REVIEW ARTICLE

Surface metrology

D J Whitehouse

Department of Engineering, University of Warwick, Coventry CV4 7AL, UK

Received 14 October 1996, in final form 21 May 1997, accepted for publication 21 May 1997

Abstract. Some important types of instrumentation for measuring surfaces both past and present are reviewed. Exhaustive lists of instruments and performance are not presented; rather more emphasis is placed on the philosophy of measurement. An attempt is made to classify the surface features and also the function of surfaces as a pre-requisite to measurement. It is revealed that, as the push towards miniaturization is being taken beyond the nanometrology scale, some theoretical restrictions are likely to be encountered.

1. Why measure surfaces?

In recent years surface texture has been recognized as being significant in many fields. In particular the surface roughness is an important factor in determining the satisfactory performance of the workpiece, in tribology for example or in coatings. Also in engineering applications the surface roughness has been found useful in machine-tool monitoring. These aspects will be discussed presently. It is, however, pertinent to consider how the importance of surface roughness is changing with the passage of time and how the importance of roughness depends on the actual scale of size of the workpiece and the process used to make it.

In very general terms the requirements for energy transfer and information transfer and storage have dominated the development of technology. This will no doubt also be true in the future but the factors governing such transfer depend themselves on size. As objects get smaller changes emphasizing the importance of the surface take place. The energy and force equations experience a change of balance, figure 1(a), as the scale of size of the moving object decreases. In rotational situations the dependence on scale is even more pronounced. For example the moment of inertia decreases by a factor of $L^5$.

Of the above factors, the elastic properties can be controlled by varying the properties of materials, which are well understood, and inertial forces can be controlled by design. This leaves areal or damping forces which cannot be controlled easily. Energy losses and transfer and information storage and transfer are all sensitive to uncontrolled areal properties and the areal property which is most likely to be influential is the surface roughness.
It is also the least understood and manageable. Hence the greater emphasis being placed on its measurement in recent years. The pressure to do so is due to the search for a better quality of goods and also to achieve better manufacturing control.

Surface metrology is as much concerned with the nature of the surface and its use as it is with the practical aspects of measurements, so some clarification of surface use will be given as a pre-requisite to measurement. Sections 2 and 3 describe some of the uses of surface measurement and the type of parameter used. There is no description of parameters as such in which follows but more a justification for the measurement. Also topics such as measurements of roundness and cylindricity have been left out because these techniques have the same general problems of data collection as does roughness. They differ in the way the pick-up is presented, and moved relative to the surface. They have been described in detail in the references, for example Whitehouse (1994).

Certain topics have deliberately been left out in this review. These are points concerned with data processing such as filtering methods, parameter characterization, digital data analysis and also methods of calibration. These omissions should not be taken to indicate their lack of importance; all of these points are necessary in order to validate the values of any parameter measured. However, the emphasis here has been placed on illustrating the reason for the use of techniques. All the other points have been covered comprehensively elsewhere, for example by Whitehouse (1994).

2. Surfaces and their importance in manufacture

There are two basic ingredients involved in manufacturing a workpiece, the manufacturing process and the machine tool or production technique. How they relate to surface measurement is shown in figure 2. At one time measurement of the surface was considered largely irrelevant but it soon became apparent that the finish on the surface was extremely sensitive to any changes in the process. Hence it became logical to assume that measurement of the surface could be used to control the process of manufacture. The argument was that, if the surface parameter being measured remained constant from workpiece to workpiece, then the process must be under control. Any change in the surface parameter should initiate a review of the process parameters. In the UK and USA (Page 1948) the average roughness \( R_a \) was used as the control parameter whereas in Germany and USSR peak parameters were used (Schlesinger 1942, Schorsch 1958). The UK approach was pragmatic in the sense that the parameter specified on the drawing had to be measurable. In Germany the approach was to use parameters such as \( R_v \), the maximum peak-to-valley height on the surface, in an attempt to impose functional constraints on the surface as well as manufacturing control. The peak parameters, however, are inherently divergent—they become larger as the sample size increases. Also, sometimes, the maximum values of a parameter are difficult to find over a large area of the surface. Extreme peak–valley measures soon degenerated into measurements of average peak–valley heights \( R_{TM} \), \( R_z \) and so on simply in order to make them reliable.

Although the use of a single parameter of the surface roughness could be used to indicate a change in the manufacturing process, it is not sufficiently discriminating to pinpoint where the changes in the process have occurred. Even using a number of simple parameters rather than just one fails to provide the necessary discrimination. It is only recently that surface metrology has become comprehensive enough to be used as a diagnostic tool. This capability arose because of the advent of random-process analysis. That is, the use of autocorrelation, power spectra and probability density functions.

These are functions rather than numbers such as the \( R_a \) (average roughness) of the profile and so can reveal much more of the underlying statistics of the surface. They are more reliable because in the case of autocorrelation and power spectral density any random phase shifts between the sinusoidal components making up the profile are eliminated. The autocorrelation function is particularly useful for looking at random surfaces and the power spectrum is more useful for looking at periodic surfaces, as will be seen shortly. Neither is particularly difficult to generate. The autocorrelation function is simply a plot of the correlation coefficient between the surface profile and the same profile shifted in space by a set amount. The power spectrum is the Fourier transform of this (see for example Papoulis (1965)). These were introduced because of the needs of functional prediction (to be discussed later) but have now been incorporated into the manufacturing control problem (Peklenik 1967).

