Paradigm shifts in surface metrology.
Part I. Historical philosophy

By X. Jiang1,*, P. J. Scott1,3, D. J. Whitehouse1,2,3 and L. Blunt1

1 Centre for Precision Technologies, School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK
2 School of Engineering, University of Warwick, Coventry CV4 7AL, UK
3 Taylor Hobson Ltd, PO Box 36, 2 New Star Road, Leicester LE4 9JQ, UK

Surface texture and its measurement are becoming the most critical factors and important functionality indicators in the performance of high precision and nanoscale devices and components. Surface metrology as a discipline is currently undergoing a huge paradigm shift: from profile to areal characterization, from stochastic to structured surfaces, and from simple geometries to complex free-form geometries, all spanning the millimetre to sub-nanometre scales.

This paper builds a complete philosophical framework for surface metrology through a review of the paradigm shifts that have occurred in the discipline of surface metrology, tracing the development of fundamental philosophies and techniques. The paper starts with a brief overview of the historical paradigm shifts and builds an up-to-date foundational philosophy, capable of rapid and effective development. The growth in interest in surface metrology stems mainly from the need to control the manufacture of armaments during the Second World War and the production of domestic goods and appliances since that time. The surfaces produced by manufacture seemed to offer the possibility of being useful for process control. Unfortunately, only a few tentative investigations had been carried out to establish usable relationships between the processes, the machine tools and the available surface parameters (with their limitations). Even fewer investigations had been carried out to relate surface geometry to the performance of manufactured products. The result was that the metrology was unprepared and, consequently, the progress was sporadic.

This overall review is given in two parts. Part I focuses on the historical philosophy of surface metrology and Part II discusses the progress within the current paradigm shift.

Keywords: surface texture; measurement; instrumentation; filtration; characterization; standardization

1. Introduction

Surface metrology is the science of measuring small-scale geometrical features on surfaces: the topography of the surface. Traditionally, surface metrology was seen as important owing to its strong association with other disciplines such as...
control of machine tools and manufacturing processes, quality of optical components, tribology and surface engineering in general.

For example, the control of surface texture allows automobile engines to have reduced running-in times and to be more fuel efficient with reduced emissions. It allows orthopaedic implants to last longer through optimized surface topography. It also enables bearings in machines such as hard disk drives to run more efficiently and wear less. In addition, when optical components have smoother surfaces they scatter less light and have better optical qualities. The usefulness of surface metrology is not in doubt: it has been viewed as a means of obtaining a fingerprint of the surface to characterize and understand how the unique surface topography plays a dominant role in the functional performance of components in a system.

One of the biggest problems for instrumentation was the wide range of texture obtained during manufacture, from milling, which produced sub-millimetre roughness, to polishing, which produced nanometre values. Accordingly, instrumentation had to be versatile, and although it is true that specific instrumentation was being developed, for example interferometric methods for nanometrology and micrometrology, and pneumatics- and capacitance-based methods for rougher surfaces, they could not fulfil the overall requirement. How stylus methods prevailed for general use and how methods were developed for processing the signal are explained in detail in this part of the paper.

The basic thesis of this paper is that new technology has led and continues to lead to the paradigm shifts in surface metrology. What is meant here is that technical developments not explicitly concerned with surface metrology (in sensors, fast and low-cost computing, mathematical concepts, etc.) have prompted disruptive but beneficial shifts leading to a fundamental change in the approach to surface metrology. This form of paradigm shift corresponds to that of Kuhn (1970) in that the shift results in a change of exemplar for the field. The science of surface metrology has followed this path and during its lifetime as a discipline, there have been three historical paradigm shifts together with the current paradigm shift, which are as follows.

(i) **Instrumentation.** The first instruments permitted quantitative assessment of surface texture as opposed to qualitative assessment based on experience.

(ii) **Digital methods.** Digital computation facilitated the application of digital theory to surface texture analysis and enabled signal processing methods, like random theory in discrete form, to be used.

(iii) **Form and texture instruments.** Separation of form and texture mathematically rather than mechanically, permitting the application of surface texture assessment to non-planar and non-circular geometries.

The current paradigm shift in surface metrology is expected to have the greatest impact of all. Unlike the previous shifts, it is not based on a single technological development. The current shift can be categorized as follows:

(i) profile to areal characterization,

(ii) stochastic to structured surfaces, and

(iii) simple geometries to complex free-form geometries.
This paper has two parts. In Part I, all three historical paradigm shifts are described in detail, together with the subsequent technological evolutions. In Part II (Jiang et al. 2007), the authors describe the current shift and discuss the progress made and the problems yet to be resolved. The overall aim of the paper is to provide a philosophical framework for the further development of the discipline of surface metrology.

2. Historical philosophy

(a) Background: general

Surface metrology, although a relatively new discipline, has had a tortuous and complex development. There have always been several basic issues needing to be addressed when surface metrology has been applied, and they are as follows.

