2

Rough Surface Topography

Abstract Approaches to solid surface topography characterization are discussed in this chapter, including experimental methods used in the conventional, nano-, and biotribology. Basic concepts of the statistical and fractal analysis of random rough surfaces and surface contact are reviewed. Common ways of surface modification, such as texturing and layer deposition, are discussed.

In this chapter, rough surface topography will be discussed with emphasis on the traditional engineering surfaces and their multiscale nature. Biological and biomimetic surfaces will be examined in detail in the third part of this book.

2.1 Rough Surface Characterization

A solid surface (or, more exactly, solid–liquid, solid–gas, or solid–vacuum interface) has complex structure and properties depending upon the nature of the material and the method of surface preparation. All solid surfaces, both natural and artificial, irrespective of the method of their formation, contain irregularities. No machining method can produce a molecularly flat surface on conventional materials. Even the smoothest surfaces, obtained by cleavage of some crystals (such as graphite or mica), contain irregularities, heights of which exceed interatomic distances. Engineering surfaces often have different types of random derivation from the prescribed form: the waviness, roughness, lay, and flow (Fig. 2.1). The waviness may result from machine vibration or chatter during machining as well as the heat treatment or warping strains. It includes irregularities with a relatively long (many microns) wavelength. Roughness is formed by fluctuation of the surface of short wavelengths, characterized by asperities (local maxima) and valleys (local minima). Lay is the principal direction of the predominant surface pattern, ordinarily determined by the production method. Flows are unintentional, unexpected, and unwanted interruptions in the texture [30, 32].

The distinction between various roughness features is somewhat conditional and may depend upon application and upon the resolution of the measuring equipment.