2.1. Autocorrelation and manufacture

These statistical parameters are effective because they provide a large enhancement of the signal over the noise introduced into the system. For example, each point on the autocorrelation function of a profile taken from a surface is a result of a great deal of averaging. Small changes between surfaces became significant. As a general rule the autocorrelation function can best be used to reveal changes in random processes such as grinding whereas power
Figure 3. The significance of autocorrelation in manufacturing.

Spectral analysis can be used to best advantage in processes which are fundamentally periodic or repetitive, such as in turning or milling. Both the autocorrelation function and the power spectrum hunt for the unit machining event. In the case of grinding the unit event is the impression left on the surface by an average grain on the grinding wheel. In power spectral analysis it is the periodic signal left on the surface by a clean cutting tool on a perfect machine.

Take grinding, for example, as a case in which autocorrelation is useful. Figure 3(a) shows the impression left on the surface by a sharp grain (Whitehouse 1978). Figure 3(b) is a typical profile and figure 3(c) is the correlation function. Notice that the correlation length (the distance over which the correlation drops almost to zero) is a direct measure of the effective grain hit width. For a grinding wheel in which the grains are blunt, figure 3(d), there is a considerable piling up of material as well as the formation of a chip. By examining the correlation function it is apparent that the piling up or ploughing is revealed by lobing in the autocorrelation function. The width of the central lobe is a measure of the amount of material removed. At a glance, therefore, the shape of the autocorrelation function reveals the efficiency of the grinding in all its aspects, figure 3(g). Notice that this would not be revealed by looking at the profile or by using simple parameters. In figure 3 any longer waves in the autocorrelation function of the surface show that there are other problems such as the need to dress the wheel.

2.2. The power spectral density in manufacturing

Another example shows how the power spectrum can be used to identify problems in turning. As the tool wears and the machine tool deteriorates significant changes occur in the spectrum of the surface, as shown in figures 4 and 5.

Figure 4(a) shows a profile of turning produced by a good tool, together with its spectrum. As would be expected for good turning, the spectrum shows some line frequencies, the fundamental corresponding to the feed and a few harmonics due to the shape of the tool.

As the tool wears the ratio of the harmonic amplitudes to that of the fundamental increases (see figure 4(c)). This is due to the imposition on the surface of the wear scars on the tool. Also, on this right-hand side of the fundamental spectrum, the base line can rise, due to random effects of the chip formation and microfracture of the surface. See figures 4(b)–(e).
To the left-hand side of the fundamental wavelength there appear periodicities whose wavelengths are much greater than that of the fundamental. These are due to machine-tool problems such as bearing wear, slideway error or even lack of stiffness in the machine tool itself, which may cause chatter. Identifying these effects by using the surface texture is an important first step in remedying the problem.

The spectrum can therefore be split up into two parts, one to the right-hand side of the fundamental frequency and one to the left, figures 5(a)–(e). On the right-hand side there appear process problems and on the left, in the sub-harmonic region, machine-tool problems. These advances in machine monitoring and diagnostics stem from the realization that the surface generated by the manufacturing process constitutes a very extensive data bank of information. The surface is in effect a fingerprint of the manufacturing process.

3. The surface and function

The surface is obviously important in many practical situations. This has been known for years. The problem is that of knowing how important it is. For many years very simple parameters were used to describe the surface and, it was hoped, its properties. These included the average value $R_a$, the RMS value $R_q$ and various peak height estimates. Investigators in Germany and the USSR used peak measurements rather than average values because they argued that peak measurements correlated more with tribological situations than did average values (Perthen 1949). Also the peak measurements of roughness could be measured equally well with optical and stylus methods. This philosophy proved to be practically unsound.

The notion of being able to predict the performance of a workpiece from the geometry of the surface has been attractive for some time. Early investigators used simple models of the surface. These usually involved modelling peaks on the surface as hemispherical spheres scattered on a plane. Then these hemispheres or ‘bosses’ were assumed to be distributed in a random Gaussian way in height (Greenwood and Williamson 1964, Greenwood 1982). This development was closer to real surfaces than previous ones had been but it had the disadvantages that two surface descriptions were needed, one deterministic to describe the shape and size of the hemispherical ‘peaks’ and one statistical to describe their distribution in space (Archard 1957).

This confusing model was eventually replaced by the random-process model mentioned earlier which was based...
Figure 5. Some space–frequency kernels and their relationship: A, machine; B, process; and C, material properties.

totally on communication theory (Whitehouse and Archard 1970). This allowed all salient features of the surface to be described with one model. This random-process model was and still is considered to be a big breakthrough in surface characterization.

However, current thinking indicates that even this model needs modifying to reflect better the mechanical situation that occurs, for example, in contact where the surfaces contact top down on each other. Another problem which has to be considered is that contact occurs in a parallel mode rather than a serial one. The conclusion has been reached that random-process analysis is adequate for monitoring the manufacturing process, but is rather inadequate for some cases of functional prediction.

Figure 6 shows the type of parameter of the surface that could be useful in various functional situations. The $x$ axis corresponds to the relative lateral movement between two surfaces and the $y$ axis to the inverse of their separation. On the extreme right-hand side is a column indicating the type of parameter which is most significant. These are (i) unit event characteristics, (ii) profile parameters, (iii) areal parameters, (iv) spatially variable characteristics and extremes of distributions and finally (v) defects. It is a fact that often combinations of these parameter types are needed in order to predict the performance of the surface. Figure 6 shows the requirement for surface measurement which in turn decides the specification for the instrumentation in hardware and software. One point to notice is that it is rarely the individual parameter $R_u$ or $R_q$ for example which is important but often the type of parameter. Little or no convincing evidence is available to link very specific surface parameters to function.