(i) Is the surface geometry important functionally?
(ii) Can surface metrology be used as a simple quality control tool in industry?
(iii) Can surface metrology lead to a deeper understanding of surface interaction and the consequent function?
(iv) Is current surface metrology adequate?

Throughout the development of surface metrology, these issues have been interwoven and have caused considerable confusion because their objectives are different: for instance, should surface metrology be used to gain a deeper understanding, or should it be used as a certification tool in the context of quality control? In what follows, these issues will be explored and new concepts emerge. In measurement technology, in general, there is always a conflict between the need to measure and understand the significance of a measured feature, and the actual means to measure and control it. Surface geometry and its measurement are no exception!

(b) Historical: function

Initially, surfaces were considered important in terms of friction and optical reflection. Much emphasis was placed on friction because it is fundamental in the performance of moving systems. Wear is often the result of too much friction, and lubrication is mainly devoted to reducing friction. Contact mechanisms are fundamental and the geometry of contact depends strongly on the surface topography.

Amonton (1699) and Coulomb (1785) were the first investigators to study friction, although da Vinci (1500) displayed an appreciation of its consequences. The original investigations concentrated on finding the relationship between the load and the area of contact. Amonton could not understand why the force required to move one block of material over another in the horizontal plane was, for a given mass of the upper block, independent of the apparent contact surface area. He was not to know that the actual area of contact was much smaller (by factors of order 1000) than the apparent contact surface area. The actual area of contact is determined by the surface topography.
Two historical factors influenced the general interest in surface geometry in England. One was the need to make accurate cannon and pulley blocks for ‘ships of the line’ of the Royal Navy (sixteenth and seventeenth centuries). The other was the need to make accurate measuring machines to help build machine tools for the Industrial Revolution in Great Britain from the mid-eighteenth to the mid-nineteenth centuries. The requirement for cannon was a smooth, circular and straight bore. In addition, it was found that adding a spiral groove to musket barrels improved their accuracy. This rifling of the barrel is probably the first ‘structured’ surface of modern technology. The only way to be sure of roundness was to rotate a shot in the bore to see where the gaps were; it was known at the time that, since it is easier to make a near-perfect sphere than it is to make a cylinder, the sphere proved to be the more accurate gauge. It has to be remembered that although roundness, straightness and flatness are the result of different aspects of manufacturing, they still constitute surface geometry, albeit on a different scale.

3. Paradigm shift: first instrument

(a) History: instrumentation

For many years, the only way to clarify surfaces was by using the thumbnail and the eye, both of which were subjective but effective only if used by an experienced practitioner. These rudimentary methods, however, illustrated a very basic problem in surface geometry measurement, namely whether to measure normal to the surface or across it. This dilemma even applies today for modern contact instrumentation and for some optical instrumentation. One point is certain: some form of instrumentation was needed because the textural features of the surface were too small to be assessed quantitatively by ‘touch and see’. Hence, the emphasis shifted to ways of magnifying geometrical features to the extent that they could be seen, not using a microscope, which magnified laterally, but using ‘instruments’ that magnified normal to the surface.

Measurements made normal to the surface were the highest priority because in industry the height of surface roughness features affected the size assessment and the tolerance requirement in the assembly process. In addition, the geometry of roughness measured normal to the surface was deemed to be most important in contact situations. At this time, lateral measurements were usually taken using microscopes to assess grain size and other metallurgical properties.

One of the first methods was devised by Tomlinson (1919) at the National Physics Laboratory. He succeeded in using a galvanometer incorporating a mirror to give a magnification of approximately 30×. However, Schmalz (1929, 1936) was first in seriously considering instrumentation, using both optical and tactile methods of measuring surfaces. His methods, although original, were not capable of high magnification. His optical method, for example, simply projected a line of light across the surface at an angle $\theta$; the magnification became $\csc \theta$ when the strip of light was viewed normal to the surface.

At this time, manufacturers considered optical methods as being too sensitive to be used near machine tools. So, around 1933, simple tactile methods began to be used, namely a stylus on a beam connected to a transducer, amplifier and meter.
It is probable that Abbott et al. (1938) of Surface Physics Ltd, Detroit, MI, USA was the first to make serious use of such an instrument. The rudimentary transducer was based on a ‘hot wire’, which was connected to a meter that read nominally the root mean square (r.m.s.) value of the signal current. The roughness was estimated by looking at the fluctuating meter pointer. The reading obtained was difficult to verify due to the fluctuations. In addition, outputting the signal onto a chart hardly helped. It was soon realized that the magnitude of the signal relative to a mean line to give an average value was preferable. An estimate of magnitude was made by dividing the hot wire output by the factor 1.11 and calling the result the average roughness, designated by arithmetic average (AA). Unfortunately, the factor 1.11 is not generally accurate except for a sine wave, as approximated generated by machine turning. For a random wave, typical of machine grinding, the factor should be 1.25. In the UK, a rectified moving coil system for the transducer provided an average roughness value called centre line average (CLA). Without the rectification, the roughness signal on the chart recorder could be assessed by using a planimeter to measure areas above and below the mean line. Validation of the average value was then possible.

