in which the tip is brought into close proximity (within a few nanometers) to, and not in contact with, the sample (Martin et al., 1987). A very weak van der Waals attractive force is present at the tip–sample interface. Although in this technique the normal pressure exerted at the interface is zero (desirable to avoid any surface deformation), it is slow and difficult to use and is rarely used outside research environments. In either mode, surface topography is generated by laterally scanning the sample under the tip while simultaneously measuring the separation-dependent force or force gradient (derivative) between the tip and the surface. The force gradient is obtained by vibrating the cantilever (Martin et al., 1987; Sarid and Elings, 1991) and measuring the shift of resonance frequency of the cantilever. To obtain topographic information, the interaction force is either recorded directly or used as a control parameter for a feedback circuit that maintains the force or force derivative at a constant value. Force derivative is normally tracked in noncontact imaging.

With AFM operated in the contact mode, topographic images with a vertical resolution of less than 0.1 nm (as low as 0.01 nm) and a lateral resolution of about 0.2 nm have been obtained. With a 0.01-nm displacement sensitivity, 10 nN to 1 pN forces are measurable. These forces are comparable to the forces associated with chemical bonding, e.g., 0.1 μN for an ionic bond and 10 pN for a hydrogen bond (Binnig et al., 1986). For further reading, see Bhushan (1999b).

STM is ideal for atomic-scale imaging. To obtain atomic resolution with AFM, the spring constant of the cantilever should be weaker than the equivalent spring between atoms on the order of 10 Nm. Tips have to be as sharp as possible. Tips with a radius ranging from 5 to 50 nm are commonly available. “Atomic resolution” cannot be achieved with these tips at the normal force in the nanoNewton range. Atomic structures obtained at these loads have been obtained from lattice imaging or by imaging of the crystal periodicity. Reported data show either perfectly ordered periodic atomic structures or defects on a large lateral scale, but no well-defined, laterally resolved atomic-scale defects like those seen in images routinely obtained with STM. Interatomic forces with one or several atoms in contact are 20 to 40 or 50 to 100 pN, respectively. Thus, atomic resolution with AFM is possible only with a sharp tip on a flexible cantilever at a net repulsive force of 100 pN or lower.

The first commercial AFM was introduced in 1989 by Digital Instruments. Now there are a number of commercial AFMs available on the market. Major manufacturers of AFMs for use in an ambient environment are: Digital Instruments Inc., Santa Barbara, CA; Park Scientific Instruments, Mountain View, CA; Topometrix, Santa Clara, CA; Seiko Instruments, Japan; Olympus, Japan; and Centre Suisse D’Electronique et de Microtechnique (CSEM) S.A., Neuchâtel, Switzerland. Ultra-high vacuum (UHV) AFM/STMs are manufactured by Omicron Vakuumphysik GmbH, Germany. Personal STMs and AFMs for ambient environment and UHV/STMs are manufactured by Burleigh Instruments Inc., Fishers, NY.

We describe here the commercial AFM for operation in ambient air, with scanning lengths ranging from about 0.7 μm (for atomic resolution) to about 125 μm (Figure 2.44a) (Anonymous, 1992b). This is the most commonly used design and the multimode AFM comes with many capabilities. In this AFM, the sample is mounted on a PZT tube scanner which consists of separate electrodes to scan precisely the sample in the *X-Y* plane in a raster pattern as shown in Figure 2.44b and to move the sample in the vertical (*Z*) direction. A sharp tip at the end of a flexible cantilever is brought into contact with the sample. Normal and frictional forces being applied at the tip–sample interface are measured using a laser beam deflection technique. A laser beam from a diode laser is directed by a prism onto the back of a cantilever near its free end, tilted downward at about 10° with respect to a horizontal plane. The reflected beam from the vertex of the cantilever is directed through a mirror onto a quad photodetector (split photodetector with four quadrants). The differential signal from the top and bottom photodiodes provides the AFM signal, which is a sensitive measure of the cantilever vertical deflection. Topographic features of the sample cause the tip to deflect in the vertical direction as the sample is scanned under the tip. This tip deflection will change the direction of the reflected laser beam, changing the intensity difference between the top and bottom photodetector (AFM signal). In the AFM operating mode of the “height mode,” for topographic imaging, or for any other operation in which the applied normal force is to be kept a constant, a feedback circuit is used to modulate the voltage applied to the PZT scanner to adjust the height of the PZT, so that the cantilever vertical deflection (given by the intensity difference

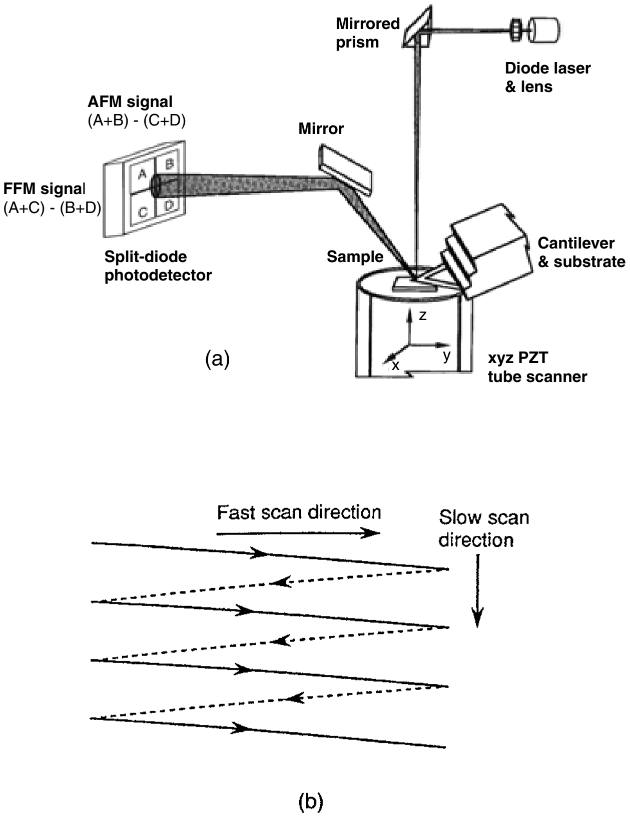

FIGURE 2.44 (a) Principle of operation of a commercial atomic force/friction force microscope, sample mounted on a piezoelectric tube scanner is scanned against a sharp tip and the cantilever deflection is measured using a laser beam deflection technique, and (b) schematic of triangular pattern trajectory of the AFM tip as the sample is scanned in two dimensions. During imaging, data are recorded only during scans along the solid scan lines.

between the top and bottom detector) will remain almost constant during scanning. The PZT height variation is thus a direct measure of surface roughness of the sample.

This AFM can be used for roughness measurements in the "tapping mode," also referred to as dynamic force microscopy. In the tapping mode, during scanning over the surface, the cantilever is vibrated by a piezo mounted above it, and the oscillating tip slightly taps the surface at the resonant frequency of the cantilever (70 to 400 kHz) with a 20 to 100 nm oscillating amplitude introduced in the vertical direction with a feedback loop keeping the average normal force constant. The oscillating amplitude is kept large enough so that the tip does not get stuck to the sample because of adhesive attraction. The tapping mode is used in roughness measurements to minimize the effects of friction and other lateral forces and to measure the roughness of soft surfaces.

