Resolution and sensitivity in surface plasmon microscopy and sensing

Eric M. Yeatman

Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2BT, UK
Tel: [44] (171) 594 6204 Fax: [44] (171) 594 6308 e-mail: e.yeatman@ic.ac.uk

Abstract: The use of surface plasmons on planar surfaces for the characterisation of those surfaces is analysed. In particular, analytic expressions are derived for the sensitivity of such characterisation to the variation of bulk refractive indices, and to the addition of thin dielectric layers on the plasmon supporting surface. Comparison between surface plasmon and dielectric waveguide techniques is made. Surface plasmon microscopy is then considered, and relationships between sensitivity and resolution are derived. The interaction of plasmons with surface features is considered, and the effects in imaging of step and periodic features are specifically analysed. Other aspects relating to implementation and possible improvements to the technique are discussed.

Keywords: surface plasmon microscopy, plasmons, surface features

1. INTRODUCTION

1.1. Application of surface plasmons in biosensing

The potential of surface plasmon resonance as a highly sensitive probe of thin film characteristics has been recognised for many years. For example, the detection of organic monolayers deposited on metal surfaces by the Langmuir–Blodgett method was demonstrated by (Pockrand et al., 1977). This potential was later applied to biosensing, an early example being the detection of antibody-antigen interactions (Liedberg et al., 1983). Commercial instruments for such applications have now been available for several years, as described by Löfils et al. (1991), and interest in these techniques appears to be increasing (see, e.g., Leung et al., 1994; and references therein). The extension of surface plasmon methods to the imaging of biological layers has also been reported (Okamoto & Yamaguchi, 1990). In general, optical excitation of plasmons on a planar metallic surface allows the detection of adsorbed thin films of sub-nanometer thickness, while giving lateral resolution of tens of microns or better.

The purpose of this paper is to examine critically the characteristics and potential advantages of these over comparable techniques, and in particular to present some quantitative analysis of the resolution and sensitivity limits. To introduce this analysis, the relevant theory is reviewed below.

1.2. Optical measurement of thin films

The extension of human vision by artificial optical instruments has long been a principal method of investigation. Essentially, such instruments respond to the index of refraction and coefficient of absorption of materials, and the related properties of absorption and reflection by surfaces. Spatial distribution of these quantities can be determined to high resolution, and thus a great deal of information can be obtained. Further information is available by measuring the variation of optical properties with wavelength (dispersion), intensity
(nonlinearity) and polarisation (anisotropy). Processes which convert radiation from one form to another, such as fluorescence, can also be the basis of valuable measurement tools.

A particular challenge applies, however, when characterisation is wanted of a layer of material whose thickness is on the order of, or less than, the optical wavelength. Unless the material is particularly strongly absorbing, light will pass through such a layer with little effect on its intensity or phase, the two most important bases of measurement. To make optical examination of such layers possible, some method is needed to enhance the optical effect, and thus the signal. Another factor in choosing such a method is that whereas with bulk materials the measurement of reflected or transmitted intensity is favoured for ease of implementation, with thin films the potentially high sensitivity offered by measurement of refractive index might be necessary. The phase change of a coherent beam passing through a medium will be proportional to its index, and this phase change can be measured with great precision interferometrically. However, such measurements are generally difficult to implement and impose stringent restrictions on the optical source (particularly on coherence length and collimation). We shall see that a method which probes refractive index without requiring an interferometric instrument could be highly advantageous.

A conceptually straightforward way to improve the sensitivity of optical measurements of films is to increase the optical path length, and this can be done in two ways: by making the light pass through the film effectively more than once, and by making the light travel laterally through the film. These options are illustrated schematically in Figure 1. The former will always happen to some extent if there are index differences between the film and the surrounding media, since these will cause multiple reflections. Usually the reflectance is, however, weak, and therefore the enhancement in optical path length is modest. Arranging for large enhancement effectively means making the film part of a Fabry-Perot resonant cavity of high finesse, i.e. one with highly reflecting mirrors at both sides. This is not normally an inherent property of the sample, and will generally be difficult to realise.

The second method requires that the light be coupled into the film somehow, and then travel along it for some distance before leaking away. The latter requirement means that a guided mode must exist, even if a leaky one. Generally this means that the light must travel in a medium of higher refractive index surrounded by a media of lower refractive indices. This guiding medium could be the film of interest, but as the evanescent field spreads into, and is influenced by, the adjacent media, the measured film could also be one of these. Coupling into and out of the guiding film can be effected in three ways, as illustrated in Figure 2: by prism, grating, or end-fire coupling. One disadvantage of such an approach is that the measured signal depends, in effect, on an integration of the optical properties along its full path, and therefore spatial resolution is lost. This may or may not be of importance, as will be discussed later.

It is unlikely that the film to be measured will be of suitable index, thickness, and uniformity to form the guiding layer, and therefore it is more practical to consider it for use as an adjacent or ‘cladding’ layer, as mentioned above. The sensitivity of the guided light to the properties of a cladding layer depend very much on what fraction of the optical power propagates in this layer. This means that to be effective in this application, the structure must be designed such that this fraction is very high, and this can be achieved if the guiding layer is very thin and of very high index.

This is illustrated schematically in Figure 3, and is called a planar or slab dielectric waveguide. Such a structure is necessary if all media are non- or weakly absorbing. However, light can also be guided in a dielectric between two highly absorbing media, such as metallic conductors, and in special conditions a single metallic surface can also support propagation of a guided optical mode. This latter case is what is known as a surface plasma wave or surface plasmon (the name

![Fig. 1. Methods of extending optical path length in thin films: (a) multiple pass; (b) transverse path.](image-url)
'plasmon' came into use as originally these oscillations were generated by electron irradiation; their quantum nature is in this context of no significance).