A few years earlier, there was a feeling among engineers that ‘smooth’ surfaces were ideal and that ‘rough’ surfaces were poor. This view came to a climax in around 1930 when Bentley cars of UK began to manufacture very smooth cylinder bores. The result was that in the Le Mans 24-h race their car engines seized, giving the clearest indication that smoothness did not imply good performance. This incident and similar ones alerted the manufacturing community to the importance of the surface finish and stimulated the development of instruments.

As a result, a number of companies began to develop instruments in the second half of the 1930s, including Bendix and Brush in the USA, as well as Taylor Hobson in the UK, and Perthen and Hommel in Germany. The Talysurf 1 was the first commercial instrument and also the first to include a chart showing a roughness profile, although Brush first used the term ‘profilometer’. It should be mentioned that not all of these instruments used the stylus method as a basis; Perthen used capacitance and Nicolau in France used pneumatics.

(b) Early characterization

There were still, however, some troublesome issues. Engineers began to think about the quantification of surface texture by means of surface texture parameters. The original idea was that a single number should be used to represent surface texture, with the scale for that number corresponding to an interval ranging from ‘goodness’ to ‘badness’. Other people regarded this concept as simplistic; to what parameter or surface attribute should the number be attached? Consequently, the profile itself began to be regarded as the method of characterization. The reason behind this view was that a magnified representation of the surface (relative to a nominal, smooth surface) could be shown on the chart. Judgements concerning the surface could be made from the chart. There were thus two extreme camps: one representing the surface by too little data—the number, and the other representing the surface by too much data—the profile.

It was left to one man, Dr Abbott, to clarify and rectify the situation: Abbott & Firestone (1933) suggested a simple curve to represent the surface, from which realistic numbers could be determined depending upon the surface application.

Furthermore, he made contact the basis for the curve, constituting a plot of the material to air ratio of the surface as a function of depth.

This inspired idea was originally given the name the Abbott–Firestone curve, as shown in figure 1, and for the first time, linked function, that is performance based on contact, to numbers simple enough to control manufacture. This idea proved to be a fundamental step forward because it could be linked also to the statistical descriptions of the surface, which arose later.

The two basic problems with this ‘material ratio’ curve were, however,

(i) it did not convey any spatial information about the surface and
(ii) it should have started at the highest peak of the surface (which cannot be found in practice).

The first problem was serious because the curve could not be used as a basis for separating spatial characteristics of the surface and the meter reading associated with them. For this a profile graph had to be used.

At this time (1940–1950) some very important books were published, which brought together the subject as a more formal discipline. These were the work by Schlesinger (1942, UK), Reason et al. (1944, UK), Perthen (1949, Germany), Schorsch (1958, Denmark) and Page (1948, UK).

(c) Mean line, envelope and standard filters

From these books and other sources, various components of texture were identified, namely form, waviness and roughness arising from design, the machine tool and the manufacturing process, respectively. Attempts were also made to decompose the components according to wavelength bands as shown in figure 2. The most difficult of these bands was that of waviness because it has no natural wavelength boundaries, either high or low pass, whereas roughness and form have the probe size and part size, respectively.

The 1950s saw two attempts to separate the waviness from the profile so that the roughness could be calibrated. One was graphical, simulating electrical filters in the meter circuit. The other was mechanical, simulating the contact of a converse surface, e.g. a shaft, with the face of the anvil of a micrometer gauge. The latter appeared as a large circle rolling across the profile. The first of these methods proposed by Reason (1961) was designated the mean line system.
(M-system), and the other, the envelope system (E-system) was proposed by von Weingraber (1956) in Germany. Neither was entirely satisfactory because in the M-system the graphical procedures did not accurately depict the electrical filters in the system, and in the E-system no practical instrument using mechanical filters could be made. Arguments about the relative methods were sometimes bitter, but around 1960, with the advent of practical computer systems, the M-system became pre-eminent, for the following two reasons.

(i) An exact simulation of the filtering used in the M-system was made by Whitehouse & Reason (1963).

(ii) Theoretical improvements to the analogue and digital filters were also made by Whitehouse (1967/68).

Basically, for (i), the 2CR network (figure 3) could be represented by the impulse function of equation (3.1). Note that the two CR circuits are buffered...
(shown by the dotted line),

\[ h(t) = \delta - \frac{1}{RC} \left[ 2 - \frac{t}{RC} \right] e^{-t/RC}, \quad (3.1) \]

\[ h(t) = \delta(t - t_0) - \frac{\omega_c}{\pi \sqrt{3}} \exp \left( - \frac{\omega_c}{\sqrt{3}} |t - t_0| \right), \quad (3.2) \]

\[ H(\omega) = k(\omega) e^{-j\omega t_0}, \quad (3.3) \]

where \( h \) is the impulse response of the 2CR circuit; \( R \) is resistance; \( C \) is capacitance; \( \omega \) is angular frequency; \( t \) is time; \( t_0 \) is time delay; \( k \) is a constant; and \( H \) is the transfer function.