There are several AFM designs in which force sensors using both the optical beam deflection method and scanning unit are mounted on the microscope head; then these AFM designs can be used as standalones. Lateral resolution of these designs is somewhat poorer than the designs in which the sample is scanned instead of the cantilever. The standalone AFMs can be placed directly on the large samples which cannot be fitted into the AFM assembly just described. There are other designs which head can adopt larger samples. A schematic of one such design is shown in Figure 2.45. The head both scans and generates the cantilever deflection. The beam emitted by the laser diode reflects off the cantilever and is detected by a quad photodetector.

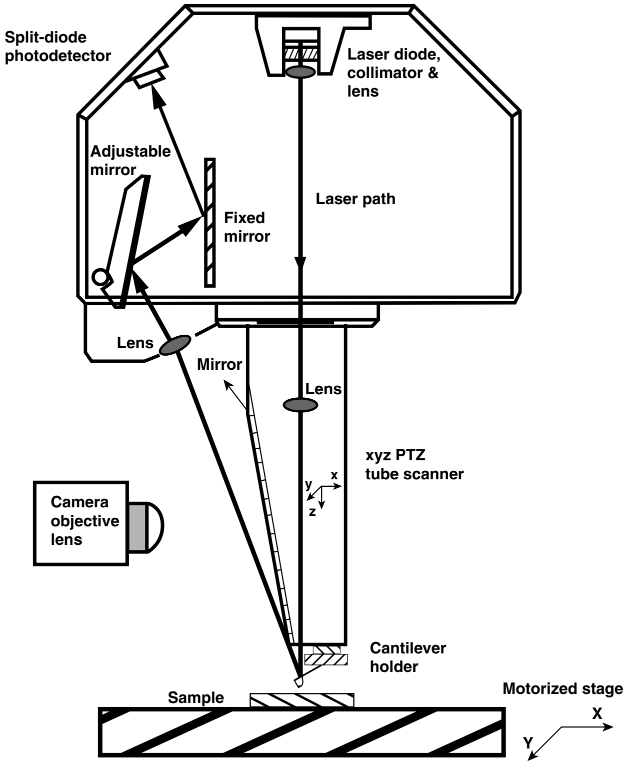

FIGURE 2.45 Principle of operation of a commercial atomic force/friction force microscope; the head both scans and generates the cantilever deflection. (From Anonymous (1994), *Dimension 3000 Instruction Manual,* Digital Instruments, Santa Barbara, CA. With permission.)

Roughness measurements are typically made using a sharp tip on a cantilever beam at a normal load on the order of 10 nN. The tip is scanned in such a way that its trajectory on the sample forms a triangular pattern. Scanning speeds in the fast and slow scan directions depend on the scan area and scan frequency. The scan sizes available for this instrument range from 0.7 μm × 0.7 μm to 125 μm × 125 μm. A maximum scan rate of 122 Hz can typically be used. Higher scan rates are used for small scan length. 256 × 256 data points are taken for each image. For example, scan rates in the fast and slow scan directions for an area of 10 μm × 10 μm scanned at 0.5 Hz are 10 μm/s and 20 nm/s, respectively. The lateral resolution at larger scans is approximately equal to scan length divided by 256. At first glance, scanning angle may not appear to be an important parameter for roughness measurements. However, the friction force between the tip and the sample will affect the roughness measurements in a parallel scan (scanning along the long axis of the cantilever). Therefore, a perpendicular scan may be more desirable. Generally, one picks a scanning angle that gives the same roughness data in both directions; this angle may be slightly different than that for the perpendicular scan.

The most commonly used cantilevers for roughness measurements in contact AFM mode are microfabricated plasma enhanced chemical vapor deposition (PECVD) silicon nitride triangular beams with integrated square pyramidal tips with a radius on the order of 30 to 50 nm. Four cantilevers with different sizes and spring stiffnesses (ranging from 0.06 to 0.6 N/m) on each cantilever substrate made of boron silicate glass are shown in Figure 2.46a. Etched single-crystal n-type silicon rectangular cantilevers with square pyramidal tips with a radius of about 10 nm are used for contact and tapping modes (Figure 2.46b). The cantilevers used for contact mode are stiff. For imaging within trenches by AFM, high-aspect ratio tips (HART) are used. An example of a probe is shown in Figure 2.46c. The probe is approximately 1 μm long and 0.1 μm in diameter. It tapers to an extremely sharp point (the radius is better than the resolution of most SEMs).

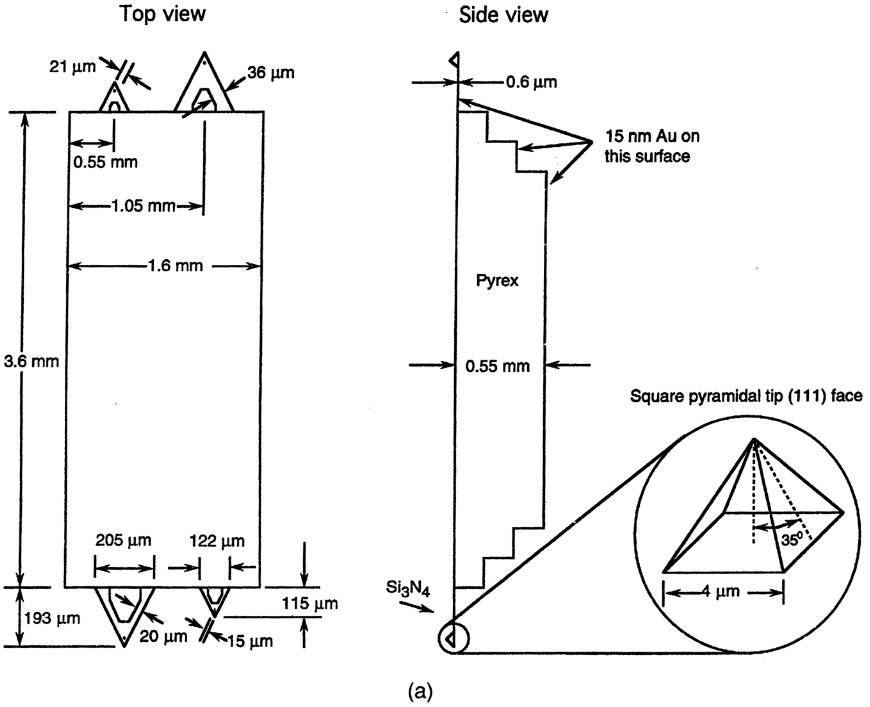

FIGURE 2.46 Schematics of (a) triangular cantilever beam with square pyramidal tips made of PECVD Si_3N_4, (b) rectangular cantilever beams with square pyramidal tips made of single-crystal silicon, and (c) high-aspect ratio Si_3N_4 probe.

2.3.4 Fluid Methods

Such techniques are mainly used for continuous inspection (quality control) procedures in service as they function without contact with the surface and are very fast. These provide numerical data that can only be correlated empirically to the roughness. The two most commonly used techniques are the hydraulic method and the pneumatic gaging method.

In the hydraulic method, sometimes called the outflow meter method, an open-bottomed vessel with a compliant annulus at its lower end is placed in contact with the surface to be measured and filled with water to a predetermined level. The time taken for a given volume of water to escape through the gap between the compliant annulus and the rough surface is measured (Thomas, 1999). A simple relationship exists between the standard deviation of asperity heights, σ_p, and the flow time, t,

$$\sigma_p = at^n \tag{2.53}$$

where *a* and *n* are constants determined by the characteristics of the method employed. This method was initially developed to measure road surfaces but can be used for any large roughness pattern.