2. SURFACE PLASMON RESONANCE

Surface plasmons are the electromagnetic waves associated with longitudinal oscillations of the free charges on the surface of a plasma. These can be propagated up to a frequency of $\omega_p/\sqrt{2}$ where $\omega_p$ is the bulk plasma frequency. Since $\omega_p^2$ is proportional to the charge density, the very dense plasma of free electrons in a metal allow plasma oscillations at frequencies up to those of ultraviolet light. At the same time, the frequency must be high enough so that individual charges oscillate with a period small compared to the mean time between collisions, otherwise the oscillations will be highly damped. In the absence of collisions, the acceleration of charges will be in phase with an externally applied oscillating field, and thus the displacement of the charges (the polarisation) will be 180° out of phase with the applied field. The implication is that the permittivity in such a situation is negative. We can also deduce that there will be electric field components in the $z$ and $x$ directions as illustrated in Figure 4, but not in the $y$ direction, and therefore that the electromagnetic field in the dielectric medium adjoining the metal has TM (transverse magnetic) polarisation.

These latter conditions can also be found by looking for a solution to Maxwell's equations at the boundary between two media. Analysis of surface plasmons by such an approach has been reported in detail by a number of authors; here, we briefly review the theory as an introduction to the analysis which follows. Generally the notation will be that of (Raether, 1977), to which the reader is directed for a more comprehensive treatment.

Let us consider again the geometry of Figure 4. An arbitrary field distribution can be written as a superposition of TM and TE (transverse electric) polarisations, these having the electric and magnetic fields in the $x$-$z$ plane, respectively. A surface wave solution to Maxwell's equations cannot be found for TE, so we can restrict ourselves to TM. The electric fields

![Fig. 3. Symmetric slab waveguide profiles: (a) refractive index (dashed line); (b) optical intensity (solid line).](image-url)
values must be imaginary (with opposite sign), their magnitudes then giving the inverse decay distance of the evanescent fields. We can then rewrite (1) as:

\[
\begin{align*}
E_x(x,z,t) &= E_{x0} \exp[i(k_{x0} x + k_{z0} z - \omega t)] \quad (z>0) \\
E_x(x,z,t) &= E_{x1} \exp[i(k_{x1} x + k_{z1} z - \omega t)] \quad (z<0)
\end{align*}
\] (4)

Maxwell’s equations give us the boundary conditions between the fields in the two media (discussed in more detail in Section 4), from which we obtain an expression for \( k_x \):

\[
\frac{k_x}{k_0} = \left( \frac{\varepsilon_{1} \varepsilon_{1}}{\varepsilon_{0} + \varepsilon_{1}} \right)^{1/2}
\] (5)

Let us take for the metal

\[
\varepsilon_{1} = \varepsilon_{r} + i\varepsilon_{i}
\] (6)

Equation (5) will give us both a real and imaginary part for \( k_x \). For the field to be evanescent in both media we also need \( \varepsilon_{i} < \varepsilon_{0} \).

The natural frequency of oscillation of surface charges (for \( \varepsilon_{r} = 1 \)) is \( \omega_{0} \sqrt{2} \), but propagation of surface plasma waves can be driven by an external source at arbitrary frequencies below this. The crucial factor in determining their period and decay distance is the complex permittivity of the metal at the frequency of interest. In general surface plasmons are generated at or near frequencies in the range of visible light, with the loss increasing rapidly as wavelengths go into the infra-red. However, plasmons cannot be generated simply by irradiating the metal surface with visible light, since it is impossible to phase match the incident beam with the plasmon along the surface.

This restriction was first overcome by using coupling prisms, in configurations devised by (Otto, 1968) and (Kretschmann, 1971), as indicated in Figure 5. Here the spatial frequency of the incident beam is increased by a factor \( n^2 \) (the prism index), and an evanescent field is generated by total internal reflection at the prism base. The parallel component of the incident beam wave vector is then given by:

\[
k_{i} = n_{2} k_{0} \sin \theta
\] (7)

This evanescent field will couple into the surface plasma.
mode when phase matching is achieved, i.e. when $k_x$ of the incident beam equals the real part of $k_x$ as given by (5). This is equivalent to prism coupling with a dielectric waveguide, but there is a difference in the effect of this coupling. In either case energy coupled into the guided mode travels in the $x$ direction, but gradually radiates ('leaks') back into the prism, since this coupling is inherently reciprocal. In the case of the dielectric guide, there is no intrinsic absorption in the mode, so that all the light coupled in also radiates out. To determine that coupling has taken place, one must measure either the displacement of the reflected beam (effectively a modified Goos–Hänchen shift) or the phase change. The first is imprecise, and the latter introduces experimental
difficulties as mentioned above. But with the surface plasmon there is intrinsic loss, and so when coupling is achieved, the reflected intensity drops. This is a rather straightforward measurement, and has the valuable feature of giving information on refractive index (since $k_f/k_0=n_0$) simply by a measurement of intensity.

The reflectance for the Kretschmann configuration, which is most commonly used, can be calculated precisely using modified Fresnel equations, and there are now a number of commercial software packages with which this problem can be investigated. However, it is still useful to have an analytic approximation for the form of the resonance curve, as this can be used to give insight to the physical phenomenon and to determine approximate sensitivity limits.