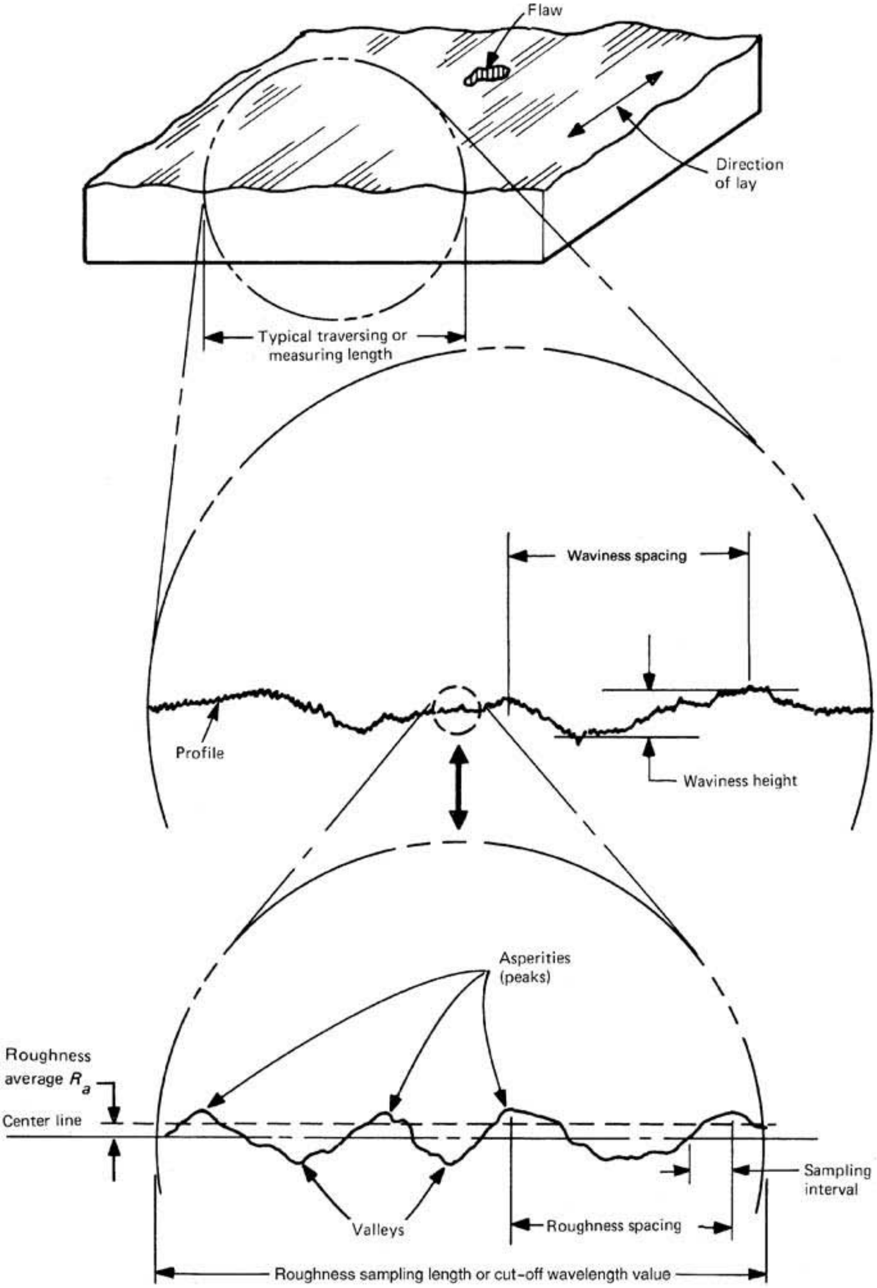


Fig. 2.1. Rough surface texture [32]

It is generally not possible to measure all the features at the same time. As will be discussed in the following, the very definition of the “asperity” involves serious problems. This is because a feature that may be a maximum of the surface profile at a given measurement resolution may involve numerous asperities and valleys when scrutinized at a higher resolution.

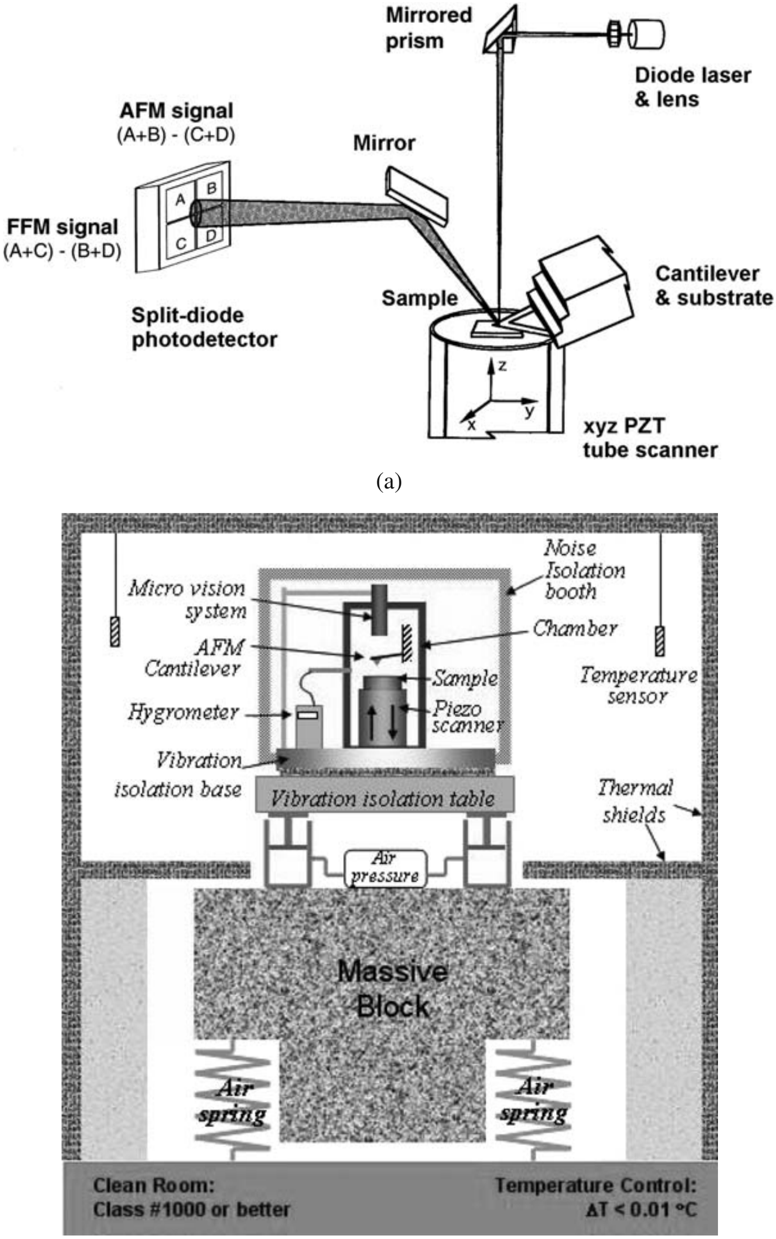
Various instruments are available to measure surface roughness. Mechanical (contact) and optical (noncontact) profilers are used to measure macro- and microscale roughness [30, 32]. The mechanical stylus method involves the amplifying and recording of vertical motions of a stylus tip displaced at a constant speed by the surface to be measured. The stylus is mechanically coupled mostly to a linear variable differential transformer or to an optical or capacitance sensor. As the stylus is scanned against the surface or the sample is transported relative to the stylus, an analog signal corresponding to the vertical stylus movement is amplified, conditioned, and digitized. The resolution of the profiler depends upon the dimensions of the tip—the sharper the tip is, the more fine details of the profile can be captured.