The pneumatic gaging method is used for finer-scale roughness, such as machined metal surfaces. An outflow meter is used with air rather than water as the working medium, and surface roughness is measured by means of pneumatic resistance between the compliant annulus and the surface. For a constant rate of air flow, the pressure drop is determined by the overall surface roughness (Thomas, 1999).

2.3.5 Electrical Method

An electrical method used is the capacitance method based on the parallel capacitor principle. The capacitance between two conducting elements is directly proportional to their area and the dielectric

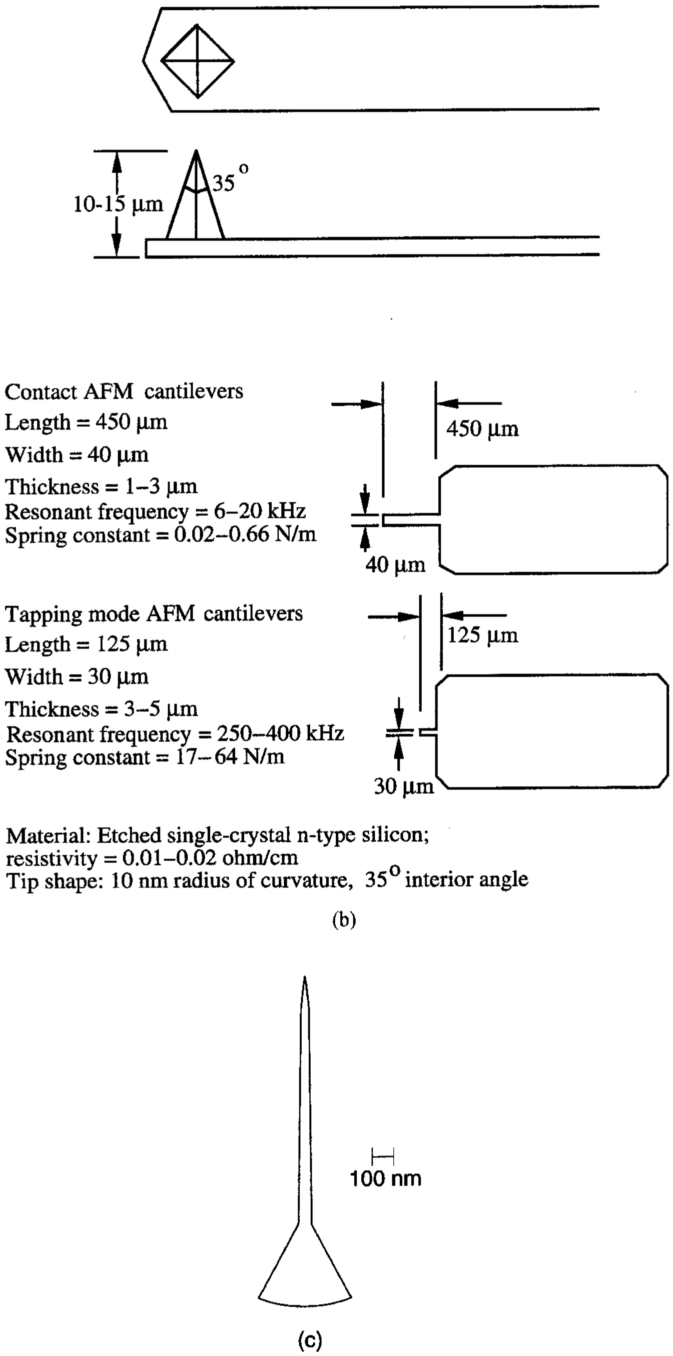
10-15 μm
35°
Contact AFM cantilevers
Length = 450 μm
Width = 40 μm
Thickness = 1–3 μm
Resonant frequency = 6–20 kHz
Spring constant = 0.02–0.66 N/m
450 μm
40 μm
Tapping mode AFM cantilevers
Length = 125 μm
Width = 30 μm
Thickness = 3–5 μm
Resonant frequency = 250–400 kHz
Spring constant = 17–64 N/m
125 μm
30 μm
Material: Etched single-crystal n-type silicon;
resistivity = 0.01–0.02 ohm/cm
Tip shape: 10 nm radius of curvature, 35° interior angle
(b)
100 nm
(c)

FIGURE 2.46 (continued)

constant of the medium between them, and inversely proportional to their separation. If a rough surface is regarded as the sum of a number of small elemental areas at different heights, it is fairly easy to work out the effective capacitance between it and a smooth surface disk for various deterministic models. The capacitance between a smooth disk surface and the surface to be measured is a function of the surface roughness. A commercial instrument is available based on this principle (Brecker et al., 1977). The capacitance method is also used for the continuous inspection procedures (quality control).

2.3.6 Electron Microscopy Methods

2.3.6.1 Reflection Electron Microscopy

Electron microscopy, both reflection and replica, can reveal both macroscopic and microscopic surface features (Halliday, 1955). But they have two major limitations: first, it is difficult to derive quantitative data; and second, because of their inherently limited field of view, they show only few asperities, whereas in fact the salient point about surface contact is that it involves whole populations of contacting asperities.

The use of SEM requires placing specimens in vacuum. In addition, for insulating specimens, conductive coating (e.g., gold or carbon) is required.

2.3.6.2 Integration of Backscattered Signals

Sato and O-Hori (1982) have shown that the profile of a surface can be obtained by processing backscattered electron signals (BES) using a computer connected to a scanning electron microscope (SEM). A backscattered electron image is produced by a BES, which is proportional to the surface inclination along the electron beam scanning. This means that the profile of the surface roughness can be derived by integrating the intensity of a BES, which varies along the scanning. Three-dimensional measurements of roughness are possible by making several scans. Disadvantages of the technique are that it requires the sample to have a conductive coating and the time taken to make measurements is fairly long.

2.3.6.3 Stereomicroscopy

The application of stereomicroscopy to obtain surface roughness information is based on the principle of stereo effects. The stereo effects can be obtained by preparing two images of the same surface with slightly different angular views (typically less than 10°). The result is a parallax shift between two corresponding image points of the same feature relative to some reference point, due to a difference in the elevation between the feature and the reference point. By measuring the parallax shift, one can extract the height information from these stereo-pair images.

Consider a point P on the specimen (Figure 2.47). Point O is an arbitrary reference point. After a clockwise rotation of angle θ, the point of interest is P′. The horizontal position of the feature is x_1 before rotation and x_2 after rotation. The distance $x_2 - x_1$ is known as parallax p. Simple trigonometry shows that the height of the feature P″ relative to the reference point O, z is given as (Boyde, 1970)

$$z = \frac{p}{2\sin\left(\theta/2\right)} \tag{2.54}$$

Image matching is the major step in stereomicroscopy. Given a stereo pair of images, we have to select a picture element (pixel), for example, the left image, and locate the corresponding conjugate pixel of

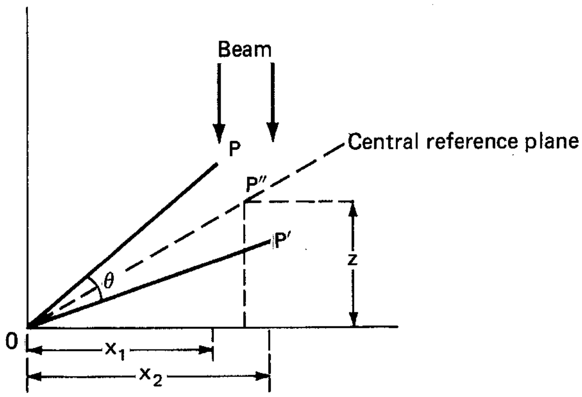

FIGURE 2.47 Surface height calculation model from stereo-pair images.

the right image. From the corresponding *x* positions, the *z* position of the pixel can be determined (relative to some arbitrary reference). This procedure is then repeated for all pixels of interest.