The derivation of this equation is not straightforward, and so will not be presented here. It has, however, a simple form with a small number of parameters, which we shall now introduce. First, we have the propagation constant of the plasmon propagating between semi-infinite media, as given by equation (5). Let us now call this the resonance wave vector $k_{m}$. This will have real and imaginary parts which we will define by:

$$\frac{1}{2} k_{m} = \frac{\varepsilon_0 \kappa_1}{\varepsilon_0 + \varepsilon_1}$$

(8)

Here $\Gamma_i$ is the intrinsic loss term. The resonance wave vector $k_{m}$ is perturbed by the introduction of the prism, the small amount of mode energy in the prism raising the effective index (or equivalently the real part of $k_f$) by an amount $\Delta k_{m}$, and leakage into the prism adding a re-radiative loss term $\Gamma_r$. Then we have:

$$k_{r} = \left(k_{m} + \Delta k_{m}\right) + i(\Gamma_i - \Gamma_r)$$

(9)

Here we have introduced the notation $k_{r}$ to indicate the complex wave vector of the resonant surface plasmon oscillation in the absence of an external excitation field. In practice, the spatial frequency is usually determined by such an exciting field, according to $k_f=n_0 k_0 \sin \theta$. The reflectance in the immediate vicinity of the plasmon resonance can then be written as:

$$R = 1 - \frac{4\Gamma_i \Gamma_r}{\left(k_{f} - k_{m}\right)^2 + (\Gamma_i - \Gamma_r)^2}$$

(10)

taking $k_{sr} = k_{sr} + \Delta k_{sr}$. To obtain an equation in terms of angle we introduce the resonance angle:

$$\sin \theta = \frac{k_{sr}}{n_2 k_0}$$

(11)

Then:

$$R = 1 - \frac{4\Gamma_i \Gamma_r}{n_2^2 k_0^2 (\sin \theta - \sin \theta)^2 + (\Gamma_i - \Gamma_r)^2}$$

(12)

3. SENSING WITH PLASMONS AND DIELECTRIC WAVEGUIDES

Let us now consider what we are really measuring. First of all, from the phenomenal point of view, the dielectric/metal interface is being probed with high sensitivity because the effective path length is increased as the plasmon propagates along the surface. Associated with this path length is an increase in the optical field intensity with respect to the intensity of the incident beam, effectively because the guided mode gains in strength as it propagates, as more and more energy leaks into it. These considerations indicate that the degree of enhancement of sensitivity depends very much on the propagation length, which in turn is determined by the rate of decay of the plasmon due to absorption and re-radiation.

With this relationship in mind, it is important to note that the greatest sensitivity will be obtained with the greatest possible propagation length, and that from this point of view plasmons are not the ideal candidate. Dielectric waveguides can be constructed with negligible intrinsic loss, and thus are inherently a much more sensitive probe than the surface plasmon. They can be used for sensing in a variety of geometries (e.g. Kunz et al., 1994). Furthermore, an intensity based measurement for dielectric guides can be obtained in several ways (see Fig. 6). A symmetrical coupling structure, using either a grating or prisms on both sides of the guide, will result in coupled light leaking out into both reflected and transmitted beams, so that the relative intensity of these two beams becomes an indication of resonant coupling. Alternatively, light can be coupled in and then out at two separate locations, using prisms or gratings, such that resonance is indicated by propagation between the two. And interferometric measurements, while less straightforward, have been made in both plasmon and dielectric waveguide imaging (Herminghaus et al., 1994).

There are, however, limitations to the sensitivity that can be obtained by such methods. These are best
explained by considering the measurement as a coherent (phase matched) coupling between a propagating mode and a probe beam, in order to measure the phase velocity (or k_x) of the mode. The precision of this measurement depends on the path length over which the probe and guided mode are mutually coherent; in the case of plasmons, the decay distance is usually the limiting factor. However, if this limitation is removed, others appear. For the probe beam to have a long coherence length, it must be very well collimated, i.e. have a narrow angular spectrum. This is all the more clear when we consider that the information is obtained by measurement of intensity vs angle, and that even for plasmons the width of the resonance curve is usually less than 1°. If the angular variation of the beam is not considerably less, the resonance will be 'smeared out' (effectively giving a convolution of the ideal curve with the beam angular spectrum), with a resultant loss of precision. Furthermore, the structure, including any measurand, must be uniform over the coherence length, to a degree increasing as this length increases.

We can summarize the situation as follows. Surface plasmons, being guided waves, offer a possibility of probing surfaces with high sensitivity. However, this sensitivity is inherently limited by the nature of the plasmon and its supporting material. Dielectric waveguides offer potentially much higher sensitivity, and do not suffer from intrinsic absorption losses. Enormous improvements in sensitivity may not be easy to obtain in practice, but for production of sensors based on a well developed fabrication technology, dielectric guides may be more appropriate. For many laboratories, however, the plasmon approach may offer sufficient sensitivity and therefore may be chosen for its ease of realisation. If imaging is also desired, the issue of propagation length takes on a new significance, as we shall see.

4. SENSITIVITY IN SURFACE PLASMON SENSING

Let us look again at the surface plasmon resonance curve, eqn (10). In most cases we wish in effect to determine k_x, and this is done by finding the position of lowest reflectance. The precision with which the resonance position can be measured depends on the width of the curve, which is determined by the total loss term Γ总计=Γ_i+Γ; the less loss, the sharper the resonance and thus the more precise the measurement. This was recognised by (van Gent et al., 1990) in a relative figure of merit. Another important consideration is the depth of the resonant 'dip', and it can be quickly seen that this is maximised in the case where Γ_0=Γ, for which resonance R=0. This is in fact an example of impedance matching.