Optical methods are based upon measuring the light reflected from the surface. This includes the specular reflection methods that are used in glossmeters, diffuse reflection (scattering) methods, and optical interference methods that are used in various commercially available interferometers. Noncontact methods do not damage the measured surface, which is possible in the case of contact methods.

Several methods have been developed to measure roughness at the micro- and nanoscale. The family of instruments based on scanning tunnel microscopy (STM) and atomic force microscopy (AFM) is called scanning probe microscopy (SPM). In the STM, which was developed in early 1980s, a sharp tiny metal tip is brought very close (0.3–1 nm) to the sample surface. As the voltage between the tip and the sample is applied, the tunneling current is measured, which is proportional to the gap between the tip and the sample. As the tip is scanned against the sample, the sample height profile can be measured with subnanometer resolution [31, 34].

The AFM combines the principle of the STM and the stylus profiler. In the AFM, the tip (with the radius of few nanometers) is placed at the end of a long (dozens of micrometers) stiff cantilever (Fig. 2.2). The cantilever deflection is measured by determining the position of a laser beam reflected from the cantilever surface. In the contact mode, the tip scans the sample and the height map can be obtained with subnanometer resolution. In the noncontact mode, the van der Waals adhesion force acts upon the tip and results in cantilever deflection. As the stiffness of the cantilever is known, the deflection can be converted into the force unit (with subnanonewton resolution). The AFM can operate in ambient air as well as in vacuum [31, 34].

Scanning electron microscopy (SEM) can also be used for studying surface features; however, it has several limitations. First, it is difficult to obtain quantitative data from the SEM, and second, the field of view in SEM is limited. The use of the SEM requires placing the specimen in vacuum. In addition, a conductive coating is required to insulate samples [30, 32]. For biological specimens, there is the technique known as environmental scanning electron microscopy (ESEM), which allows one to conduct measurements in controlled humidity and pressure conditions.



(a)

(b)

Fig. 2.2. Atomic force microscope (AFM). **a** Principle of operation. A sample mounted on a piezoelectric tube (PZT) scanner scanned against a sharp tip and the cantilever deflection is measured using a laser beam [32]. **b** Vibration isolated clean-room setup for the AFM used at the NIST (credit to Dr. S.H. Yang, NIST)

Modern methods of surface structure analysis include X-ray spectrometry, Raman spectroscopy, electron diffraction, and others.

In addition to surface irregularities, the technical solid surface itself involves several zones or layers, such as the chemisorbed layer (0.3 nm), physisorbed layer (0.3–3 nm), chemically reacted layer (10–100 nm), etc. [30, 32]. In the chemisorbed layer, the solid surface bonds to the adsorption species through covalent bonds with an actual sharing of electrons. In the physisorbed layer, there are no chemical bonds between the substrate and the adsorbent, and only van der Waals force are involved. The van der Waals force is relatively weak (under 10 kJ/mol) and long range (nanometers) as opposed to strong (40–400 kJ/mol) and short range (comparable with the interatomic distance of about 0.3 nm). Typical adsorbents are oxygen, water vapor, or hydrocarbons from the environment, which can condense at the surface. While the chemisorbed layer is usually a monolayer, the physisorbed layer may include several layers of molecules. The chemically reacted layer is significantly thicker and involves many layers of molecules. The typical example of the chemically reacted layer is the oxide layer at the surface of a metallic substrate.

2.2 Statistical Analysis of Random Surface Roughness

There are several quantitative parameters commonly used to characterize random solid surface roughness, i.e., a random derivation from the nominal (prescribed) shape. These parameters include the amplitude (or height) parameters and the spatial parameters [30, 32, 316] (Fig. 2.3). The most commonly used amplitude parameter is the root mean square (RMS) or the standard deviation from the center-line average. For a 2D roughness profile $z(x)$, the center-line average is defined as the arithmetic mean of the absolute value of the vertical deviation from the mean line of the profile (Fig. 2.4)