As these and other filter characteristics could be implemented in digital form, all practical objections to the M-system were dropped.

Equation (3.1) represents a fundamental change in the approach of manufacturers of surface metrology instruments. Up to 1963, the design of instruments was carried out by mechanical engineers who, although very talented, were unable to calibrate the meter or chart. This inability posed a very serious problem which threatened to undermine the excellent mechanical designs. Many procedures were attempted, to no avail. The M-system at that time established a reference by using an ‘averaging box’ that provided a ‘waviness’ line on the profile graph, from which various parameters could be determined. The box function was, however, a poor approximation to the filter. The E-system was another attempt to produce a reference line. Motif methods, described in §4c, were yet another. Although a profile signal could be obtained using, for example, a stylus-based instrument, it could not be processed satisfactorily.

This deadlock was resolved in 1963 by using electrical filter theory developed by Whitehouse and Reason. Although with hindsight this seems to be an obvious step to take, at that time it was a revelation. It enabled the simulation of various types of filter. In particular, with the advent of digital methods, discussed in §4, it became possible to correct faults in the standard 2CR filter. Equation (3.2) illustrates the ‘weighting function’ or impulse response used to correct phase distortion caused by convolving the profile with the standard filter weighting function. This phase-corrected filter is made possible by introducing a delay \( (t_0 \) or \( x_0 \)) to the real-time profile signal. Whitehouse (1967/68) presented the theory behind the phase-corrected filters. The significance of these filter developments is not overstated: it enabled standardization of instrument characteristics to be achieved and provided a traceable path from the stylus to the chart (or computer store) and ultimately to the various National Standards Institutes. It should be pointed out that although the theory used was conventional to electrical engineers, its application by mechanical engineers was not.

The standard filter used in surface metrology has a cut-off transmission value that is not standard. Its value has been changed a number of times. It was given the value 80% in the 1950s by the British. On the other hand, the USA opted for 70.7%, this being derived from the half-power point. These two values were subsequently averaged to give a value of 75%, which was adopted by the International Standards Organization (ISO) for general use. This value has since
been changed to 50% to make the roughness cut-off and the waviness (+ form) symmetrical and complementary. These changes in the cut-off value demonstrate the somewhat arbitrary evolution of the instruments.

4. Paradigm shift to the digital age

(a) Emergence of digital methods

The use of digital methods in surface metrology as outlined in the theory by Whitehouse & Archard (1969) was paralleled in a practical sense by Williamson & Greenwood (1966), who overcame the difficulty of ‘seeing’ between two surfaces in contact by mapping the surfaces digitally and simulating contact using a computer. This simple stratagem had important implications. Visualization of inaccessible and minute geometry interaction was possible from then on. It showed that surface peaks hardly ever contacted the tips of peaks on the opposing surface, but that contact was usually on the shoulders of the peaks.

The demonstration of digital computing in theory (Whitehouse & Archard 1969) and practice (Williamson 1967/68) boosted the subject of surface metrology to an extent never envisaged. Up to the early 1960s, models of surfaces were invariably deterministic, being in the form of simple geometrical shapes such as sine waves or triangular waves. These forms to some extent were imitating machining processes such as turning or milling used to remove material, but not the machining processes used for finishing such as grinding or polishing. However, advances in communication theory by Rice (1944) and others in the 1950s and in oceanography by Longuet-Higgins (1957) enabled researchers such as Nayak (1971) to introduce random surface models into tribology and by Peklenik (1967) into manufacturing. Greenwood & Williamson (1966) inserted random surfaces into contact theory and also made a significant extension to contact mechanism theories by the introduction of the concept of a plasticity index $\phi$, where

$$\phi = k \frac{E}{H} \sqrt{\frac{\sigma}{\beta}}.$$  \hfill (4.1)

This concept brought together the geometrical properties $\sigma/\beta$ of the surface, where $\sigma$ is the r.m.s. value of the roughness and $\beta$ is the average peak curvature with the physical properties of the subsurface $E/H$, where $E$ is the elastic modulus and $H$ is the hardness. Whitehouse & Philips (1978, 1982) introduced the concept of discrete random surfaces in a way that linked for the first time the calculations of tribological parameters carried out in the computer with theoretical parameters of contact.

Formulae (4.2)–(4.4) are the digital forms of some parameters that are useful in contact theory. These include $N$, the probability that a digital ordinate is a peak, where $p_1$ and $p_2$ are the values of the autocorrelation function of the surface, which correspond to $h$, the sample interval length, and $2h$, respectively. The other parameters are the mean peak height as measured from profile information and the mean curvature.