We can think of the surface plasmon as a damped oscillator which is driven by a transmission line (the probe beam) of some characteristic impedance (the coupling strength). Only when the two impedances are equal will there be no reflected energy back towards the source. We might reasonably expect that the resonance 'frequency' would be best measured in the case where the maximum energy is launched into the resonant mode. However, in plasmon sensing we are often not concerned with an absolute measurement of the plasmon properties but only with a change of reflectance due to some temporal or spatial variation of the properties of a measurand.

This affords us the first opportunity of an analytical result relating to sensitivity. The most straightforward measurement mode is to choose a specific angle of incidence and observe the variation in reflected intensity as the properties of the resonance curve alter with the measurand. In this case the sensitivity is greatest where the slope dR/dθ is greatest. This can be determined analytically, but the analysis is greatly simplified if we consider instead dR/dk_x. Noting that k_x is proportional to sinθ rather than θ, resonance curves plotted as R(θ) do not have the symmetry implied by eqn (10) and become increasingly distorted as the curves get wider (greater loss).

The maximum dR/dk_x is found by setting dR²/dk_x²=0, giving:

\[
\left(\frac{dR}{dk_x}\right)_{max} = \frac{3\sqrt{3}}{2} \frac{\Gamma_i\Gamma_r}{(\Gamma_i+\Gamma_r)^3}
\]

for

\[
k_x-k_x'=\pm(\Gamma_i-\Gamma_r)\sqrt{3}
\]

If we now vary Γ_r, by varying the coupling strength to maximise the slope given by (13), we find that the highest sensitivity is achieved when:

\[
\Gamma_r=\Gamma/2
\]

That is, if the coupling strength is reduced from the matched case Γ_r=Γ, initially the maximum slope of the curve rises although the contrast (depth of resonance curve) is reduced.

This result has been reported previously (Yeatman & Ash, 1988b), along with other considerations on relative variations in SPR sensitivity. It would also be valuable, however, to have some analytical expressions for the absolute sensitivity of the technique, which we can
obtain in a similar way. The expressions given above can only be used to indicate the relationship between precision in different quantities. To form absolute expressions we must determine a precision in the directly measured quantity, the reflectance, and this we do by introducing a noise term, \( N \). Taking \( R \) to be a dimensionless measure of fractional reflectance, \( N \) is also a dimensionless quantity, and is defined as:

\[
N = \frac{\text{noise equivalent (optical) power at receiver}}{\text{optical power of incident probe beam}}
\]

In a typical experiment with 1 mW of incident light, \( N \) might be on the order of \( 10^{-3} \) or 30 dB, but this is highly dependent on experimental conditions.

By combining eqns (13) and (15) we obtain an expression for the maximum slope when the coupling strength is optimised:

\[
\frac{dR}{dk_x} = \frac{2\sqrt{3}}{9\Gamma_i} \tag{16}
\]

To obtain sensitivity expressions from this, it is useful to express \( \kappa_i \) and \( \Gamma_i \) in terms of the physical parameters \( n_2 \), \( \varepsilon_i \), and \( \varepsilon_r \). We start with eqn (5), with the notation of (9), replacing \( \varepsilon_i \) using (6) and making the approximations \( -\varepsilon_r > > \varepsilon_i \) and \( -\varepsilon_r > > \varepsilon_0 \). Then, after some algebraic manipulation, we can obtain:

\[
\frac{k_{x,m}}{k_0} = n_0 k_0 \varepsilon_i \left( 1 + \frac{\varepsilon_0}{2\varepsilon_r} \right) \left( 1 + \frac{\varepsilon_r}{2\varepsilon_r} \right) \left( 1 + \frac{\varepsilon_r}{2\varepsilon_r} \right) \left( 1 + \frac{\varepsilon_r}{2\varepsilon_r} \right) \tag{17}
\]

This gives directly:

\[
\Gamma_i = \frac{n_0^3 k_0 \varepsilon_i}{2\varepsilon_r} \tag{18}
\]

We can obtain a sensitivity limit by combining (18) and (16) and recognising that the minimum detectable \( dR \) is just \( N \); the change in \( R \) is the same for a small change in the measurement wave vector \( dk_x \), as for an equal change in the resonance wave vector \( dk_{x,m} \), so that:

\[
\lim_{\Delta k_{x,m}} \left[ \frac{dR}{dk_x} \right] = \frac{9N}{4\sqrt{3}} \frac{n_0^3 k_0 \varepsilon_i}{\varepsilon_r^3} \tag{19}
\]

Let us take as an example the case of a silver/air interface interrogated at 633 nm wavelength, a very common experiment. Then \( n_0 = 1 \), \( \varepsilon_i = 0.5 \), \( \varepsilon_r = -15 \), and if we use \( N = 10^{-3} \) we obtain a sensitivity to the air index of \( d\varepsilon_0 = 3 \times 10^{-6} \). This is indeed a very precise measurement.

Normally, plasmon resonance will not be used to measure the index of a bulk medium. Of more practical interest is the detection of a thin dielectric film, and therefore an estimate of sensitivity for this case could be of greater value. The development of the approximation is rather complex, and is presented here in an abridged form.