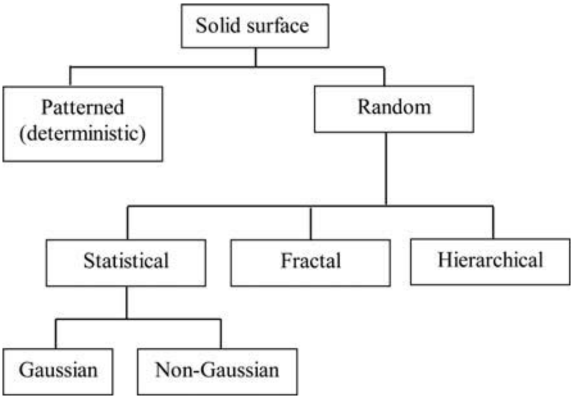


Fig. 2.3. Typology of rough surfaces (based on [32])

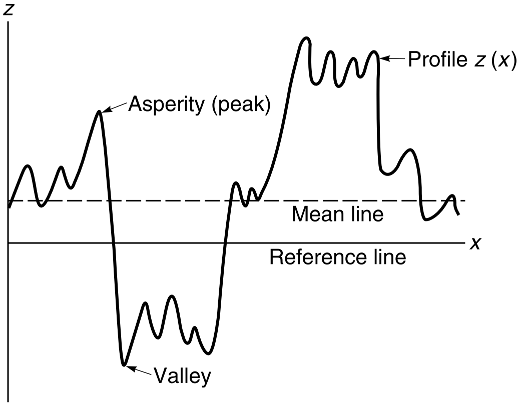


Fig. 2.4. Schematics of a rough surface profile [32]

$$R_\mathrm{a} = \frac{1}{L}\int_0^L |z - m|\,\mathrm{d}x, \tag{2.1}$$

where L is the sampling length,

$$m = \frac{1}{L}\int_0^L z\,\mathrm{d}x. \tag{2.2}$$

The square RMS is given by

$$\sigma^2 = \frac{1}{L}\int_0^L (z - m)^2\,\mathrm{d}x. \tag{2.3}$$

Since different rough surface profiles can have the same RMS, additional parameters are required to characterize details of surface profile. Two additional statistical parameters are the skewness and kurtosis, which are given in the normalized form by

$$Sk = \frac{1}{\sigma^3 L}\int_0^L (z - m)^3\,\mathrm{d}x, \tag{2.4}$$

and

$$K = \frac{1}{\sigma^4 L}\int_0^L (z - m)^4\,\mathrm{d}x. \tag{2.5}$$

A surface with a negative skewness has a larger number of local maxima above the mean, whereas for a positive skewness the opposite is true. Similarly, a surface with a low kurtosis has a larger number of local maxima above the mean as compared to that with a high kurtosis. Note that we defined these parameters for a 2D profile, but they can easily be generalized for a 3D surface [30, 32].

The cumulative probability distribution function, $P(h)$ associated with the random variable $z(h)$, is defined as the probability of the event that $z(x) < h$, and is written as

$$P(h) = \text{Probability } (z < h). \tag{2.6}$$

It is common to describe the probability structure of random data in terms of the slope of the distribution function, known as the probability density function (PDF) and given by the derivative

$$p(z) = \frac{\mathrm{d}P(z)}{\mathrm{d}z}. \tag{2.7}$$

The integral of the PDF is equal to $P(z)$, and the total area under the PDF must be unity [30, 32].

In many practical cases, the random data tend to have the so-called Gaussian or normal distribution with the PDF given by

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-m)^2}{2\sigma^2}\right), \tag{2.8}$$

where m is the mean and σ is the standard deviation. For convenience, the Gaussian function is often plotted in terms of the normalized variable $z^* = (z - m)/\sigma$ as

$$p(z^*) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^{*2}}{2}\right). \tag{2.9}$$

The Gaussian distribution has zero skewness $Sk = 0$ and kurtosis $K = 3$ [32] (Fig. 2.5).

The Gaussian distribution is found in nature and in technical applications when the random quantity is a sum of many random factors acting independently of each other. When an engineered surface is formed, there are many random factors that contribute to the roughness, and thus in many cases roughness height is governed by the Gaussian distribution. Such surfaces are called Gaussian surfaces.