The important point to notice in figure 6 about the surface geometry characteristics important in function is their diversity. They fall into two basic sets. One is the statistical type required, of which there are a number of options. The questions which have to be asked concern whether an average value, an extreme value, the presence or absence of a feature or perhaps even the spatial variability is required. The other basic issue is that of whether the feature of importance is basically height orientated, in which case profile information will often suffice, or areal. It should be noticed that areal information reveals structure which is most often important when movement is involved. Under these circumstances it is often extremes of parameters rather than averages which are important.

Although profile-height information has been available for many years with stylus instruments (Reason et al 1944) it has been only recently that areal information has been recognized as important. Serious attempts are now being made to integrate all aspects of areal (sometimes called 3D) measurement (Stout et al 1994). The reason for this swing to areal measurement is that the possibility of using surface geometry for functional prediction is now recognized. Previously, control of manufacture could be achieved reasonably well with just profile information.

One factor which has been omitted from figure 6 is the importance of physical parameters such as nano-hardness and elasticity. Also missing is the presence of thin chemical films on the surface. These factors are known to be important. These non-geometrical properties of the surface should also be measured and included in the surface characterization if a true estimate of the surface performance is to be obtained. There is evidence that multidisciplinary parameters are now being incorporated into some instruments (Marti et al 1990). However, their use is non-standardized and at present unsatisfactory calibration methods have held back their general implementation.

It should be emphasized that these non-geometrical parameters refer not to bulk properties of the materials but
rather to the properties of the outermost skin of the surface, where the skin thickness is taken to be of the same scale as the roughness, if not smaller. The reason for this distinction between bulk and skin properties is primarily that all the action, that is energy transfer, takes place at the surface boundary rather than in the bulk of the material (Gant and Cox 1970).

Whether asperities are crushed or deform elastically is very dependent on the skin property. Apparently soft materials such as copper have a considerably harder surface skin than had previously been thought, which is why stylus measurement is possible despite having stylus pressures greater than the nominal hardness of the material. It is when trying to measure surface parameters which are significant in fatigue or corrosion that the greatest problems arise because in these examples of function it is the isolated deep scratch or defect which often initiates the breakdown. Finding such cracks over a wide area on the surface is a real problem, requiring an instrument to cover a wide area yet with high resolution.

It seems obvious from that which has been reported above that surfaces can be very important. The question is that of how to measure the surface most effectively.

4. Conventional instrumentation

4.1. Introduction

The earliest ways of measuring surfaces were using the thumb nail and the eye. Both of these are highly effective but completely subjective. Demand for quantitative results led to the development of two parallel branches of instrumentation: one following the tactile example of the nail, the other mimicking the eye.

As will be described in that which follows, the two methods actually evolved to measure different things. The optical methods looked for lateral structure, namely spacings and detail in the plane of the surface, whereas the stylus method examined heights in the plane perpendicular to the surface. Optical methods were developed to help the metallurgist or biologist whereas the stylus method was for engineers’ use.

4.2. Stylus methods

The stylus method essentially uses a caliper, the two arms of which touch a reference surface and the surface under test respectively. The arm towards the test piece ends

---

**Figure 6.** Surface geometry and function.
with a diamond stylus whose tip dimension is such that it can penetrate the detailed geometry of the surface (Reason 1970).

The other arm contacts a reference surface, figure 7(a) and (b) by means of another stylus. All surface-measuring instruments have this basic configuration in one form or another. In some cases it is difficult to spot the reference. One such variant has both caliper arms contacting the surface, as shown in figure 7(c). Here the ‘skid’ technique provides an ‘intrinsic’ reference. In other words the surface itself generates the reference by having one stylus much blunter than the other. The blunt stylus integrates the surface. It acts as a mechanical filter having a lower cut-off than that of the sharp stylus, see figure 8.

The actual position of the reference relative to the test surface is not important; it is not the position or dimension which is being measured but rather the deviation from an intended surface shape. There are three basic trends in conventional stylus instruments. One is the increasing use of transducers which have a large range-to-resolution value. Another is the use of areal scanning and the third is the coming into being of a generation of hand-held instruments.

Simple statements like the above are not sufficient to describe the advantages and disadvantages of the various instruments so various merit criteria have been devised. One can be considered to be mainly applicable to research instruments, (figure 9) and the other to industrial uses (figure 10).

The former wavelength–amplitude method (Stedman 1987) shows the limits of the amplitude which can be measured and the limits on spatial wavelengths that are imposed by such features as the physical dimension of the stylus tip and its angle. The size, shape and position of the envelope reveal the usefulness and application of the particular instrument. Obviously the best instruments on this diagram encompass a wide area. Instruments suited to nanotechnology work best near the origin whereas instruments more suitable for general engineering would cover larger amplitudes and wavelengths and not necessarily be concentrated near the origin. Figure 9 shows a typical plot for some nanotechnology instruments. It shows how, for example, the effect of the stylus can be taken into account.

The range-to-resolution response due to Whitehouse takes into account the dynamics of the measuring system, which is important if fast scanning speeds are to be achieved. This is required when areal scans are needed or when the instrument is being considered for in-process measurement. The horizontal axis is also important because it determines the types of surface feature which can be measured with one trace. One example is shown in figure 11, which shows an arcuate profile. If the range to resolution of the instrument is high then both the radius and the roughness can be measured at the same time. Good modern instruments have values of 10^6:1. Table 1 gives some typical values for instruments. The actual performance values shown in figure 9 and table 1 are continually changing as the technology develops. It is the general pattern which remains virtually the same. Other advances in the dynamics of stylus instruments include the use of optimum damping.