Let us find the solutions for a guided TM mode in a three layer system, as illustrated in Figure 7 but without the coupling medium \( \varepsilon_2 \), with \( \varepsilon_0 \) and \( \varepsilon_3 \) giving need to choose an experimental structure, and the simplest is the detection of the variation of index of a uniform dielectric medium, \( n_0 \), for example in a gas sensor. From (17) we can say that, to a first order, \( \frac{d \varepsilon_{m}}{n_0} = k_0 \), so that the minimum detectable change in the index of the dielectric is given by:

\[
\lim_{\Delta \varepsilon_1} \left[ \frac{d \varepsilon_1}{\varepsilon_0} \right] = \frac{1.3Nn_0^2 \varepsilon_{m}}{\varepsilon_0^3} \tag{20}
\]

Let us take as an example the case of a silver/air interface interrogated at 633 nm wavelength, a very common experiment. Then \( n_0 = 1 \), \( \varepsilon_i = 0.5 \), \( \varepsilon_r = -15 \), and if we use \( N = 10^{-3} \) we obtain a sensitivity to the air index of \( d\varepsilon_0 = 3 \times 10^{-6} \). This is indeed a very precise measurement.

Normally, plasmon resonance will not be used to measure the index of a bulk medium. Of more practical interest is the detection of a thin dielectric film, and therefore an estimate of sensitivity for this case could be of greater value. The development of the approximation is rather complex, and is presented here in an abridged form.

Let us find the solutions for a guided TM mode in a three layer system, as illustrated in Figure 7 but without the coupling medium \( \varepsilon_2 \), with \( \varepsilon_0 \) and \( \varepsilon_3 \) giving
respectively the relative permittivities of the bulk dielectric and a film of thickness \( a \). The magnetic fields will have the form:

\[
\begin{align*}
H_y &= B \exp(i k_z z) + C \exp(-i k_z z) \quad 0 < z < a \\
D \exp(i k_x z) &= z 
\end{align*}
\]  (21)

Here, the common factors \( \exp(i k_z z) \) and \( \exp(-i k_z z) \) have been omitted. The \( z \) components of the propagation constants may be real or imaginary in general, although for the guided mode they will certainly be imaginary in the bulk media. The exact solution to the form of the guided mode is found by eliminating the constants \( A-D \) using the boundary conditions. These are that \( H_y \) and \( E_x \) are both continuous at the interfaces \( z=0 \) and \( z=a \). Noting that \( E_x \) is proportional to \( \delta H_y / \delta z \) we obtain a set of simultaneous linear equations in \( A, B, C \) and \( D \). These can be greatly simplified if we assume that the middle layer is very thin, such that:

\[
\exp(i k_z a) = 1 + i k_z a 
\]  (22)

Using only this approximation, and algebraic manipulation, we can obtain:

\[
\frac{k_x n_0}{k_0 n_1} - 1 = i k_z a \begin{pmatrix} k_x n_0 & k_x n_1 \\ k_x n_1 & k_x n_0 \end{pmatrix} = 0 
\]  (23)

Here, the right-hand-side of the equation, which we have labelled \( p \), is a perturbation on the simpler two medium case; by setting \( p=0 \) we can obtain the standard solution, equivalent to eqn (5). Note that since \( k_x^2 + k_z^2 = k_0^2 \) in each medium, (23) can be converted to a transcendental equation for \( k_z \).

This analysis does not include the effect of the prism on the position of the resonance. However, it is a reasonable approximation that this effect and the effect of the perturbation \( p \) on the resonant value of \( k_x \) are independent. Then by combining the minimum detectable resonance shift (19) with the dependence of the resonance position on the additional layer, we can determine the detection limits for this layer. The minimum detectable thickness ‘\( a \)’ will be given by:

\[
\lim(a) = \frac{1}{(d k_x/\delta p)(d p/\delta a)} 
\]  (24)

We can find \( d k_x / \delta p \) by converting (23) to an equation for \( k_x^m \), from which we obtain:

\[
\frac{d k_x^m}{\delta p} = \frac{(k_x^m/k_0)^3}{\varepsilon_r - \varepsilon_0} \frac{n_0 k_0}{k_x^m} 
\]  (25)

Here we have used only the first term of (17) for \( k_x^m/k_0 \), and the approximation \( |\varepsilon_r| > > |\varepsilon_0| \). Finding an approximation for \( \delta p/\delta a \) using (23) requires some careful manipulation and yields:

\[
\delta p/\delta a = -2 k_0 n_0 |\varepsilon_r| \Delta n/\varepsilon_0 
\]  (26)

Here \( \Delta n \) is the index difference between the bulk dielectric and the thin film, i.e. \( \Delta n = n_3 - n_0 \) (where \( n^2 = \varepsilon \)).

By combining (19) and (24-26) we arrive at an expression for the minimum detectable thickness of a dielectric film with an index difference \( \Delta n \) from the bulk dielectric medium:

\[
\lim(a) = 0.1 \frac{\varepsilon_r}{|\varepsilon_r|^{1/2}} \frac{n_0}{\Delta n} \lambda N 
\]  (27)

Let us take as an example a gold film having a relative permittivity \( \varepsilon_r = -11 + i 1.3 \) at the measurement wavelength 633 nm. With the Kretschmann geometry in the maximum slope case, we wish to detect a film having an index difference 0.1 from the surrounding medium \( n_0 = 1.33 \), with a detectable change in reflectance \( N=0.01 \). Then (27) gives a detectable thickness of 0.3 nm. This prediction is borne out by numerical simulation. Figure

![Figure 8. Surface plasmon resonance (at 633 nm wavelength) for a gold surface contacting a dielectric medium of refractive index 1.33 (solid line), and for an additional 3 nm layer between these two media of refractive index 1.43 (dashed line).](image)
8 shows the resonance curve for this case with and without an added dielectric layer 3 nm thick, which gives a maximum reflectance difference of about 10%. This was modelled using a commercial optical design package (Film Wizard).