In order to represent spatial distribution of random roughness we use the autocorrelation function, defined as

$$C(\tau) = \lim_{L\to\infty} \frac{1}{\sigma^2 L} \int_0^L \big[z(x) - m\big]\big[z(x+\tau) - m\big]\,\mathrm{d}x. \tag{2.10}$$

The autocorrelation function characterizes the correlation between two measurements taken at the distance τ apart, $z(x)$ and $z(x + \tau)$. It is obtained by comparing the function $z(x)$ with a replica of itself shifted for the distance τ. The function $C(\tau)$ approaches zero if there is no statistical correlation between values of z separated by the distance τ; in the opposite case $C(\tau)$ is different from zero. Many engineered surfaces are found to have an exponential autocorrelation function

$$C(\tau) = \exp(-\tau/\beta), \tag{2.11}$$

where β is the parameter called the correlation length or the length over which the autocorrelation function drops to a small fraction of its original value. At the distance β, the autocorrelation function falls to $1/\mathrm{e}$. In many cases the value $\beta^* = 2.3\beta$ is used for the correlation length, at which the function falls to 10% of its original value [30, 32].

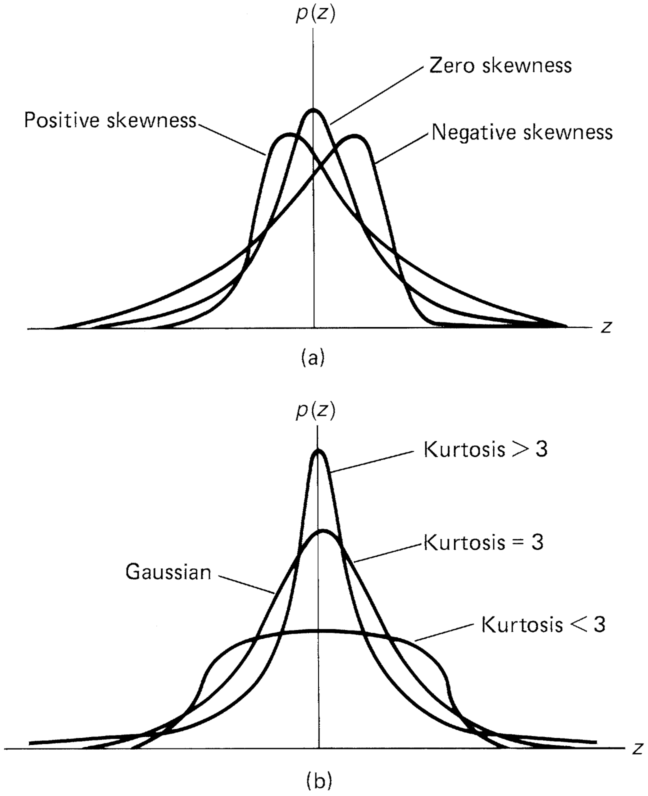


Fig. 2.5. Typical **a** skewness and **b** kurtosis [32]

For a Gaussian surface with the exponential autocorrelation function, σ and β^* are two parameters of the length dimension which conveniently characterize the roughness. While σ is the height parameter that characterizes the height of a typical roughness detail (asperity), β^* is the length parameter that characterizes the length of the detail. The average absolute value of the slope is proportional to the ratio σ/β^*, whereas the average curvature is proportional to β^*/σ^2. For a Gaussian surface, σ is related to the RMS as $\sigma = (\sqrt{\pi/2})R_a$ [30, 32]. These two parameters, σ and β^*, are convenient for characterization of many random surfaces. Note that a Gaussian surface has only one inherent length scale parameter, β^*, and one vertical length scale parameter, σ, and thus it cannot describe the multiscale roughness.

2.3 Fractal Surface Roughness

A measurement of the roughness parameters, such as σ and β^*, shows that they are sensitive to the scale, that is to the resolution of a measuring device (the sampling interval or the short wavelength limit) as well as to the scan size (the long wave-

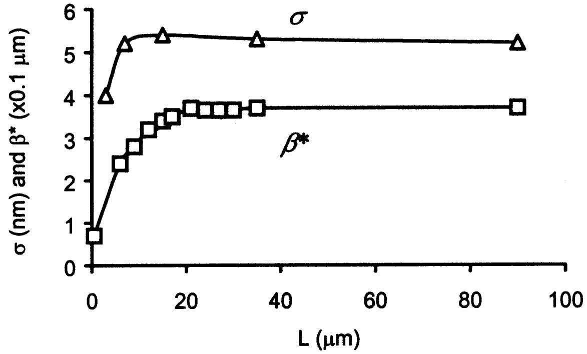


Fig. 2.6. Dependence of measured σ and β^* upon scan size L for a glass disk (based on [268])

length limit) (Fig. 2.6). And understandably so, since the roughness is composed of many wavelengths superimposed upon each other which all affect the cumulative values of σ and β^*, and the wavelengths smaller than the sampling intervals or larger than the scan size cut off and do not contribute to the roughness parameters [268]. Thus, the measured roughness parameters depend upon the short- and long-wavelength limits. This consideration is not only an artifact of the measurement or a result of the measuring devices' limitations. For practical contact problems, asperity may be defined as a roughness detail that participates in the contact and forms a contact spot. Therefore, the size and length of the contact spots are important for the contact of rough surfaces and may provide wavelength limits relevant for the contact problem.