It has recently been realized that there is a more realistic way of assessing the performance of instruments. Conventionally the response of the instrument to sinusoidal inputs had been measured and recorded as the transfer function. This method is misleading because most surfaces have significant random features or sharp edges, both of which have a broad-band frequency component rather than a single-wavelength sine wave. It is possible (Whitehouse 1988) to utilize this fact by optimizing what is in effect an integrated transfer function, the inverse of which is shown in figure 12. The instrument’s performance can thereby be optimized over the whole frequency band. Figure 12 shows the relative importance attached to each component in the signal. It can be seen that all signals can have the same importance (1.0) right up to the resonant frequency, at which \( w/w_n = 1 \) for a damping ratio of 0.59 except for a small amount which is averaged out near the resonant frequency.

It also follows that the reaction on the surface caused by the stylus pressure can be made substantially constant on average over the total frequency range. This approach has produced the interesting conclusion that the traditional damping factor used for instruments is too low and that the optimum damping should be 0.59, which is close to the
Figure 9. Instrument criteria: P, profile; NP, profile; SIP, slope integration; N/T, Talysstep; and EM, electron microscope.

Table 1. The range/resolution response for typical methods. (Note that ultrasonic pneumatics are suitable for rougher surfaces only.)

<table>
<thead>
<tr>
<th>Method</th>
<th>Spatial resolution</th>
<th>Z resolution</th>
<th>Range z</th>
<th>Frequency</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stylus</td>
<td>0.1 μm to 1 mm</td>
<td>0.3 nm</td>
<td>50 μm</td>
<td>20 Hz</td>
<td>Contacts workpiece; easy to use; traceable</td>
</tr>
<tr>
<td>Optical probe</td>
<td>0.5 μm</td>
<td>0.1 μm</td>
<td>20 μm</td>
<td>10 Hz to 30 kHz</td>
<td>Non-contacting; less traceable; with servo drive; range extended to 50 μm</td>
</tr>
<tr>
<td>Heterodyne</td>
<td>2.5–200 μm</td>
<td>0.2 μm</td>
<td>0.5 μm</td>
<td>10 Hz</td>
<td>Requires computer unravelling</td>
</tr>
<tr>
<td>TIS</td>
<td>0.6–12 μm</td>
<td>1 nm</td>
<td>0.1 μm</td>
<td>Seconds</td>
<td></td>
</tr>
<tr>
<td>Scatterometer</td>
<td>≃ 10 μm</td>
<td>1 nm</td>
<td>λ/100 nm</td>
<td>Seconds</td>
<td>Resolution depends on aperture; insensitive to movement</td>
</tr>
<tr>
<td>TEM</td>
<td>2 nm to 1 μm</td>
<td>2 μm</td>
<td>100 nm</td>
<td>Minutes</td>
<td>Replication needed can destroy surface</td>
</tr>
<tr>
<td>SEM</td>
<td>10 nm</td>
<td>2 nm</td>
<td>2 μm</td>
<td>Minutes</td>
<td>Vacuum needed</td>
</tr>
<tr>
<td>STM</td>
<td>2.5 nm</td>
<td>0.2 nm</td>
<td>100 nm</td>
<td>Minutes</td>
<td>Vibration-free mounting</td>
</tr>
<tr>
<td>Normarsky</td>
<td>&gt; 0.5 μm</td>
<td>1 nm</td>
<td>100 nm</td>
<td>Minutes</td>
<td>Needs interpretation and certain minimum reflectivity</td>
</tr>
<tr>
<td>Capacitance</td>
<td>2 μm</td>
<td>1 nm</td>
<td>50 nm</td>
<td>2 kHz</td>
<td>Needs conductors</td>
</tr>
<tr>
<td>Interferometry</td>
<td>2 μm</td>
<td>1 nm</td>
<td>1 μm</td>
<td>Minutes</td>
<td></td>
</tr>
</tbody>
</table>

Figure 10. Instrument criteria.

critical damping. This development allows higher speeds to be obtained without the stylus lifting off the surface. Another example of modern developments is the use of neural networks to help in optimum measurement. One example is the use of neural networks to ‘de-convolute’ the stylus shape from the profile measurement (Wang and Whitehouse 1995). The result of these advances is that the stylus technique is moving in a direction which takes in the advantages of both axes, namely faster measurement yet a larger range-to-resolution ratio too.

The methods outlined above illustrate advances in integrated and high-precision measurement using the stylus technique. Attempts have been made to adapt the stylus method to in-process measurement or at least to reduce the measurement cycle time. One example of stylus in-process measurement is shown in figure 13 (Deutschke et al 1973).

This is comprised of a drum which is forced onto the workpiece. A single probe projects through the drum wall to contact the test surface. This probe has a displacement transducer connected to it within the drum. As the
4.3. Other (non-optical) methods

Other possible methods for measuring surfaces are numerous but on the whole they are matched to a particular process or workpiece shape. One consideration is that, for the transduction to have a high signal-to-noise ratio, a high change in energy per unit displacement of the transducer is required.

One technique which has this high energy is the pneumatic method (Von Weingraber 1942). In this air is blown on to the surface via a measuring head and skirt. The amount of air which escapes can be related to the surface finish. In practice the leakage is proportional to the average distance between the skirt lip and the rough surface. One possibility is to measure the flow of air; the other is to measure the back pressure. The latter is most sensitive, but neither is highly sensitive compared with other methods. Range-to-resolution ratios of about 100:1 are common, which is rather poor. However, they are cheap and robust. Difficult shapes cause problems because the shape of the skirt has to match the general shape of the surface. Also the air supply has to be dry otherwise water condenses out of the air on to the surface.