It should be noted that certain features appear in several of the discrete formulae. For example, the probability that an ordinate is a peak designated by $N$ appears in equations (4.2) and (4.3). In other words, the ingredients constituting the parameter are preserved and the physics is retained.
Thus, the peak density $n$ is given by

$$\frac{n}{h} = \frac{1}{\pi} \tan^{-1}\left[\left(\frac{3 - 4p_1 + p_2}{1 - p_2}\right)^{1/2}\right], \quad (4.2)$$

and the mean peak height is given by

$$z' = \frac{(1 - p_1)^{1/2}}{2N}, \quad (4.3)$$

whose mean curvature $C$ is given, where $R_q$ is the r.m.s. roughness, by

$$C = \frac{3 - 4p_1 + p_2}{2Nh^2\left[\pi(1 - p_1)^{1/2}\right]} \frac{R_q^2}{C}. \quad (4.4)$$

Greenwood (1984) approached contact theory in an alternative way, but basically came to the same conclusions. So, for the first time, the necessary ingredients for functional prediction were in place, namely surface metrology instrumentation, tribology, digital theory and random process analysis. Furthermore, although they were available for use for research workers in the late 1960s, it was not until around 1970 that commercial instruments were available. Up to this time, the value of some excellent work had been compromised by a lack of one of the ingredients above. One investigator, Tallian et al. (1964), carried out some careful experiments with bearings, but had limited instrumentation and used the simplest of surface parameter AA, now $R_a$, which omitted vital information. The result of the experiment was inconclusive, so carrying out a good experiment is insufficient. The three main ingredients all have to be present.

(b) Proliferation of parameters

The early analogue instruments had limited parameter options, with the restriction being imparted by the electrical circuit elements. Parameters such as the average roughness value $R_a$ and slope were possible. In some respects, this limitation was useful because only simple parameters could be used for process control and functional prediction, the basic philosophy being that parameters should be proved to be useful before incorporation into instruments. An effect of this limitation was that the bandwidth of instruments could be specified at all stages from the stylus to the pen recorder, as shown in figure 4.

However, one problem with this approach was that a carrier signal was added to the transducer signal to shift the working point to a more linear region. This carrier signal was automatically removed by the low-pass characteristics of the meter unit and the chart recorder. Unfortunately, when researchers attached analogue/digital equipment between the amplifier and the filter, the carrier ruined all signals due to aliasing. A separate low-pass filter had to be inserted as shown in figure 5.

Breaking the signal path left the researchers free to develop software for any parameter. This freedom presupposes that the researcher had expertise in digital analysis, at least in integration, differentiation, interpolation and extrapolation. The problem was that the necessary knowledge was only just becoming available (see Whitehouse (1994) for background).

---

The digital breakthrough occurred when the waviness filter (sometimes called the meter cut-off) could be accurately simulated in the computer. From then on, the chart lost its primary role and the emphasis shifted to the computer output listings. As a consequence, the possibility arose of being able to implement any parameter in software, whether sensible or not, resulting in the ‘Parameter Rash’ (Whitehouse 1982). In effect, the onus on the instrument maker to provide significant parameters was shifted to the researcher or process engineer to develop his or her own ‘preferred’ parameter. So, instead of a single value of $R_a$ for a surface, a whole spectrum of parameters was allowed and the situation became difficult to manage in terms of standardization.

The transformation from analogue to digital systems represents the second paradigm shift. It led to the possibility of matching specific parameters to performance and also the possibility of better process control. Although in principle this transformation was beneficial, it bred some problems. For example, digital simulations began to replace actual experiments. There was and still is the possibility of losing sight of the physics. At about the same time as the digital revolution in surface metrology, two other advances were made. One was the realization that surface texture could be represented not only as a periodic signal.

Figure 4. Schematic of an analogue stylus instrument.

Figure 5. Schematic of an early digital stylus instrument.
derived from turning or milling, but also as a stochastic or random signal derived from abrasive processes like grinding. The other advance was the progress in the theory and practice of tribology, namely friction, wear and lubrication, and contact theory. Taken altogether, the subject became more complete when digital characterization was applied to the random surfaces and phase-corrected filters were used to produce more realistic images of surfaces (Whitehouse 1978a).

At about this time, physicists and mathematicians began to realize that surface metrology and tribology were very fruitful research topics. The result was that not only the number of parameters proliferated, but also the use of somewhat arbitrary mathematical functions. These functions initially were based on profiles or probability density functions of profiles. One of the main aims was to characterize the surface and also to obtain, from the surface, evidence of poor machining or forewarnings of functional disasters.

There are many ways to characterize a profile using a mathematical function. These can have an advantage over previously mentioned filters because there is less run-up, which results in the loss of signal at one or both ends. However, some prior knowledge of the surface is beneficial in obtaining a good match. Functions like Fourier, Legendre, Chebyshev and beta (Whitehouse 1978b) were tried.

(c) Filtration

Filtration has always been important in surface metrology: it is the means by which the surface features of interest are extracted from the measured data for further analysis. As discussed above, there were two competing filter systems: the M-system and the E-system.