A surface is composed of a large number of length scales of roughness that are superimposed on each other. The variances of surface height and other roughness parameters depend on the resolution of the roughness measurement instrument. As the resolution increases, more small details of the rough profile can be observed. When a rough surface is repeatedly magnified, increasing details of roughness are observed down to nanoscale. The roughness at all magnifications appears quite similar in structure. Such self-affinity can be characterized by fractal geometry.

Archad [12] suggested we present a rough surface as one covered by asperities of a certain size, which have much smaller asperities on the top of them and even smaller asperities on the top of those. He showed that an elastic surface with such a hierarchical structure, which is similar to fractal geometry, leads to an almost linear dependence of the real area of contact with a flat upon the normal force. This, along with the linear proportionality of the friction force to the real area of contact due to the adhesion, could explain the well-known linear proportionality of the friction force to the normal load.

Self-similar curves and surfaces have been studied by mathematicians since the first half of the 20th century. A remarkable property of these curves and surfaces is that they have a fractional "dimension," D, in a sense that when the linear scale is magnified by a certain factor α, the length of the curve or the area of the surface changes proportional to α^D. This is because more fine details are observed with the

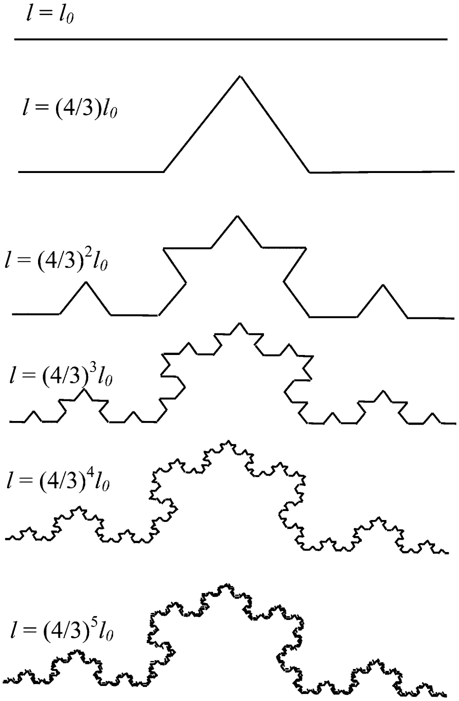


Fig. 2.7. The Koch curve with fractal dimension $D = 1.26$. The curve is built by an iterative procedure, and at each step its length l is increased by the factor $4/3$. If the linear length scale l_0 is increased by 3 times, the total length is increased by $4 = 3^D$

magnification. Thus, when the so-called Koch curve (Fig. 2.7) of length l is magnified by the factor $\alpha = 3$, its length becomes equal to $4l$. Thus, the fractal dimension, $D = \ln(4)/\ln(3) = 1.26$, is between 1 and 2. Unlike most mathematical functions used in engineering, the fractal curves do not have a derivative at any point. Although self-similarity implies equal magnification in all directions, the term self-affinity has a broader meaning and implies that a curve can scale in a certain manner during magnification.

In the 1970s, the term “fractal” was introduced and the concept of self-similar and self-affine objects was widely popularized. It was recognized that fractal geometry could be applied to various physical phenomena, ranging from the coastal line of oceans to the turbulent flow in fluids. The fractals were thought to be a universal tool which could be applied in the situation of noncontinuum behavior that cannot be studied by the continuum functions of traditional calculus.

Since the 1980s, it was suggested that fractal geometry can be applied for the characterization of rough surfaces in tribology (Fig. 2.8) [119, 120, 205, 212, 213, 215]. Majumdar and Bhushan (1990) suggested that the Weierstrass–Mandelbrot self-affine function captures significant features of a self-affine rough profile