Another possibility is the use of capacitance methods (Perthen 1936). One of these uses an electrode positioned above the surface. The capacitance between it and the surface is a measure of the roughness and the average distance from electrode to surface. Some dielectric such as air has to be between the two.

If capacitance is to be used in this way there are some complications because the value of the capacitance is not simply related to the surface roughness. In fact it follows an inverse law with nonlinear terms. The capacitance method can also be used in scanning mode but it is no real contender.

In general the capacitance method is very sensitive and can be used effectively in certain applications. However, it does have problems, one of which is that the signal from the transducer is susceptible to extraneous noise. So for high sensitivity shielding has to be used. Also the electrode shape should follow the general shape of the workpiece being measured. It should be remembered that the capacitance method integrates the surface roughness (Sherwood and Crookall 1968). It has therefore two attractive features; it measures over an area and it is a non-contacting method. Scanning methods are currently being considered but they do not have the same potential for versatility as do the stylus and optical methods (Matey and Bland 1985).

A further possible technique for measuring surfaces is to use ultrasonics (Berry 1973). In this method ultrasonic waves are projected on to the surface and the scattered waves picked up. In principle this method has potential because it is possible to get phase as well as amplitude is measured. Over many revolutions a picture of the surface is built up, at the same time that the workpiece is being machined. Problems with the method such as debris and surface damage caused by the drum rims limit its usefulness.
information, in contrast to optical methods, with which phase information is lost unless recovered indirectly.

Some snags with this method are that very high (unrealistic) frequencies are needed to measure the roughness on fine surfaces and also that ultrasonic waves only propagate through air with difficulty so that the attenuation rate is high. The directionality can also be poor. Another wave technique is that using optics.

4.4. Optical methods

Some optical methods mimic the eye whereas some mimic the profile produced by a stylus instrument. Such is the potential of optical methods that the variety of possibilities is large. In that which follows the simplest method will be considered first; then the more complicated methods will follow.

All optical methods involve projecting light on to a surface. Sometimes the light is focused but often it is not. Also the light can pass by an obstruction on the way to the surface or after scattering. In addition the light can be used coherently or in an incoherent mode and finally polarization properties can be used (Beckmann and Spizzichino 1963, Bass and Fuchs 1963, Ogilvy 1991). More practical assessments by optical means have been treated thoroughly in the books by Bennett (1994) and Whitehouse (1996).

Schmaltz in 1927 was the first person to use optical techniques (Schlesinger 1942, Schmalz 1929). He projected a collimated beam of light past a knife edge, the shadow of which was made to intersect the surface at an angle. When viewed from a different angle a profile of the surface roughness was apparent. In this method there is no horizontal magnification. The vertical magnification is provided by the obliquity angle of the projected edge. The magnification is only about $\times 20$ so that the method is restricted to viewing rough surfaces. The roughness measurement is made by eye using the graticule in the viewing optics.

The other simple way is to view the scattered light directly or, more practically, sense the scattered light with one or more detectors. This is the basis of the gloss meter or scatterometer (Whitehouse and Bowen 1994, Young et al 1980).

Two detectors A and B pick up the scattered light (figure 14). The detector A is positioned at the specular (reflected) angle. In addition another detector B is used, which is called the diffuse detector. The surface roughness can be estimated from the ratio of the scattered light which hits the two detectors. For a rougher surface whose $R_a$ of texture is about half the wavelength of light the ratio is about one half and for very rough surfaces there is no specular component and the light detected at A is equal to that at B; that is, the surface number is zero. This scattering-based number can be used as a basis for surface quality measurement or quantification. The problem is that the light-scattering characteristic as a function of roughness varies with the manufacturing process so that the method cannot be recommended for general use (Wang and Wolfe 1983). The scattering method can only be used as a comparator. However, if the surface is very fine all the light scattered can be collected by an integrating sphere. This total light scattered can be related to the RMS value of the surface (figure 15) (Beckmann and Spizzichino 1963).

In a more refined version of the scattering method coherent light is used (figure 16) and the light scattered from the surface is collected in the Fourier plane of the collecting lens, namely the back focal plane.

In this method the light source is imaged in the Fourier (transform) plane (Konczakowski 1983). The image in the transform plane is modulated principally by the slopes on the surface. Under certain conditions, for example, when the surface roughness is small compared with the wavelength of light, the image in the transform plane can be taken as the power spectral density of the surface. This particular configuration has enormous advantages as a metrological tool. The first is that small detail tends to scatter over a very wide angle. This produces a signal on the detector which is far from the centre so that a large measurement on the detector corresponds to a small distance on the surface; this is a metrological freak, figure 16. The situation is usually the other way round; that is, in the image of the surface, small detail in the object produces small detail in the image.

From the power spectral density it is possible to evaluate the surface roughness and the various moments of the spectrum leading to slope and curvature information.
Because this method is essentially a slope-sensitive one, the scattering pattern is insensitive to movement of the surface. A point to notice is that, because of this independence of speed the method lends itself to in-process gauging. Furthermore, it allows random-process analysis, that is, power spectral analysis, of the surface geometry to be carried out (Welford 1977).

In another mode which utilizes the coherent property of the light the interference between light scattered from the test surface and that scattered from a reference surface such as a glass flat (Tolansky 1948) produces fringes. The reference flat can be positioned as in a Twyman Green interferometer or a multiple-beam interferometer by laying it on the surface (figure 17). Fringe contours can then be examined. The contour pattern when viewed can give a very clear picture of the general surface geometry. If the fringes are sharpened up by means of multiple-beam interferometry it is possible to consider the roughness.