Figure 8 also illustrates that the major effect of adding a non-lossy dielectric film to the sensing surface is to shift its position; the shape is not by comparison sufficiently altered. Indeed the analysis above neglects any change to the imaginary part of \( k_x \), and simulations show that this assumption holds in all but extreme cases. Therefore eqn (27) provides a simple but reasonably accurate expression for the detection limits of thin films in a wide range of practical cases.

5. SURFACE PLASMON MICROSCOPY

The information available from the interrogation of a surface by plasmon resonance can be greatly increased if the variation of response with position is used. This can quite easily be done; the first report of surface plasmon microscopy simply used an expanded collimated beam in the Kretschmann configuration with an imaging lens placed in the reflected path (Yeatman & Ash, 1987). Similar results were later reported by another group (Rothenhausler & Knoll, 1988b), and various instruments have been constructed (Okamato & Yamaguchi, 1990, 1992). Measurements using a scanned focused beam have also been investigated (Yeatman & Ash, 1988b).

As discussed in Section 4, the sensitivity of the surface plasmon measurement is a function of the length of coherent interaction between the probing field and the plasmon. This will be limited by the smaller of the plasmon decay length, the probing field coherence length, and the length along which the physical system is uniform in the propagation direction. Thus in cases where the plasmon decay length is otherwise the limiting factor, full sensitivity will be achieved if the variation of the measurand is less than this sensitivity limit over the propagation length. If the variation is on a smaller lateral scale, sensitivity will decline because the measurement effectively averages over the decay length.

The plasmon decay length for the simpler geometries is fairly short, typically a few tens of \( \mu \text{m} \) at most. As a resolution limit this is quite acceptable for many applications. Guided dielectric modes will usually have much longer propagation lengths, and so while offering potentially higher sensitivity their spatial resolution is poor. The same considerations apply to so-called long range surface plasmons, which are generally hybrids of plasmon and dielectric modes, or are produced by the interactions of plasmons on two metal surfaces. The opposite trade-off can also be made: resolution can be enhanced by the use of shorter-range plasmons, as extensively discussed in (Berger et al., 1994). This can be done by changing the probe wavelength, or by choosing a different metal; in any case it will always be at the expense of sensitivity.

It should also be noted that those resolution limits apply to only one of the directions along the surface. The orthogonal direction to that in which the plasmon mode propagates, the \( y \) direction in the notation used here, is not limited in the same way. Thus extra information can be obtained by rotation of the object (De Bruijn et al., 1993).

A simple analytic approach to characterising the relationship between resolution and sensitivity is to consider excitation of a plasmon by a focused beam. Where the detector measures the total reflected intensity as a function of the axial incident angle of the beam, the response will be a convolution of the ideal resonance curve with the angular spectrum of the incident beam. For this analysis it is useful to have the resonance curve in angular units. If the resonant dip is narrow, we can make the approximation that \( \Delta \sin \theta = \cos(\theta) \Delta \theta \); we can then convert (12) to:

\[
R(\theta) = \frac{1}{1 + (\Delta \theta / \omega)^2} h
\]

(28)

where \( \Delta \theta = \theta - \theta_r \), \( \theta_r \) being the resonant angle. Here we have parameterised the dip by its height, given by:

\[
h = 4 \Gamma_p \Gamma_r / (\Gamma_p + \Gamma_r)^3
\]

(29)

and by its width:

\[
\frac{\Gamma_p + \Gamma_r}{w} = \frac{n_r k_0 \cos \theta_r}{n_r k_0 \cos \theta_r}
\]

(30)

We can repeat the analysis of Section 4 in this formulation to find equivalently that the maximum slope is obtained when \( \Delta \theta = \omega / \sqrt{3} \). At this position, the value of the slope is proportional to \( h/w \), and therefore the sensitivity is also proportional to this parameter.

Let us now consider a focused incident beam of Gaussian angular spectrum:

\[
A(\theta) = A_c \exp \left( -\left( \theta - \theta_c \right)^2 / b^2 \right)
\]

(31)

where \( \theta_c \) is the axial angle and \( b \) is the characteristic half
width. We can show (Yealman, 1989) that the measured resonant width $w'$ in this case will be approximately given by:

$$w'_Z \approx \sqrt{w^2 + b^2}$$

(32)

In this case the sensitivity will be proportional to $h/w'$. Since the amplitude distribution incident at the probed surface is just the Fourier transform of $A(\Theta)$, we can relate $b$ to a spot half width $x_0$, defined at the half-power point. This requires some numerical calculation and yields:

$$x_0 = \frac{0.675}{n_2 k_0 \cos \Theta} b$$

(33)

Equivalently we can relate the resonance width $w$ to a characteristic length parameter $x_\epsilon$, where:

$$x_\epsilon = \frac{0.675}{n_2 k_0 \cos \Theta} w$$

(34)

If we now define a sensitivity parameter $S = h/w'$, we can obtain a sensitivity versus spot size relationship:

$$S = \frac{h}{w} x_\epsilon$$

(35)

Using (34) and (35) we find that $x_\epsilon = 0.675/(\Gamma_\epsilon + \Gamma_\epsilon)$; taking the case $\Gamma_\epsilon = \Gamma_\epsilon/2$ and expanding $\Gamma_\epsilon$ using (18) we obtain:

$$x_\epsilon = \frac{0.14 \epsilon_2}{n_0 \epsilon_1^3}$$

(36)

As an example, for the case of silver/air at 633 $\mu$m we get a characteristic length $x_\epsilon = 40 \mu$m. In fact $x_\epsilon$ only differs from the natural decay length of the plasmon, $1/(\Gamma_\epsilon + \Gamma_\epsilon)$, by the factor 0.675 which is introduced by our way of defining the spot size. Thus for the same example the decay length is 59 $\mu$m, while 40 $\mu$m is the incident spot half-width at which the sensitivity is reduced by $1/\sqrt{2}$.