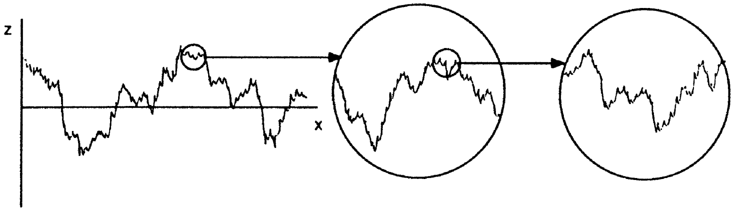


Fig. 2.8. Self-affinity of a surface profile [32]

$$z(x) = G^{(D-1)} \sum_{n=n_j}^{\infty} \frac{\cos 2\pi \gamma^n x}{\gamma^{(2-D)n}}; \quad 1 < D < 2; \ \gamma > 1, \tag{2.12}$$

where D is the fractal dimension, G is a nondimensional scaling constant, and γ^n determines the frequency spectrum of the profile roughness. Nondimensional D and G with the dimension of the length are two parameters that characterize a fractal profile. Ganti and Bhushan [120] extended that analysis and considered the lateral resolution of the measuring instrument as an intrinsic length unit. This generalized analysis allows surface characterization in terms of two fractal parameters—fractal dimension and amplitude coefficient—which, in theory, are instrument independent and unique for each surface. Ganti and Bhushan [120] developed a technique for the simulation of fractal surface profiles. A number of engineered surfaces were measured to validate the generalized fractal analysis, in particular, magnetic tapes, thin-film rigid disks, steel disks, plastic disks, and diamond films, all of varying roughness. For a given surface with varying roughnesses, the fractal dimension essentially remains constant, while the scaling constant varies monotonically with variance of surface heights (σ^2) for a given instrument. Simulated σ shows similar trends in the measured σ for small scan lengths. The coefficient of friction of all surfaces has reasonable correspondence with the scaling constant.

In practice, the profile demonstrates self-affine behavior down to a certain scale length (e.g., of the molecular scale) or a high frequency (short wavelength) limit, ω_h. With a further magnification of the profile, no self-similarity can be found. In a similar manner, there is a low frequency (long wavelength) limit, ω_l, of the fractal behavior [212]. Note that a fractal profile has no characteristic parameters of the length scale. However, short and long wavelength limits effectively provide such parameters of the length dimension, $1/\omega_\mathrm{h}$ and $1/\omega_\mathrm{l}$. During the contact of two rough surfaces, relevant parameters—such as the number of asperity contacts and the real area of contact—depend upon the short- and long-wavelength limits as power functions with power exponents depending upon D and G.

2.4 Contact of Rough Solid Surfaces

When two rough surfaces come into a mechanical contact, the real area of contact is small in comparison with the nominal area of contact, because the contact takes place only at the tops of the asperities. For two rough surfaces in contact, an equivalent rough surface can be defined for which the values of the local heights, slopes, and local curvature are added to each other. The composite standard deviation of profile heights is related to those of the two rough surfaces, σ_1 and σ_2 as

$$\sigma^2 = \sigma_1^2 + \sigma_2^2. \tag{2.13}$$

The composite correlation length is related to those of the two rough surfaces, β_1^* and β_2^*, as

$$1/\beta^* = 1/\beta_1^* + 1/\beta_2^*. \tag{2.14}$$

Using of the composite rough parameters allows us to effectively reduce the contact problem of two rough surfaces to the contact of a composite rough surface with a flat surface [30, 32].

Two parameters of interest during the elastic and plastic contact of two rough surfaces are the real area of contact, A_r, and the total number of contact spots, N. In most cases, only the highest asperities participate in the contact. This allows us to linearize the dependence of A_r and N upon the roughness parameters during the elastic contact as

$$A_\mathrm{r} \propto \frac{W\beta^*}{\sigma E}, \tag{2.15}$$

$$N \propto \frac{1}{\sigma\beta^*}, \tag{2.16}$$

where W is the normal load force and E is the composite elastic modulus. Qualitatively, the higher the asperities, the larger σ is and the smaller A_r is; the wider the asperities, the larger β^* is and smaller A_r is. The larger and wider the asperities, the smaller A_r is [44].

For plastic contact, N, which depends upon the contact topography and thus is independent on whether the contact is elastic or plastic, is still given by (2.16) for a given separation between the surfaces [43], whereas the real contact area is found by dividing the load by the hardness

$$A_\mathrm{r} \propto W/H. \tag{2.17}$$

For fractal surfaces, the roughness and contact parameters are related to the high and low frequency limits as [212]

$$\sigma \propto \omega_\mathrm{l}^{(D-2)}, \tag{2.18}$$

$$A_\mathrm{r} \propto \frac{\omega_\mathrm{l}^{(2-D)/2}}{\omega_\mathrm{h}^{D/2}}, \tag{2.19}$$

$$N \propto \omega_\mathrm{l}^{3(2-D)/2}\omega_\mathrm{h}^{D/2}. \tag{2.20}$$

2.5 Surface Modification

As discussed in the preceding sections, surface properties including the topography have significant effect upon the mechanical contact. There are many ways to modify surfaces in order to obtain desirable properties. Two such methods are surface texturing and layer deposition.