If the reference surface is kept at a small angle to the optical slab and a tilted reference is used, the fringes are sharpened up by means of multiple-beam interferometry (figure 18). The fringe pattern is stored in a computer and then some subsequent movements are made between the objective and the test surface (hence the reference) and the test surface (Wyant 1975). The total movement is restricted to about one wavelength. The form and roughness of the surface are then computed from the sets of stored data. This is a good example of how computation can help out in a difficult metrology situation. There are still problems, however; if the surface is complicated, the computer can get confused! For reasonably flat surfaces such as semi-conductor wafers this method works admirably. Incidentally the instrument design has to be good so that when the small movements in the objective position are made there is no twisting or yawing.

White light can be used in some cases to generate fringes of equal chromatic order. This technique (FECO) produces coloured fringes close to any one diffraction order. This method relies upon multiple reflections and can be used for very fine surfaces (Hodgkinson 1970). Obviously, complicated roughness or form cannot be examined by this method.

One large group of optical methods involves scanning as part of the technique. The obvious starting point is the flying-spot microscope (Young and Roberts 1951). In this very high spatial resolution is obtained by restricting the illumination on the surface to a very small spot. Light collected at this time can obviously originate from the illuminated point on the surface only, which has the effect of not allowing extraneous light to hit the detector.

A variation on this is the confocal microscope (Sheppard and Wilson 1978) in which the imaging and receiving optics are exactly the same. Like in the flying-spot method, the optics is used to produce a spot on the surface. The reflected light is picked up by a point detector. The resulting signal is used to modulate the brightness of a spot on a TV screen. The point detector is in effect a pin hole which completely blocks out all light from an out-of-focus plane which could otherwise mix in with the signal (figure 19).

As the object is defocused the light signals fall off dramatically. This method produces very sharp images. Resolutions of 1 µm in x and y and 0.1 µm in depth are possible. The absolute roughness is found by measuring the axial movement needed to keep the spot focused.

Optical ‘follower’ methods utilize the null technique outlined above. Essentially the sensor follows the surface during a scan, maintaining a fixed distance from it. Some optical means has to be adopted whereby the de-focus of the optical system is detected and an error signal generated which corrects the focus. The first attempt at this was due to Mme Dupuy who used the Foucault knife-edge test as a basis for the focus-correction mechanism (Dupuy 1968). In this configuration (figure 20) a knife edge is situated at the position where the surface spot re-focuses in the image plane. Further from the objective two photodetectors are positioned laterally and symmetrically relative to the optical axis. When, during the scan, the spot on the surface is higher than the in-focus plane due to a peak, the de-focused image causes a difference between the signals on the two detectors. This references the spot via the objective. If a valley is encountered exactly the opposite occurs.

The movement of the objective lens to maintain focus is measured and used as the surface profile. The
most successful optical methods are heterodyne techniques utilizing common-path interference in which two different types of illumination are projected on to the surface at the same time.

Invariably the light is split either into two different wavelengths or into different polarizations, (Sommargren 1981, Simon 1970). The resulting two beams of light are focused at different axial or transverse positions. It is possible to follow the surface geometry by observing how the roughness affects the two positions where focus occurs for the different rays (figure 21). Some of the very finest surface measurements have been achieved using these methods. Differential interference microscopy, such as by Nomarski methods (figure 22) (Tolansky 1948), is used to examine fine surfaces but, as the name implies, the technique is essentially a differentiator. This has the effect of enhancing edges and generally distorting heights. Such techniques are best used for visualization rather than to obtain quantitative information. There are other optical techniques which have not been examined in detail here. These include speckle techniques (Erf 1974) and holographic techniques. They have been useful occasionally for roughness measurement but not generally so. As versatile tools for metrology they are somewhat lacking (Whitehouse 1996).

The problem is that the value of the roughness is obtained indirectly rather than directly as in the stylus method. In methods such as those using speckle the roughness value is inferred from the speckle contrast; in other words some model of the surface has to be assumed before the roughness value can be obtained; the credibility of the result can depend heavily on the validity of the model. For this reason speckle and similar techniques, particularly those based on scattering, have been restricted largely to experimental instruments.

For form and vibration measurements these methods have had more success (Ribbins 1974, Jones and Wykes 1983). There are some problems, one of which is that follower methods tend to be relatively slow. More
importantly, problems can arise because the surface reacts to different polarizations in different ways which do not correspond to the geometrical features of the surface. This problem is common. In principle the mechanical stylus and optical methods need not agree when measuring fine surfaces. The former measures mechanical geometry whereas the optical methods measure the optical path length or the optical path difference. If any thin films are present they will contribute to the value of the roughness measured optically. Such films can be ignored by stylus instruments. Another problem with optical methods is that sharp edges in the surface can produce misleading diffraction spikes which can be mistaken for real peaks. This effect can be troublesome in calibration. Hence the practical situation is that optical methods tend to enhance the noise whereas stylus instruments tend to reduce the signal (the fine detail of the geometry is integrated). The best result is somewhere in between. For a given signal-to-noise ratio the stylus reduces the signal, and the optics increase the noise.

5. Unconventional methods

The natural constraints apparently governing the resolution of surface measurement have been the wavelength of light and the rather blunt stylus used in tactile instruments. Recent advances in scanning microscopy have largely removed these restrictions. Before examining these methods their forerunner the scanning electron microscope (SEM) will be considered.