In the procedure of pixel matching, a window (array of pixels) is first set up in the left image. Then, a window array of the same size is placed around a potential pixel in the right image and a measure of image agreement between these two windows is computed. This procedure is repeated by moving the window in the right image until the best image content agreement is reached. The degree of image matching can be obtained by calculating the sum of the squared difference between the two comparing windows. This method calculates the sum of the squared intensity difference for all the pixels within the two windows. The smallest sum corresponds to the maximum image agreement.

The stereomicrographs are taken using an SEM with suitable resolution. The measurement technique requires several steps: obtaining SEM stereo pairs, stereo-pair image digitization (conversion of the analog data into digital form in the image so that they can be processed by a computer), and finally parallax analysis from which roughness information is deduced.

Since an SEM is typically used to obtain the pair of stereo images, the lateral resolution is limited by the electron beam size, which is typically 5 nm. Vertical resolution is a function of lateral parallax resolution and the angle ϕ.

2.3.7 Analysis of Measured Height Distribution

The measured height distribution across the sample can be analyzed to determine surface roughness statistics and geometrical parameters of a surface. The following surface roughness statistics can be obtained from the height distribution data: surface height distributions; surface slope and curvature distributions in x, y, and radial directions; heights, absolute slopes, and curvatures of all summits and the upper 25% summits; summit density and the upper 25% summit density; number of zero crossings per unit length in x, y, and two dimensions; and a three-dimensional plot of the autocovariance function with a contour of the autocovariance function at 0 and 0.1 (Wyant et al., 1986; Bhushan, 1996). The following geometrical parameters of a surface, for example, the radii of spherical curvature and cylindrical curvature, can be measured by fitting spherical and cylindrical surfaces, respectively.

2.3.8 Comparison of Measurement Methods

Comparison of the various methods of roughness measurement may be made on a number of grounds, such as ease of use, whether quantitative information can be obtained, whether three-dimensional data of topography can be obtained, lateral and vertical resolutions, cost, and on-line measurement capability. Table 2.4 summarizes the comparison of the relevant information.

The final selection of the measurement method depends very much on the application that the user has in mind. For in-process inspection procedures, measurement methods employing specular reflection, diffuse reflection, or speckle pattern are used. For continuous inspection (quality control) procedures requiring limited information, either fluid or electrical methods can be used. For procedures requiring detailed roughness data, either the stylus profiler, digital optical profiler or atomic force microscope is used. For a soft or super-finished surface, the digital optical profiler or AFM is preferred.

Roughness plots of a disk measured using an atomic force microscope (spatial resolution ~15 nm), noncontact optical profiler or NOP (spatial resolution ~1 µm), and a stylus profiler (spatial resolution ~0.2 µm), are shown in Figure 2.48. The figure shows that roughness is found at scales ranging from millimeters to nanometers. The measured roughness profile depends on the spatial and normal resolutions of the measuring instrument. Instruments with different lateral resolutions measure features with different length scales. It can be concluded that a surface is composed of a large number of length scales of roughness that are superimposed on each other. Figure 2.49 shows the comparison of AFM, NOP, and SP profiles extracted from the measurements with about the same profile lengths and sampling intervals. The roughness measurements are affected by the spatial (lateral) resolution of the measuring instrument.

TABLE 2.4 Comparison of Roughness Measurement Methods

Method	Quantitative Information	Three-Dimensional Data	Resolution (nm)		On-line Measurement Capability	Limitations
			Spatial	Vertical		
Stylus instrument	Yes	Yes	15–100	0.1–1	No	Contact type can damage the sample, slow measurement speed in 3D mapping
Optical methods						
Taper sectioning	Yes	No	500	25	No	Destructive, tedious specimen preparation
Light sectioning	Limited	Yes	500	0.1–1	No	Qualitative
Specular reflection	No	No	10^5–10^6	0.1–1	Yes	Semiquantitative
Diffuse reflection (scattering)	Limited	Yes	10^5–10^6	0.1–1	Yes	Smooth surfaces (<100 nm)
Speckle pattern	Limited	Yes			Yes	Smooth surfaces (<100 nm)
Optical interference	Yes	Yes	500–1000	0.1–1	No	
Scanning tunneling microscopy	Yes	Yes	0.2	0.02	No	Requires a conducting surface; scans small areas
Atomic force microscopy	Yes	Yes	0.2–1	0.02	No	Scans small areas
Fluid/electrical	No	No			Yes	Semiquantitative
Electron microscopy						Expensive instrumentation, tedious, limited data, requires a conducting surface, scans small areas
Reflection/replication	No	Yes	5	10–20	No	
Integration of backscattered signal	Yes	Yes	5	10–20	No	
Stereomicroscopy	Yes	Yes	5	50	No	

It refers to the stylus size of AFM and stylus profiler and the pixel size used in NOP for roughness measurement. For AFM and stylus profiler instruments, the ability of the stylus to reproduce the original surface features depends on the stylus size. The smaller the stylus, the closer it will follow the original profile. The stylus tip radius of AFM is smaller than SP and therefore AFM measurement is expected to be more accurate. A profile measured by AFM is used to assess the effect of the stylus size on the accuracy of roughness measurements (Poon and Bhushan, 1995a,b). Figure 2.50 shows the loci of different stylus radii on an AFM profile. By increasing the stylus size, the original profile is distorted, resulting in the underestimation of σ and the overestimation of β*. σ drops from 4.70 nm to 4.06 nm by 14% and β* increases from 0.16 µm to 0.44 µm by 175% when the stylus tip radius increases to 5 µm. NOP is an optical technique to measure surface roughness using the optical interference technique. The light intensity of the fringes is related to the surface height. In the optical system, the fringe pattern is discretized into pixels. Within one pixel or one sampling interval, the light intensity represents the average value of surface heights within the pixel. Effectively, the optical probe acts as an optical filter to remove high-frequency details using a cutoff length equal to the sampling interval. In the NOP measurement, the sampling interval is 1 µm. Therefore, the AFM profile in Figure 2.50a can be used to simulate the profile given by NOP by splitting the profile into a number of cutoff lengths equal to 1 µm. The mean of each cutoff length represents the surface height measured by NOP. A cubic spline curve is obtained to go through the mean points and shown in Figure 2.50e. σ for the simulated NOP profile is about 50% underestimated, and β* is 45% overestimated as compared with the AFM profile. Various roughness parameters of the disk measured using the AFM with two scan sizes are presented in Table 2.5.