The M-system started with the fully analogue 2CR filter implemented as a two-stage capacitor–resistance network as shown in figure 3. This filter could badly distort profile features due to phase shifting of the different harmonic components in the profile. In an analogue embodiment of a phase-corrected system, the signal is passed through a single-stage CR filter network. It is then stored on magnetic tape. The tape is reversed and passed through the filter again. Phase effects are cancelled, but this is not viable as a workshop instrument, mainly because handling the tape in order to turn it around introduced a large delay.

Digitally, the same effect is achieved by convolving the surface signal by a digital filter having an impulse response with even symmetry (Whitehouse 1967/68).

The phase-corrected 2CR still had problems, one being that it badly distorted the profile at the ends. In 1984, the three main surface texture instrument manufacturers (Hommelwerk, Mahr and Taylor Hobson) held a meeting in Zurich, Switzerland. Several alternative phase-corrected profile filters were suggested. By 1986, at a joint meeting held in Hanover, Germany, the three manufacturers had reached a consensus, with the Gaussian filter being chosen as the new filter for separating differing surface wavelengths. This recommendation was adopted by ISO, resulting in ISO 11562:1996 in which the Gaussian filter is given as the standardized phase-corrected profile filter (M-system) for surface texture.

The French automobile industry adopted an alternative approach to filtration called roughness and waviness (R&W). The method began as a purely graphical approach, where an experienced operator would draw on a profile graph an upper
envelope that subjectively joined the highest peaks of the profile. This was an attempt at a simulation of the E-system. The technique was based on the concept of the ‘motif’: a motif is the portion of a profile between local peaks. The basic approach was to determine all motifs between adjacent local peaks; then a series of rules would combine ‘insignificant’ motifs with neighbouring motifs, to create larger combined motifs, until only ‘significant’ motifs were left, from which surface texture parameters could be calculated.

Unfortunately, as any problems with the approach occurred, the rules would be tweaked to overcome them. Even though one set of rules was standardized in ISO 12085:1996, the tweaking of the rules has continued. The main problem is that the method is unstable: small changes in the profile can cause large changes in the resulting motifs. Even though the basic concept is sound, the adopted approach was never based on well-founded mathematics.

In 1996, ISO set up a group, under the convenorship of Scott (ISO/TC 213, 1996) to investigate filtration. This effort was motivated by an urgent need in industry to distinguish between those features in a surface that are functionally important in new applications and manufacturing processes. The Gaussian filter, although a good general filter, is not applicable for all functional aspects of a surface, for example in contact phenomena, where the upper envelope of the surface is more relevant. This work has resulted in the establishment of a standardized framework for filters, giving a mathematical foundation for filtration, together with a toolbox of different filters. Information concerning these filters has been or is about to be published as a series of technical specifications (ISO/TS 16610 series), to allow metrologists to assess the usefulness of the recommended filters. So far, only profile filters have been published, including the following classes of filters.

(i) **Linear filters.** The mean line filters (M-system) belong to this class and include the Gaussian filter, Spline filter (Krystek 1996, 1997) and the Spline-wavelet filter (Jiang et al. 2000).

(ii) **Morphological filters.** The envelope filters (E-system) belong to this class and include Closing and Opening filters using either a disk or a horizontal line. The E-system, which had problems as a purely mechanical approach, can be simulated digitally (Scott 1992; Srinivasan 1998).

(iii) **Robust filters.** Filters that are robust with respect to specific profile phenomena such as spikes, scratches and steps. These filters include the Robust Gaussian filter (Seewig 1999) and the Robust Spline filter.

(iv) **Segmentation filters.** Filters that partition a profile into portions according to specific rules. The motif approach belongs to this class and has now been put on a firm mathematical basis, using pattern analysis, which has solved the stability problem in R&W techniques (Scott 1992, 2004).

(d) **Analysis techniques**

In many instances, the data obtained from a profile are not very convenient because there are too much of them! Characterization and filtering are attempts to reduce the profile to only a few numbers, which should enable the objective to be achieved. For example, the convolution process required for filtering spatially converts to multiplication when Fourier transformed, thereby enabling much
simpler transmission of data through electronic equipment (figure 6). In addition, other simplifying techniques using Lagrangian sampling can be used (Whitehouse 1994).

Transforms can either be time- \((t \text{ or } x)\) or frequency-based. Then,

\[
F(\omega) = \int p(t)\exp(-j\omega t)dt,
\]

\[
p(t) = \frac{1}{2\pi} \int F(\omega)\exp(j\omega t)d\omega
\]

or in series form if the profile is periodic as in turning, where \(F(\omega)\) is the Fourier transform; \(p(t)\) is the profile at time \(t\); \(\omega\) is angular frequency and \(j = \sqrt{-1}\). It is important to realize that if the profile is random, then it is better to investigate it via correlation methods. In particular, the autocorrelation function (figure 7) can unscramble the ‘unit machining event’, i.e. the average impression produced by a grit hitting the surface in grinding or similar processes (Whitehouse 1978a).