5.1. The SEM

The basic idea of the SEM is simple (figure 23). Electrons from a filament are accelerated by a high voltage between two or three magnetic lenses. These cause the electron beam to strike the surface in a raster scan. The incident electrons produce amongst other things secondary electrons which are detected. The detected signal is synchronized with the scan to produce an image of the surface on a cathode-ray tube (CRT). The SEM has no imaging lenses in the true sense of the word. The image magnification is determined solely by the ratio of the CRT scan to the surface scan. This invariably means that low magnifications are more difficult to produce than are high magnifications because of the large scanning area in the former case. One of the big advantages of the SEM is the large depth of imaging available because of the very low numerical aperture of the electron beam and the small equivalent wavelength of the electron beam. Despite the various options in display the SEM has its drawbacks. One is its rather poor lateral and vertical resolution and there is also some distortion produced by sharp edges in the object. This makes calibration a problem.

It is also ironic that its large depth of focus can cause ambiguities, as can be seen in figure 24. The large depth of focus allows large features to be measured. The secondary electron output is influenced not only by the slope but also by the curvature. As a result of this, edges become enhanced. Although this sometimes clarifies the picture and makes visualization easier, it is qualitative in nature and does not help the calibration. In fact a trace taken through a SEM picture correlates better with the differential of a profile taken by a stylus instrument than it does with the profile itself! Nevertheless the SEM has been and is extremely useful as a surface tool. It constitutes sound metrology philosophy in the sense that, for measuring to nanometres and below, a metrology unit, namely the electron wavelength, is of the same order of size as the features being measured (Suganuma 1985).

Various methods have been tried to quantify the picture. One such method is to use more than one detector to get a stereoscopic view of the surface. Another variant is to look at back-scattered electrons at the specular angle. There are many options for measuring surfaces using scattering of electrons or other particles and waves (figure 25). Most of these have been used to highlight different aspects of the surface, see for example Whitehouse (1996).

5.2. The TEM

Transmission electron microscopes (TEM) are not as important in surface examination because most engineering surfaces are opaque. However, information can be obtained if replicas are used. One such application is in measuring diamond styli. In this the stylus is indented into glass which is coated with carbon and shadowed with platinum at an angle. Asperities on surfaces can be highlighted in the same way. Other replication materials are plastics such as cellulose acetates which are pressed on to a surface previously coated with a solvent such as acetone.
5.3. The STM and the AFM

The routine study of atomic and molecular scale structure has become possible with the development of the scanning tunnelling microscope (STM) and the atomic force microscope (AFM) (Binning et al. 1982, 1986, Gehtz et al. 1988, Binning 1992). These are similar to the conventional tactile surface instrument in that there is a stylus.

The difference is that contact with the surface is not a pre-requisite. In fact gaps of the order of nanometres have been claimed. These techniques do not measure the topography necessarily; other features such as the charge density are revealed in the case of STM and forces in the case of AFM. In all these cases the information is obtained via a stylus. The great advance in technology is that these instruments are not bounded by conventional restrictions such as diffraction caused by the relatively large wavelength of light. The limit is imposed by the geometrical size of the stylus probe.

These can now be made extremely small in tip dimension. Values of a few nanometres are not uncommon. The other big difference between the scanning probe microscopes (SPM) and conventional instruments is that they rely for their signal on quantum-mechanical effects (tunnelling currents in the STM, for example). This probabilistic nature of the new instruments is inevitable when exploring surfaces on the atomic scale. Differences between these methods and the more conventional ones are to be expected (Bhushan 1990).

A typical configuration is shown in figure 26. It is basically a very simple device comprising of a pick-up and a translation stage. There are two modes of operation, open loop (figure 26(a)) and closed loop (figure 26(b)). In the latter the signal current is kept constant throughout the traverse by means of a servo system. The other mode simply moves the probe in a fixed horizontal plane. Figure 27 is an atomic eye view of what the stylus records in the cases mentioned above.

In the AFM there is an added advantage in that the surface being measured does not have to be a conductor. It simply responds to atomic-scale forces by the deflection of a simple cantilever. The movement of the cantilever was measured originally by a STM (figure 28) but is now usually measured optically. The method whereby the deflection is converted into an electrical signal is not at issue here.

One advantage of this new generation of microscopes is their ability to measure more than one feature. For example, by varying the voltage between the specimen and the probe in the STM different physical phenomena of various substances can be investigated on the atomic scale. The instrument can in effect be used both as a spectroscopic and as a topographer. The STM provides spectroscopic data by utilizing its sensitivity to the electron energy states of the sample. Measurements of the tunnelling current made under conditions of changed bias voltage can provide data on the local density of states.

In another mode (figure 26) changing the height of the tip and measuring the tunnelling current can give information about the local work function. The STM in spectroscopic mode has a considerable advantage over conventional spectroscopic techniques in that the results are not averaged. Figure 29 is a schematic diagram of a typical STM.

The versatility of instruments like the STM has allowed many different surface experiments to be carried out. These include the measurement and detection of growth in biological specimens, the initiation of some chemical reactions, the machining of substrates and the microfabrication and movement of material around the
surface. Also other physical properties can now be examined using this basic technique. (Wickramasinghe 1989). The techniques are:

(i) magnetic force microscopy (MFM), in which the probe is magnetized;
(ii) electrostatic force microscopy (EFM), in which the probe has an electrostatic charge;
(iii) scanning thermal microscopy (STM), in which the probe is designed as a thermocouple;
(iv) scanning conductance microscopy (SICM), in which a micropipette probe containing an electrode is used; and
(v) near-field scanning optical microscopy (NSOM), in which the probe is in fact a submicrometre aperture (Guerra 1990).