As stated earlier, surface roughness is generally characterized by σ, sometimes along with other parameters. From the profiles in Figures 2.49 and 2.50, vertical roughness parameters σ, R_p, and P – V

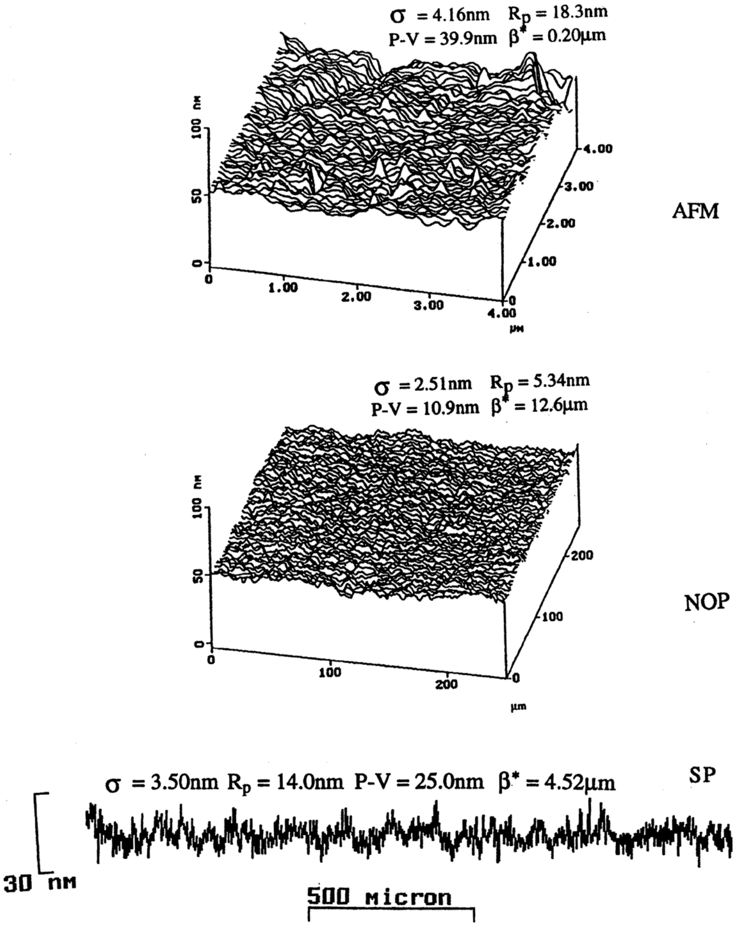

FIGURE 2.48 Surface roughness plots of a glass–ceramic disk measured using an atomic force microscope (spatial resolution ~15 nm), noncontact optical profiler (spatial resolution ~1 μm), and stylus profiler (tip radius ~0.2 μm).

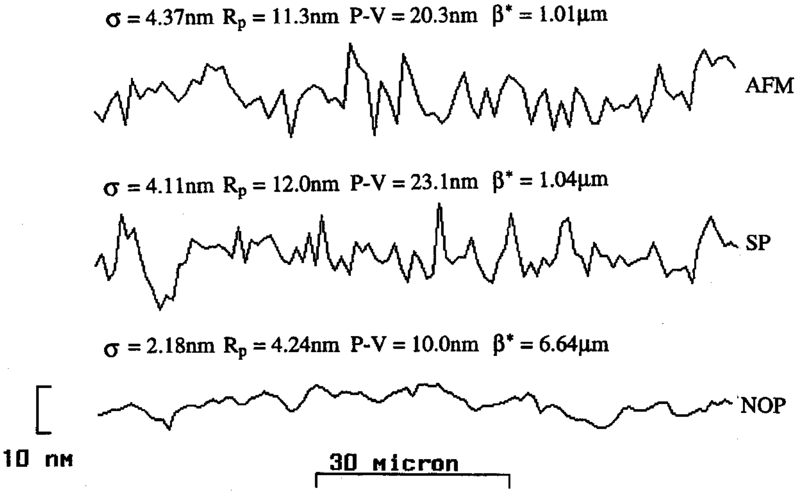

FIGURE 2.49 Comparison of surface plots of a glass–ceramic disk measured using AFM (~0.16 μm), SP (~0.2 μm), and NOP (~1 μm) drawn on a same scale.

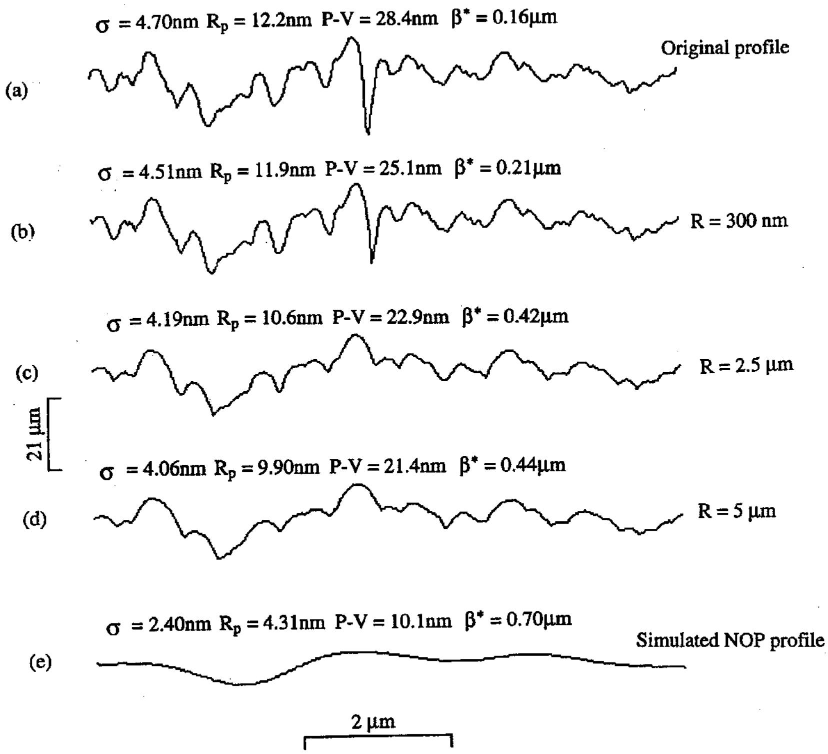

FIGURE 2.50 Simulated profiles of different stylus sizes sliding on the original AFM profile and the simulated NOP profile. (From Poon, C.Y. and Bhushan, B. (1995a), Comparison of surface roughness measurements by stylus profiler, AFM and non-contact optical profiler, *Wear,* 190, 76-88. With permission.)

are seen to increase with the measuring instruments in the following order: NOP < SP < AFM. On the other hand, the spatial parameter β* is seen to increase in the reverse order, i.e., AFM < SP < NOP. σ and β* as functions of scan size for three instruments shown in Figure 2.51 show a similar trend and are related to different instrument spatial resolutions. We also note that the σ initially increases with the scan size and then approaches a constant value, whereas β* increases monotonically with the scan size. The result of σ as a function of scan size suggests that the disk has a long-wavelength limit. It is expected that β*, which is a measure of wavelength structure, should also approach a constant value. In contrast, β* generally increases with the scan size. As the sampling interval increases with increasing scan size, high-frequency details of the original profile gradually disappear, resulting in high β*. σ is a vertical parameter not sensitive to sampling interval, but it generally increases with scan size. β* is a spatial parameter affected by both sampling interval and scan length. If the sampling interval can be kept the same for all scan sizes, β* will be expected to approach a constant value (Figure 2.52) (Poon and Bhushan, 1995a).

The question often asked is what instrument should one use for roughness measurement? For a given instrument, what scan size and sampling interval should one use? Deformation of asperities depends on the roughness, mechanical properties, and loading. Nanoasperities deform by plastic deformation, which is undesirable (Bhushan and Blackman, 1991; Poon and Bhushan, 1996). Therefore an instrument that can measure high frequency, such as AFM, should be used, particularly in low-load conditions. As stated earlier, a sampling interval equal to 0.25 and 0.50 times the correlation length at the selected scan size should be selected. A scan size equal to or greater than the value at which σ approaches a constant value, or twice the nominal contact size of the physical problem, whichever is smaller, should be used.