We see that the same parameters that determine sensitivity also indicate the attainable resolution. In fact, the product of lateral resolution, as given by (36), and index sensitivity, as given by (20), is independent of all material parameters, within the approximations used, indicating a straightforward trade-off between the two. If we consider instead sensitivity to an added layer, we see that the detectable thickness of such a layer, as given by (27), varies more slowly with $\epsilon_2$ and $n_0$ than the lateral resolution.

6. IMAGING OF FEATURES

While we have discussed in Section 5 the relationship between sensitivity and lateral resolution in surface plasmon measurements, it is also of interest to consider how the technique responds to lateral variations in the measurand. In effect, we seek a diffraction theory for surface plasmon microscopy.

A way is found to consider this problem analytically by recognising that the reflected radiation which is measured can be broken down into two distinct components—the re-radiated field which leaks from the lossy plasmon mode, and the directly reflected field which is that which we obtain if no plasmon is generated. If the amplitude coefficients of the incident and reflected fields are $E$ and $D$, respectively, then the reflected amplitude from a particular lateral position $x$ is given by:

$$D(x) = E(x) r^2_{11} + D^+(x)$$

(37)

where $D^+$ is the re-radiated component and $r_{21}$ is the amplitude reflection coefficient of the prism/metal boundary. The first term is independent of the properties of the plasmon supporting surface, so the only varying quantity of interest is the plasmon field amplitude, to which

$$D^+$

is proportional. This is determined by the cumulative effects of absorption, re-radiation, and coupling from the incident field. We can write:

$$dD^+/dx = -[(\Gamma_\epsilon + i(k_x - k_{\epsilon}))D^+(x) - 2\Gamma_\epsilon E(x)r_{21}$$

(38)

Here the first term is the total absorption loss ($\Gamma_\epsilon = \Gamma_\epsilon + \Gamma_\epsilon$), the second is the effect of phase mis-match between the incident beam and the natural plasmon oscillation, and the final term is that of the field coupled into the plasmon, noting that $\Gamma_\epsilon$ determines equivalently the rate of leakage in and out. The derivations of (37) and (38) can be found in (Yealman, 1989) and (Yealman & Ash, 1988a). The result is similar to that obtained for tapered waveguide couplers by (Ulrich, 1971).

We can use (38) to find the output field pattern for a focused beam on a uniform surface, by integrating in the
+x direction from the beginning of the beam, setting $D(x)$ initially to zero. More interestingly we can use it to determine the response to arbitrary (one-dimensional) features. Here we will consider only two important general cases: abrupt transitions (step features) and periodic variations. First we should consider the limits of validity of this analysis. We have not considered coupling directly into radiation modes on the measurand side of the metal, but such coupling is possible for features having spatial frequencies greater than $k_{x''} - k_0$. Also, reflections of plasmon modes into the -x direction are not explicitly included, although this can be added, as we will see below.

Let us then consider the effect of a step feature where there is a sudden change in $k_{x''}$, as at the border between a bare metal and one with a dielectric coating. If the boundary is at $x=0$ and the incident field is uniform, then $D^+$ will take some constant equilibrium value. We can determine this value, which we label as $D_0^+$, by setting $dD^+(x)/dx$ to zero in (38), which gives:

$$D_0^+ = -\frac{2\Gamma_1 E(x)x_1}{\Gamma_1 + i(k_x - k_{x''})}$$

(39)

We can now rewrite (38) as:

$$dD^+(x)/dx = -[\Gamma_1 + i(k_x - k_{x''})](D^+(x) - D_0^+)$$

(40)

The solution to this equation is given by:

$$D^+(x) = D_0^+ + C \exp[-(\Gamma_1 + i(k_x - k_{x''}))x]$$

(41)

where $C$ is a complex constant, the value of which will be determined by boundary conditions.

Where all conditions have been constant over a significant distance in the +x direction, $D^+$ will equilibrate, and $C$ will be zero. But for $x$ just greater than zero, immediately after the step transition in the surface's characteristics, the plasmon field does not immediately equilibrate. The only instantaneous change in $D^+(x)$ will be due to scattering into radiation modes in the (near-side) dielectric, and reverse-travelling plasmons. If we label the surfaces $x<0$ and $x>0$ as $a$ and $b$ respectively then the relationship between $D^+_a(x=0)$ and $D^+_b(x=0)$ is effectively a transmission coefficient, which we will write as $t_{ap}$. From this quantity we easily find $C$, and thus the full re-radiated field in medium 2:

$$D^+(x>0) = D_0^+ + (t_{ap} D_0^+ - D_0^b)x \exp[-(\Gamma_1 + i(k_x - k_{x''}))x]$$

(42)