5.4. X-rays

The trend towards atomic measurements and in particular various atomic measurements on the surface has pointed towards the use of x-rays (Hogrete and Kunz 1987, Wormington et al 1994). This makes sense because the wavelength of x-rays is about the same as the atomic lattice spacing. Recent work has shown that by varying the angle of incidence, the surface interface can be assessed, at least crudely (figure 30). In the same measurement an estimation of crystal flaws and residual stress is also possible. It is in effect an extension of the optical scattering methods.

6. Trends in metrology

Figure 31 shows the trends in surface metrology in recent years. The $z$ axis shows the scale of size under consideration in terms of resolution. The $x$ axis refers to the lateral information on the surface and the $y$ axis to the height information. Both $x$ and $y$ are scaled chronologically with the origin at 1950. The scale of size in $z$ is the resolution of a typical instrument.

There are three curves in figure 31 each representing a family of instruments. The left-hand curve refers to stylus instruments; its position to the left-hand side of the graph reflects the fact that, during the early stages of development, stylus instruments measured height almost exclusively. Optical methods, shown by the right-hand curve, on the other hand, were almost completely concerned with lateral structure and spacings. The middle curve represents the other, newer, techniques such as those involving the scanning microscopes.

This graph has been described as showing the ‘horns of metrology’ (Whitehouse 1988, 1991). The ‘horns’ are the curves of the stylus and optical methods which have been converging in their measurement philosophy: both are now attempting to measure height as well as lateral structure.

Other changes in the philosophy have to be recognized. These changes are concerned with the physical basis of the measurement. At the normal engineering scale of size, called macro in figure 31, measurement is deterministic; distances rely on readings of scales, fringes and so on. At the sub-micrometre level that which is measured in positioning, for example, relies on statistical averaging to produce very high resolution; distance is the separation of two planes, each determined by the average position of millions of atoms. This regime is called here statistical mechanics. Measurement to sub-nanometre accuracy in two dimensions is possible using statistical mechanics. Problems arise, however, when atomic accuracy in all three dimensions is required. This accuracy can only be achieved with a closed-loop system; there has to be evidence of proximity between the two points in space for which this accuracy is required. (Notice that this type of measurement cannot refer to planes or lines.) However, because of the limited number
of atoms involved in the points, it is difficult to get a signal between them. Quantum mechanics does not readily allow it! The loop closure is not possible on a continuous basis.

The only way to achieve a measurable signal strength is to reduce the measurement bandwidth; that is, the signal has to build up in time. The actual value of the signal
is determined by quantum mechanics which determine the point-to-point tunnelling. The probabilities only convert to a measurable signal after a considerable time lapse; hence all effects of external vibration and environment will be included in the measurement, so the signal-to-noise ratio will increase rather than decrease with time. This effect constitutes an uncertainty principle in metrology between time and spatial 3D discrimination, which has yet to be solved? Compare this restriction with that of signal to noise in the stylus and optical methods mentioned earlier.

7. Summary

Measurement techniques for surfaces have progressed rapidly during the last decade. Traditional complementary techniques such as the stylus and optical methods are converging to satisfy areal and vertical measurement requirements. This is not to say that they are competing against each other. Contact and non-contact methods both have advantages.

Contacting methods clear unwanted debris away. Also, there is the possibility of measuring the nano-hardness and even the friction at the nano-scale with tactile stylus methods.

Optical and ultrasonic methods can be quicker because they do not contact the surface. Also, neither does stylus damage to the surface occur nor is there a change in resolution due to stylus damage. Optical techniques are still plagued by the lack of suitable standards but these are now being introduced by standardization committees.

Calibration is still a problem with the newer generation of scanning microscopes. At the same time as making atoms visible the new technology has revealed a challenge to metrology in terms of achievable signal levels.

References

Beckmann P and Spizzichino A 1963 The Scattering of Electromagnetic Waves from Rough Surfaces (Oxford: Pergamon)
Bennett J 1994 Introduction to Surface Roughness ISBN 1-557521085
Berry M V 1973 The statistical properties of echoes diffracted from rough surfaces Phil. Trans. R. Soc. A 273 611
Binning G 1992 Ultramicroscopy 42/44 7
Gant N and Cox J M 1970 The hardness of metals at very low loads Phil. Mag. 22 891
Greenwood J A and Williamson J B P 1964 The contact of nominally flat surfaces Burndy Res. Rep. 15
Marti O et al 1990 Topography and friction measurement on mica Nanotechnology 2 141
Matey J R and Bland J 1985 Scanning capacitance microscanning J. Appl. Phys. 57 1439
Ogilvy J A 1991 The Theory of Wave Scattering from Random Rough Surfaces (Bristol: Hilger)
Page F S 1948 Fine Surface Finish (London: Chapman and Hall)
Perthen J 1936 Ein neues Verfahren zum Messen der Oberflächengute durch die Kapazität eines Kondensators Maschinenbau. Beitr. 15 669
——1949 Prüfen und Messen der Oberflächengestalt (Munich: Carl Hanser)
Reason R E, Hopkins M R and Garrod R I 1944 Report on the measurement of surface finish by stylus methods, Rank Organisation
Schlesinger G 1942 Surface Finish (London: Institute of Production Engineers)
Schmalz G 1929 Z. VDI 78 144–61
Schorsch H 1958 Gutebestimung an technischen Oberflächen (Stuttgart: Wissenschaftliche)
Suganuma T 1985 J. Electron Microsc. 34 328
Tolansky S 1948 Multiple Beam Interferometry of Surfaces and Films (Oxford: Oxford University Press)
Van Weingraber 1942 Pneumatische Mess- und Prüfgüte. Machinebau 121 505
——1994 Handbook of Surface Metrology (Bristol: Institute of Physics)


Young J Z and Roberts T 1951 The flying microscope Nature 167 231