TABLE 2.5 Various Roughness Parameters of a Glass-Ceramic Disk Measured Using AFM at Scan Sizes

	Scan Size (μm^2)	
Roughness Parameters	8 × 8	32 × 32
σ, surface height (nm)	5.13	5.42
Skewness	–0.24	0.24
Kurtosis	6.01	4.1
σ, profile slope x (mrad)	53.5	22
σ, profile slope y (mrad)	67.7	25.2
σ, surface slope (mrad)	86.3	33.5
σ, profile curvature x (mm^{-1})	1635	235.5
σ, profile curvature y (mm^{-1})	3022	291.2
σ, surface curvature (mm^{-1})	1950	228.3
Summit height (nm)		
Mean	2.81	4.26
σ	5.56	5.08
Summit curvature (mm^{-1})		
Mean	3550	384
σ	1514	225.5
Summit-valley distance (nm)	45.9	48.5
Summit-mean distance (nm)	22.9	24.2
Summit density (μm^{-2})	15.6	2.97
Profile zero crossing *x* (mm^{-1})	2794	1279
Profile zero crossing *y* (mm^{-1})	4157	1572
Mean correlation length (μm)	0.32	0.67

x and y are along radial and tangential directions, respectively; summit threshold is taken as 0.5 nm.

2.4 Closure

The solid surfaces, irrespective of the method of formation, contain deviations from the prescribed geometrical form, ranging from macro- to nanoscale. In addition to surface deviations, the solid surface consists of several zones having physicochemical properties peculiar to the bulk material itself.

Surface texture, repetitive deviation from the nominal surface, includes roughness (nano- and microroughness, waviness, or macroroughness and lay). Surface roughness is most commonly characterized with two average amplitude parameters: R_a or (σ) R_q and R_t (maximum peak-to-valley height). However, the amplitude parameters alone are not sufficient for complete characterization of a surface, and spatial parameters are required as well. A random and isotropic surface can be completely characterized by two functions — the height distribution and autocorrelation functions. A random surface with Gaussian height distribution and exponential autocorrelation function can be completely characterized by two parameters σ and β*; these parameters can be used to predict other roughness parameters.

A surface is composed of a large number of length scales of roughness superimposed on each other. Hence, commonly measured roughness parameters depend strongly on the resolution of the measuring instrument and are not unique for a surface. The multiscale nature of rough surfaces can be characterized using a fractal analysis for fractal surfaces.

Various measurement techniques are used for off-line and on-line measurements of surface roughness. Optical techniques such as specular reflection and scattering, are commonly used for on-line semiquantitative measurements. Commonly used techniques for off-line measurements are either contact profiler — stylus profilers and atomic force microscopes- or noncontact profilers — optical profilers based on two-beam interference. Contact–stylus based profilers are the oldest form of measuring instruments and are most commonly used across the industry. However, the stylus tip can scratch the delicate

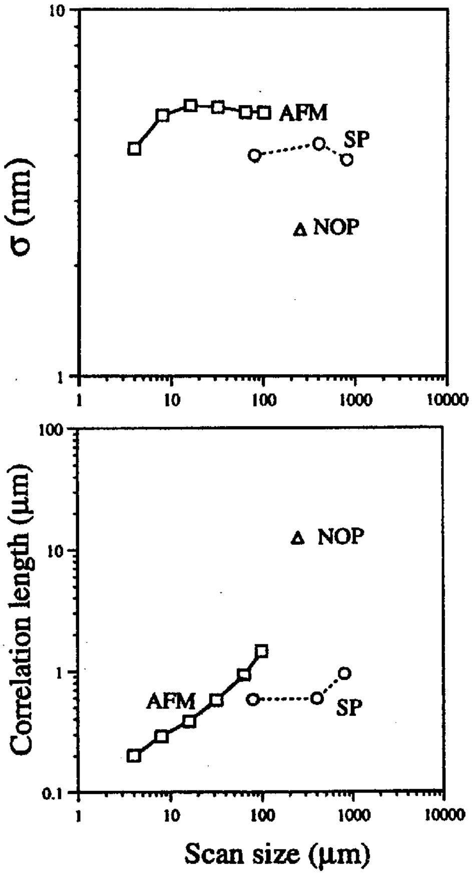

FIGURE 2.51 Variation of σ and β* with scan size for a glass–ceramic disk measured using AFM (scan length/256 data points), NOP (~1 µm), and SP (~0.2 µm).

surface during the course of the measurement. They also suffer from slow measurement speed, where three-dimensional mapping of the surfaces is required. Optical profilers are noncontact and can produce three-dimensional profiles rapidly and without any lateral motion between the optical head and the sample. Optical profilers can be used for surfaces with homogeneous optical properties, otherwise they need to be coated with a 10- to 20-nm-thick reflective coating (e.g., gold) before measurement. Lateral resolutions of profilers with sharp tips are superior to optical profilers. Nanoscale roughness with atomic-scale resolutions can be measured using atomic force microscopes which are used at ultralow loads. However, these are more complex to use.

Three-dimensional roughness height data can be processed to calculate a variety of amplitude and spatial functions and parameters. Without the use of long-wavelength filtering, waviness data can be obtained and analyzed.

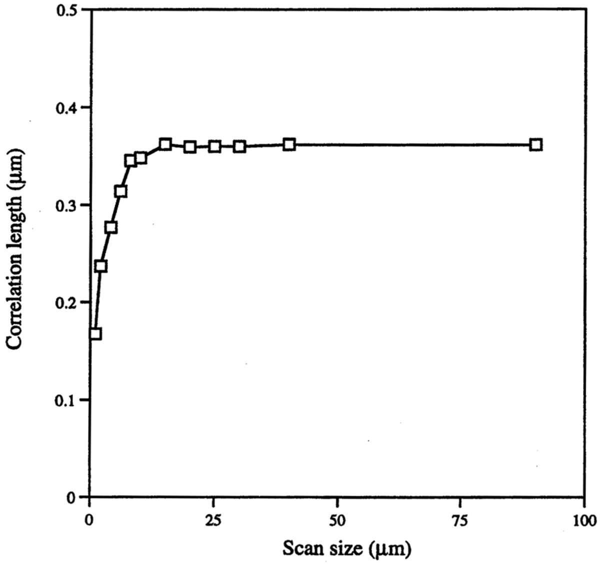

FIGURE 2.52 Variation of correlation length with scan size with a constant sampling interval (40 nm) for a glass–ceramic disk measured using AFM.

References

Abbott, E.J. and Firestone, F.A. (1933), Specifying surface quality, *Mech. Eng.,* 55, 569-572.

Alince, B. and Lepoutre, P. (1980), Plastic pigments in paper coatings, *Tappi,* 63, 49-53.

Anonymous (1975), Instruments for the Measurement of Surface Roughness by Profile Methods, ISO3274, International Standardization Organization.

Anonymous (1985), Surface Texture (Surface Roughness, Waviness and Lay), ANSI/ASME B46.1, ASME, New York.

Anonymous (1992a), *Nanoscope III Scanning Tunneling Microscope, Instruction Manual,* Digital Instruments, Santa Barbara, CA.

Anonymous (1992b), *Nanoscope III Atomic Force Microscope, Instruction Manual,* Digital Instruments, Santa Barbara, CA.

Anonymous (1994), *Dimension 3000 Instruction Manual,* Digital Instruments, Santa Barbara, CA.