Other transforms that have been found to be useful for surface analysis are the Hough transform, Walsh transform, Hadamard transform, and so on (Whitehouse 1994). They all transform completely into either spatial coordinates or frequency coordinates, each of which brings specific qualities to the analysis, but unfortunately none of them is very versatile.
There is, however, a family of functions that are two-dimensional, containing $t$ (or $x$) and $\omega$ as variables. Apart from wavelets, discussed in Part II of this paper (Jiang et al. 2007), the two transforms that are most used are the Wigner distribution and the ambiguity function. The transforms have two arguments rather than one, and link together space $x$ and spatial frequency $\omega$. They can in principle be able to identify freak surface marks as well as average characterizations, hence anticipating wavelet usage as demonstrated in Part II.

The Wigner distribution (Zheng & Whitehouse 1991) is given by

$$W(x, \omega) = \int_{-\infty}^{\infty} f(x + X/2)f^*(x - X/2)\exp(-j\omega X)dX$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} F^*(\omega + \sigma/2)F(\omega - \sigma/2)\exp(j\sigma x)d\sigma. \quad (4.6)$$

Note that equation (4.6) contain elements of both the correlation function, $\int f(x)f(x + \tau)d\tau$, and the Fourier spectrum, $\int F(\omega)\exp(j\omega x)d\omega$. So, these so-called ‘space frequency functions’ should be capable of a much fuller characterization of surface properties than either the correlation function or the spectrum.

The Wigner distribution $W(x, \omega)$ (figure 8) centres on specific $x$ position and $\omega$ frequency and scans via $\chi$ and $\omega$ over all the area. It is in effect a two-dimensional convolution with $\chi$ and $\omega$ as dummy variables. The two-dimensional integral gives complete information about the spatial frequency characteristics centred at $x$, $\omega$.

The moments of $W(\chi, \omega)$ can be used to extract information functions. In addition, the spatial moments provide the position and width of pulses. The Wigner distribution was originally used in quantum mechanics.

Another space frequency function is the ambiguity function given by

$$A(X, \sigma) = \int_{-\infty}^{\infty} f(x + X/2)f^*(x - X/2)\exp(-j\sigma x)dx,$$

$$A(x, \sigma) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega + \sigma/2)F^*(\omega - \sigma/2)\exp(j\omega X)d\omega. \quad (4.7)$$

This is different from the Wigner distribution because here instead of the box function expanding around $x$, $\omega$ as shown in figure 9, the fixed box $\chi$, $\omega$ is moved all over the $x$, $\omega$ field. Hence, the ambiguity function is good for identifying positional information, i.e. properties changing in $x$ or $\omega$. 

Figure 8. Wigner distribution function.
In shifting frequency, the other symbols have been specified. The symbol associated with the Fourier transform is $F$ and with its complex conjugate is $F^*$. Note that the ambiguity function has been used extensively in radar techniques. In equations (4.6) and (4.7), $f$ is the spatial profile.

The various moments of the Wigner distribution and the ambiguity function can be used effectively to investigate the performance of a machine tool. It is possible to detect tool vibration as well as defects in the surface (Zheng & Whitehouse 1991).

Invariably, it is the angular errors, i.e. roll, pitch and yaw, that produce most variation, and spatial angular errors, i.e. straightness along the bed and tool column, that are often ignored. Linear effects $xyz$ are not so troublesome. The possibilities for using space (time) frequency functions are numerous and can be very effective. For example, the Wigner distribution can identify chirp signals on the surface caused by yaw in the tool holder.

(ii) Fractals

The relevance of the Wigner distribution and other space frequency functions derives from the fact that they complement the proven usefulness of the spectral and correlation functions, and yet have an extra factor to deal with the more complicated demands of today. Another possibility is to use fractal analysis, first recognized as being useful in surface topography by Sayles & Thomas (1978), and in more general terms by Russ (1994).

Two parameters, the fractal dimension $D$ and the topothesy $L$, are used to characterize fractals. As in the space frequency functions, fractals can be related to the power spectrum $P(\omega)$ of surfaces. Thus

$$P(\omega) = \frac{A}{\omega^v},$$

where the fractal dimension $D$ is related to $v$ by the relationship $v = 2D - 5$. The essential difference between fractal surfaces and conventional surfaces is that $v$ can be fractional and that there is no breakpoint in the spectrum of equation (4.8) such as occurs in low-pass spectra.

A property of fractals is that the dimension $D$ is continuous between 1 and 2, and surface characterization can be scale-invariant, i.e. have the same characteristics irrespective of scale, as shown in figure 10. In other words,
fractal analysis can reveal the commonality between surfaces of different scale of size. This means that mechanisms of generation can be identified, and the analysis can also be used to identify other factors like fracture mechanisms.

A note of caution needs to be given here because although fractals can be used effectively to describe growth mechanisms (Russ 1994), they are not necessarily suited to describe conventionally manufactured surfaces. Further, they are not particularly suited to describe tribological characteristics such as lubrication or wear, which are not scale-invariant processes (Whitehouse 2001). However, there is growing awareness that very fine surface texture is likely to benefit from fractal analysis, as reported in §2g of Part II of this paper (Jiang et al. 2007).