2.5.1 Surface Texturing

Since most engineered and natural surfaces are rough, it may be advantageous not to stay with the random roughness, but to texture a surface in a certain manner so that the useful properties of the surface, such as load capacity, low friction, and wear, improve. Surface texturing has became an object of intensive study in the recent decade [28, 191]. Various techniques are used for surface texturing, including machining, ion beam texturing, etching, lithography, and laser texturing. The texturing usually produces a large number of microdimples on a surface that are effective in combination with lubrication. The dimples can serve as microhydrodynamic bearings, reservoirs for lubricant, or traps for wear debris [106]. Surface texturing is commonly used in magnetic storage devices [27, 28] and microelectromechanical systems (MEMS) to prevent adhesion and stiction (sticking of two components to each other due to adhesion) [36, 38]. It is also used in the automotive industry to hone cylinder liners. At this point, most studies in the area of texturing are experimental and concentrate on finding the optimum size and distribution of the dimples. Thus, Hsu and coworkers [335] investigated the effect of dimple size (of the order of dozens of microns) and depth (below one micron) on sliding friction under boundary lubrication conditions. They found that, for a constant dimple surface, smaller and shallower dimples are more advantageous.

Fabrication techniques for creating micro/nanoroughness include lithography (photo, E-beam, X-ray, etc.), etching (plasma, laser, chemical, electrochemical), deformation, deposition, and others.

2.5.2 Layer Deposition

Thin, artificially deposited layers of long-chain molecules can be used to lubricate microdevices. Such monolayers or thin films are commonly produced by the so-called Langmuir–Blodgett method and by chemically grafting the molecules into self-assembled monolayers. In the Langmuir–Blodgett method, a monolayer is formed at a liquid-air interface and then deposited upon the substrate, to which it is bonded by weak van der Waals forces. Self-assembled monolayer (SAM) molecules attach to the substrate by chemical bonds [267].

Besides the SAM method, there are several other techniques of deposition, including adsorption, dip coating, spin coating, anodization, electrochemical deposition, evaporation, plasma, etc.

2.6 Summary

Since fractals have been introduced into surface mechanics [213], the argument continues over whether fractal geometry provides an adequate description of physical phenomena and scaling issues. Interestingly, one of the creators of the classical Greenwood and Williamson [137] statistical model of the surface published an "apology," recognizing that a fractal description is needed instead [138]. Indeed, many rough surfaces demonstrate self-affine properties to a certain extent and at a certain range of scales. Fractals as mathematical objects obviously have a certain beauty and give us a tool to describe noncontinuous phenomena; these features attracted many physicists and other scientists. However, it is questionable whether the fractal description, which ultimately assumes that a rough profile is characterized by only two nondimensional parameters, D and G, can provide more practical information for the analysis of engineering surfaces than traditional statistical characterization. The generalized analysis by Ganti and Bhushan [120] provides an extension of the Majumdar–Bhushan model for tribological applications; however, practical usefulness of fractal analysis in tribology remains the subject of an argument. An ideal fractal surface is composed of roughness at different scales, but it does not possess parameters of length scale. Unlike the ideal surface, a real fractal surface has such parameters, $1/\omega_{\mathrm{h}}$ and $1/\omega_{\mathrm{l}}$. The contact parameters calculated from the fractal models of surfaces, given by (2.19)–(2.20), depend upon ω_{h} and ω_{l}, which, in fact, characterize *limits* of fractal behavior, rather than the fractal behavior itself.

It is important to note that the statistical description of a rough surface provides some parameters of the length scale, for example, σ and β^*. However, single length and height parameters do not provide an adequate description of multiscale surfaces that involve several scale lengths. While the roughness parameters provide only a constant scale length, we have observed in this chapter two different types of scale dependence. One is the dependence of the roughness parameters upon the scan size, as shown in Fig. 2.6. This dependence is the measurement artifact and is a result of the measuring equipment's limitations. Another type of scale dependence appears during the contact of rough surfaces. The contact spot's size provides additional length parameters that may interplay with the roughness parameters. For example, if the contact size is smaller than the long wavelength limit of roughness, ω_{l}, the roughness components with larger wavelengths do not contribute to the roughness and contact parameters, effectively changing the latter.