Anonymous (1996a), *Dektak 800 Surface Profile Measuring System,* Veeco/Sloan Technology, Santa Barbara, CA.

Anonymous (1996b), *Tencor P-12 Disk Profiler Reference,* Tencor Instruments, Milpitas, CA.

Beckmann, P. and Spizzichino, A. (1963), *The Scattering of Electromagnetic Waves from Rough Surfaces,* Pergamon, New York, Chapter 5.

Bendat, J.S. and Piersol, A.G. (1986), *Engineering Applications of Correlation and Spectral Analysis, 2nd edition,* Wiley, New York.

Bennett, H.E. (1978), Scattering characteristics of optical materials, *Opt. Eng.,* 17, 480-488.

Bennett, J.M. (1976), Measurement of the rms roughness, autocovariance function and other statistical properties of optical surfaces using a FECO scanning interferometer, *Appl. Opt.,* 15, 2705-2721.

Bennett, H.E. and Porteus, J.O. (1961), Relation between surface roughness and specular reflectance at normal incidence, *J. Opt. Soc. Amer.,* 51, 123-129.

Bennett, J.M. and Mattson, L. (1989), *Introduction to Surface Roughness and Scattering,* Opt. Soc. of Am., Washington, D.C.

Bhushan, B. (1996), *Tribology and Mechanics of Magnetic Storage Devices, 2nd edition,* Springer, New York.

Bhushan, B. (1999a), *Principles and Applications of Tribology*, Wiley, New York.

Bhushan, B. (1999b), *Handbook of Micro/Nanotribology, 2nd edition*, CRC, Boca Raton.

Bhushan, B. and Blackman, G.S. (1991), Atomic force microscopy of magnetic rigid disks and sliders and its applications to tribology, *ASME J. Trib.*, 113, 452-457.

Bhushan, B., Wyant, J.C., and Koliopoulos, C.L. (1985), Measurement of surface topography of magnetic tapes by Mirau interferometry, *Appl. Opt.*, 24, 1489-1497.

Bhushan, B., Wyant, J.C., and Meiling, J. (1988), A new three-dimensional digital optical profiler, *Wear*, 122, 301-312.

Binnig, G. and Rohrer, H. (1983), Scanning tunnelling microscopy, *Surf. Sci.*, 126, 236-244.

Binnig, G., Rohrer, H., Gerber, Ch., and Weibel, E. (1982), Surface studies by scanning tunneling microscopy, *Phys. Rev. Lett.*, 49, 57-61.

Binnig, G., Quate, C.F., and Gerber, Ch. (1986), Atomic force microscope, *Phys. Rev. Lett.*, 56, 930-933.

Boudreau, B.D., Raja, J., Sannareddy, H., and Caber, P.J. (1995), A comparative study of surface texture measurement using white light scanning interferometry and contact stylus techniques, *Proc. Amer. Soc. Prec. Eng.*, 12, 120-123.

Boyde, A. (1970), Practical problems and methods in the three-dimensional analysis of scanning electron microscope images, *Scanning Electron Microscopy*, Proc. of the Third Annual SEM Symposium, IITRI, Chicago, IL, 105-112.

Brecker, J.N., Fromson, R.E., and Shum, L.Y. (1977), A capacitance based surface texture measuring system, *Annals CIRP*, 25, 375-377.

Bruning, J.H. (1978), Fringe scanning interferometers, in *Optical Shop Testing*, Malacara, D. (Ed.), Wiley, New York, 409-437.

Buckley, D. H. (1981), *Surface Effects in Adhesion, Friction, Wear and Lubrication*, Elsevier, Amsterdam.

Budde, W. (1980), A reference instrument for 20°, 40°, and 85° gloss measurements, *Metrologia*, 16, 1-5.

Bush, A.W., Gibson, R.D., and Keogh, G.P. (1976), The limit of elastic deformation in the contact of rough surfaces, *Mech. Res. Commun.*, 3, 169-174.

Bush, A.W., Gibson, R.D., and Keogh, G.P. (1979), Strongly anisotropic rough surfaces, *ASME J. Trib.*, 101, 15-20.

Caber, P. (1993), An interferometric profiler for rough surfaces, *Appl. Opt.*, 32, 3438-3441.

Cheng, Y.Y. and Wyant, J.C. (1985), Multiple-wavelength phase-shifting interferometry, *Appl. Opt.*, 24, 804-807.

Chilamakuri, S. and Bhushan, B. (1998), Contact analysis of non-gaussian random surfaces, *Proc. Instn. Mech. Engrs., Part J: J. Eng. Trib.*, 212, 19-32.

Church, E.L. (1979), The measurement of surface texture and topography by differential light scattering, *Wear*, 57, 93-105.

Church, E.L., Vorburger, T.V., and Wyant, J.C. (1985), Direct comparison of mechanical and optical measurements of the finish of precision-machined optical surfaces, *Opt. Eng.*, 24, 388-395.

Clarke, G.M. and Thomas, T.R. (1979), Roughness measurement with a laser scanning analyzer, *Wear*, 57, 107-116.

Creath, K. (1988), Phase-shifting interferometry techniques, in *Progress in Optics*, 26, Wolf, E. (Ed.), Elsevier, New York, 357-373.

Davidson, M., Kaufman, K., Mazor, I., and Cohen, F. (1987), An application of interference microscopy to integrated circuit inspection and metrology, *Proc. Soc. Photo-Opt. Instrum. Eng.*, 775, 233-247.

Elderton, P.E. and Johnson, L.J. (1969), *System of Frequency Curves*, Cambridge University Press, London, U.K.

Fineman, I., Engstrom, G., and Pauler, N. (1981), Optical properties of coated papers in relation to base papers and coating raw materials, *Paper Tech. and Indus.*, March, 59-65.

Francon, F. (1966), *Optical Interferometry*, Academic Press, San Diego, CA.

Francon, F. and Mallick, S. (1971), *Polarization Interferometers*, Wiley Interscience, New York.

Ganti, S. and Bhushan, B. (1995), Generalized fractal analysis and its applications to engineering surfaces, *Wear*, 180, 17-34.

Gardner, H.A. and Sward, G.G. (1972), *Paint Testing Manual, Physical and Chemical Examination: Paints, Varnishes, Lacquers and Colors, 13th ed.*, ASTM Special Pub. 500, Philadelphia.

Gatos, H.C. (1968), Structure of surfaces and their interactions, in *Interdisciplinary Approach to Friction and Wear,* Ku, P. M. (Ed.), SP-181, NASA, Washington, D.C., 7-84.

Giaever, I. (1960), Energy gap in superconductors measured by electron tunnelling, *Phys. Rev. Lett.,* 5, 147-148.

Goodman, J.W. (1963), *Statistical Properties of Laser Speckle Patterns*, Tech. Rep. No. 2303-1, Stanford Electronics Lab., Palo Alto, CA.

Greenwood, J.A. (1984), A unified theory of surface roughness, *Proc. R. Soc. Lond. A,* 393, 133-157.

Gupta, P.K. and Cook, N.H. (1972), Statistical analysis of mechanical interaction of rough surfaces, *ASME J. Lub. Tech.,* 94, 19-26.

Halliday, J.S. (1955), Surface examination by reflection electron microscopy, *Proc. Instn. Mech. Engrs.,* 109, 777-781.

Haltner, A.J. (1969), The physics and chemistry of surfaces: surface energy, wetting and adsorption, in *Boundary Lubrication,* Ling, F.F. et al. (Eds.), *ASME,* New York, 39-60.