Overall, the important concept to be extracted from the background to the second paradigm is that the freedom generated by digital methods led inexorably to the use of mathematical tools, capable of processing multidimensional data, having much greater capability to extract unique information when the data are limited. It was the turn of the instruments for the next paradigm shift!

5. Paradigm shift: form and texture instrumentation

The first truly ‘wide-range’ surface texture instrument that was capable of measuring both surface texture and form with a single profile measurement had its genesis in an MSc thesis (Garratt 1977). This fundamental transducer was later commercialized by Taylor Hobson to become the Form Talysurf, which was launched in 1984 (figure 11).

Up to this point, surface texture profile instruments had a very limited usable vertical range, typically 100–300 µm. As a result, surfaces had to be manually levelled to enable the profile to stay within the vertical range of the instrument, which was very time consuming and limited the range of surfaces that could be measured. In contrast, the Form Talysurf had a vertical range of 2 mm (with a resolution of 5 nm).

It was not just the new interferometric transducer that was revolutionary; the Form Talysurf also had enhanced calibration to correct for the nonlinearities in the measurement due mainly to the arcuate movement of the stylus arm. This enhanced calibration was possible due to two developments. One of these developments was that the horizontal traverse direction had a linear grating to ensure accurate traverse positioning (previous instruments relied on timing to determine traverse position). The other development was the provision of mathematical optimization algorithms that could determine the calibration constants automatically, by the measurement of a calibrated sphere, to ensure an accurate Cartesian coordinate system in the measurement plane.

An accurate Cartesian coordinate system allowed the separation of form and surface texture through mathematical algorithms. Previously, only mechanical removal of a straight line and circular form could be achieved. Now, in principle, almost any mathematically defined shape could be removed with the appropriate mathematical algorithm and the resulting residuals characterized (Scott 1983). This advance facilitated precision profile measurement by a single instrument in many new application areas. These areas include Gothic-arc bearings, aspheric lenses and orthopaedic joint replacement; the possibilities were endless.

There was another significant improvement in range and resolution with the development of the phase grating interferometer transducer (figure 12), which was independently developed by Taylor Hobson in the UK (Mansfield & Buehring 1997) and Jiang in China (Jiang 1995). This development has enabled today’s wide-range surface texture instruments to have a range of 24 mm with a simultaneous resolution of 0.1 nm.

6. Lessons from history

This paper has outlined the three historical paradigm shifts in surface metrology, together with the subsequent evolution resulting from the shifts. It illustrates that these shifts enabled a number of types of measurement (surface texture and
form) and analytical methods (digital filtration, autocorrelation and fractals), which are widely used in surface metrology practices. It also illustrates that these shifts enabled other methods (Wigner distribution and ambiguity function), which are used in machine tool applications for precision engineering. These shifts have led to a series of specific lessons being learnt by the surface metrology community.

The basic lessons learnt from the development of surface metrology are shown in table 1; they illuminate philosophical points from history, although these are quite general for measurement science. However, surface metrology has some particular problems, for example, with disruptive technologies: parts are becoming smaller due to miniaturization, and are also becoming more complex. The results of these changes are that a surface signal contains more information with many different components of surface geometry, which have to be unravelled before characterization can begin.

The historical philosophy has also highlighted the fact that paradigm shifts must be robust and flexible. This statement implies that the current paradigm must be capable of responding to the points in table 1, and also be flexible to allow for further development. It must allow for full control of surface manufacture and provide an understanding of the surface functional performance.

Part II of this paper provides a philosophical framework and details of the current paradigm shift, as a ‘stepping stone’, building on the above historical context. However, interpolation from the past to the future can be very uncertain: disruptive technologies that lead to paradigm shifts are very unpredictable (Christensen 1997), making forecasts difficult until the new disruptive technologies are identified. In the future, more aspects of the surface geometry will have to cater for surfaces derived from disruptive application, that is to say, applications that require new technologies and methodologies are to be
developed. Structured and free-form surfaces are identified candidates and will be dealt with in Part II of this paper together with a discussion on further developments in the field of surface metrology.

X.J. gratefully acknowledges the Royal Society under a Wolfson–Royal Society Research Merit Award. The authors gratefully acknowledge the Research Committee of the University of Huddersfield for supporting funds for this work and the directors of Taylor Hobson Limited for permission to publish.

### References


Coulomb, C. A. 1785 Memorie de Mathematique et de Physics de l’Academic Royale, 161.


Garratt, J. 1977 An interferometric transducer for surface metrology. MSc dissertation, University of Aberdeen, UK.


Schmalz, G. 1936 Technische Oberflachenkunde. Berlin, Germany: Springer.

Schlesinger, G. 1942 Surface finish, report of Research Department of Institute of Production Engineers, London.


Scott, P. J. 1983 Multipivotal models with applications to a shape fitting problem. PhD dissertation, Imperial College, UK.