Hecht, E. and Zajac, E. (1974), *Optics,* Addison-Wesley, Reading, MA, 82.

Hildebrand, B.P., Gordon, R.L., and Allen, E.V. (1974), Instrument for measuring the roughness of supersmooth surfaces, *Appl. Opt.,* 13, 177-180.

Ibe, J.P., Bey, P.P., Brandon, S.L., Brizzolara, R.A., Burnham, N.A., DiLella, D.P., Lee, K.P., Marrian, C.R.K., and Colton, R.J. (1990), On the electrochemical etching of tips for scanning tunneling microscopy, *J. Vac. Sci. Technol.,* A 8, 3570-3575.

Israelachvili, J.N. (1992), *Intermolecular and Surface Forces,* 2nd ed., Academic, San Diego, CA.

Kotwal, C.A. and Bhushan, B. (1996), Contact analysis of non-gaussian surfaces for minimum static and kinetic friction and wear, *Tribol. Trans.,* 39, 890-898.

Lange, S.R. and Bhushan, B. (1988), Use of two- and three-dimensional, noncontact surface profiler for tribology applications, *Surface Topography,* 1, 277-290.

Longuet-Higgins, M.S. (1957a), The statistical analysis of a random, moving surface, *Phil. Trans. R. Soc. Lond. A,* 249, 321-387.

Longuet-Higgins, M.S. (1957b), Statistical properties of an isotropic random surface, *Phil. Trans. R. Soc. Lond. A,* 250, 157-174.

Majumdar, A. and Bhushan, B. (1990), Role of fractal geometry in roughness characterization and contact mechanics of surfaces, *ASME J. Trib.,* 112, 205-216.

Martin, Y., Williams, C.C., and Wickramasinghe, H.K. (1987), Atomic force microscope-force mapping profiling on a sub 100-A scale, *J. Appl. Phys.,* 61, 4723-4729.

Massey, F.J. (1951), The Kolmogorov–Smirnov test for goodness of fit, *J. Amer. Statist. Assoc.,* 46, 68-79.

McCool, J.I. (1984), Assessing the effect of stylus tip radius and flight on surface topography measurements, *ASME J. Trib.,* 106, 202-210.

McGillem, C.D. and Cooper, G.R. (1984), *Continuous and Discrete Signal and System Analysis*, Holt, Rinehart & Winston, New York.

Nayak, P.R. (1971), Random process model of rough surfaces, *ASME J. Lub. Tech.,* 93, 398-407.

Nayak, P.R. (1973), Some aspects of surface roughness measurement, *Wear,* 26, 165-174.

Nelson, H.R. (1969), Taper sectioning as a means of describing the surface contour of metals, *Proc. Conf. on Friction and Surface Finish,* 2nd ed., MIT Press, Cambridge, MA, 217-237.

North, W.P.T. and Agarwal, A.K. (1983), Surface roughness measurement with fiber optics, *ASME J. Dyn. Sys., Meas. and Control,* 105, 295-297.

Peters, J. (1965), Messung Des Mitterauhwertes Zylindrischer Teile Während Des Schleifens, *VDI*-Berichte 90, 27.

Poon, C.Y. and Bhushan, B. (1995a), Comparison of surface roughness measurements by stylus profiler, AFM and non-contact optical profiler, *Wear,* 190, 76-88.

Poon, C.Y. and Bhushan, B. (1995b), Surface roughness analysis of glass-ceramic substrates and finished magnetic disks, and Ni-P coated Al–Mg and glass substrates, *Wear,* 190, 89-109.

Poon, C.Y. and Bhushan, B. (1996), Nano-asperity contact analysis and surface optimization for magnetic head slider/disk contact, *Wear,* 202, 83-98

Radhakrishnan, V. (1970), Effects of stylus radius on the roughness values measured with tracing stylus instruments, *Wear,* 16, 325-335.

Ruffing, B. and Fleischer, J. (1985), Spectral correlation of partially or fully developed speckle patterns generated by rough surfaces, *J. Opt. Soc. Amer.,* 2, 1637-1643.

Sarid, D. and Elings, V. (1991), Review of scanning force microscopy, *J. Vac. Sci. Technol. B,* 9, 431-437.

Sato, H. and O-Hori, M. (1982), Surface roughness measurement by scanning electron microscope, *Annals CIRP,* 31, 457-462.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

Smirnov, N. (1948), Table for estimating the goodness of fit of empirical distributions, *Annals of Mathematical Statistics,* 19, 279-281.

Stover, J.C. (1975), Roughness characterization of smooth machined surfaces by light scattering, *Appl. Opt.,* 14, 1796-1802.

Stover, J.C. (1995), *Optical Scattering: Measurement and Analysis*, 2nd ed., SPIE Optical Engineering Press, Bellingham, WA.

Stover, J.C., Bernt, M., and Schiff, T. (1996), TIS uniformity maps of wafers, disks and other samples, *Proc. Soc. Photo-Opt. Instrum. Eng.,* 2541, 21-25.

Thomas, T.R. (1999), *Rough Surfaces*, 2nd ed., Imperial College Press, London, U.K.

Tolansky, S. (1973), *Introduction to Interferometers*, Wiley, New York.

Uchida, S., Sato, H., and O-Hori, M. (1979), Two-dimensional measurement of surface roughness by the light sectioning method, *Annals CIRP,* 28, 419-423.

Vorburger, T.V., Teague, E.C., Scire, F.E., McLay, M.J., and Gilsinn, D.E. (1984), Surface roughness studies with DALLAS array for light angular scattering, *J. Res. of NBS,* 89, 3-16.

Whitehouse, D.J. (1994), *Handbook of Surface Metrology*, Institute of Physics Publishing, Bristol, U.K.

Whitehouse, D.J. and Archard, J.F. (1970), The properties of random surfaces of significance in their contact, *Proc. R. Soc. Lond. A,* 316, 97-121.

Whitehouse, D.J. and Phillips, M.J. (1978), Discrete properties of random surfaces, *Phil. Trans. R. Soc. Lond. A,* 290, 267-298.

Whitehouse, D.J. and Phillips, M.J. (1982), Two-dimensional discrete properties of random surfaces, *Phil. Trans. R. Soc. Lond. A,* 305, 441-468.

Williamson, J.B.P. (1968), Topography of solid surfaces, in *Interdisciplinary Approach to Friction and Wear,* Ku, P.M. (Ed.), SP-181, NASA Special Publication, NASA, Washington, D.C., 85-142.

Wyant, J.C. (1975), Use of an ac heterodyne lateral shear interferometer with real time wavefront corrections systems, *Appl. Opt.,* 14, 2622-2626.

Wyant, J.C. (1995), Computerized interferometric measurement of surface microstructure, *Proc. Soc. Photo-Opt. Instrum. Eng.,* 2576, 122-130.

Wyant, J.C. and Koliopoulos, C.L. (1981), Phase measurement system for adaptive optics, *Agard Conference Proceedings*, No. 300, 48.1-48.12.

Wyant, J.C., Koliopoulos, C.L., Bhushan, B., and George, O.E. (1984), An optical profilometer for surface characterization of magnetic media, *ASLE Trans.*, 27, 101-113.

Wyant, J.C., Koliopoulos, C.L., Bhushan, B., and Basila, D. (1986), Development of a three-dimensional noncontact digital optical profiler, *ASME J. Trib.,* 108, 1-8.