All that remains is to calculate $t_{ap}$. This we can do by making the approximation that the transmission coefficient is the same in the prism coupled case as it would be for a plasmon travelling on an interface between metal and dielectric half-spaces. The determination of $t_{ap}$ in this case remains a difficult problem (Leskova, 1984), but a useful approximation has been derived by (Barlow & Brown, 1962) for radio frequency Zenneck waves which is also valid here:

$$t_{ap} = \frac{k_m^2 k_{m''}}{k_a^2 k_{a''}} \frac{2k_{ab}}{k_a + k_{ab}}$$

(43)

where $k_a$ and $k_{ab}$ are the values of $k_x$ on either side of the step, and similarly for $k_{x''}$. An interesting feature of this solution is that, in the regime beyond the step, the combination of the equilibrium response and the transient response (the $C$ term) can give a strongly oscillatory signal, i.e. fringes appear. This has been reported experimentally (Rothenhausler & Knoll, 1988a). In Figure 9, the reflected intensity distribution is plotted for a particular set of parameters. This is obtained by calculating $D^+(x)$ with $\Gamma_1 = 0.005k_0$, $k_{m''} = 1.05k_0$, and $k_{m''} = 1.03k_0$.}

Fig. 9. Reflected intensity from a uniformly illuminated plasmon coupling configuration having two resonant angles, corresponding to $k_{x''} = 1.05k_0$ for $x<0$ and $k_{x''} = 1.03k_0$ for $x>0$. For all $x$, $\Gamma_1 = 0.005k_0$. Lines labeled 11, 12, 13 correspond to values of $k_x$ for the incident beam of 1.05$k_0$, 1.04$k_0$, and 1.03$k_0$. 

-0.5 0 0.5 1 1.5 2

-0.5 0 0.5 1 1.5 2

0 500 1000

$k_{x''}$ (k_0)

$I(x)$ (a.u.)
according to (42), adding the term \( E(x)r_2 \), and taking the modulus squared. Note that \( r_2 \), is a phase term given by:

\[
r_{21} = \frac{k_x e_1 - k_y e_2}{k_x e_1 + k_y e_2}
\]  

(44)

The theory presented above can be used to analyze the case of periodic features, the simplest being sinusoidal variation of the plasmon propagation vector \( k_{\nu r} \). The response to a particular perturbation can be modelled directly from (38), and we shall see that for strong periodic perturbations, as with step features, artefacts appear. For weak features, however, the response is linear and a sensitivity vs spatial frequency relationship can be obtained. Let us consider the case of a small sinusoidal perturbation of \( k_{\nu r} \), of amplitude \( M \) and spatial frequency \( p \), about the incident centre frequency \( k_{xc} \). Then:

\[
k_{\nu r} = k_{xc} + M \cos(px)
\]  

(45)

We will define the equilibrium value of \( D^* \) as that in the absence of the perturbation, so that:

\[
D_0^* = -2\Gamma E(x)r_2/\Gamma_l
\]  

(46)

We will also introduce a variable \( \Delta = D^*(x) - D_0^* \). Then the differential equation for \( D^* \) (40) can be written as:

\[
d\Delta(x)/dx = -\Gamma \Delta(x) - iM \cos px D^*(x)
\]  

(47)

If the perturbation is small, then we can make the approximation that \( D^*(x) = D_0^* \) in the last term of (47) and then taking the derivative we obtain:

\[
d^2 \Delta(x)/dx^2 = \Gamma \Delta(x) + iM \Gamma D_0^* \cos px
\]  

\[-iMpD_0^* \sin px
\]  

(48)

We can find a solution to this equation of the form:

\[
\Delta(x) = P \cos(px + \phi)
\]  

(49)

where the amplitude and phase of the perturbation on the

reflected signal are given by:

\[
P = \frac{-iMD_0^*}{(\Gamma_l^2 + p^2)^{1/2}}
\]  

(50)

\[
\tan \phi = -\rho/\Gamma_l
\]  

(51)

We can conclude that for weak periodic perturbations the reflected intensity has a matching periodic perturbation whose modulation transfer function (amplitude of response vs spatial frequency) has the form \( (\Gamma_l^2 + p^2)^{-1/2} \). That is to say, the spatial period at which the response has dropped by 1/2 is \( 2\pi/\Gamma_l \). This can be compared to the results of Section 5, where the same reduction of sensitivity is found for a spot diameter \( 2x_c = 1.35/\Gamma_l \).

Finally, in Figure 10 we see the spatially nonlinear effects obtained as the perturbation amplitude increases, calculated using (38). The condition for linear response can be shown to be a simple relation between the range of variation of the local value of \( k_{\nu r} \) and the resonance width \( \Gamma_r \), i.e. linearity is obtained (for \( \rho = k_{xc} \)) where the condition \( M \Gamma_r \) is satisfied.

7. DISCUSSION AND CONCLUSIONS

Principally, this paper has been concerned with the interrogation of planar surfaces by surface plasmons, and in analysing the attainable sensitivity and lateral resolution for such measurements and the relationship between them. Comparison has also been made with the more general technique of optical guided wave sensing. One related method that has not been discussed is that of imaging a surface using a scanned metallic tip. Here plasma oscillations particular to the tip geometry can be optically excited, and the behaviour of these will depend on the optical characteristics of the immediate vicinity. In this case sub-wavelength resolution can be obtained (Fischer & Pohl, 1989).

Another aspect we have not considered is that of techniques for optimising the noise figure \( N \). Large improvements can be had by modulating the incident beam and using phase sensitive detection of the electronic signal; a particularly sensitive method has been reported (Bass et al., 1994) in which the wavelength is modulated acousto-optically and the resonance slope locked onto by a feedback control loop. It should also be noted that our considerations of sensitivity relate essentially to relative changes in the signal and thus the measurand. For identifying image
Fig. 10. Reflected intensity from a uniformly illuminated plasmon coupling configuration having a resonant angle distribution \( k_x = k_{x0} + M \cos(\Gamma x/4) \), with \( k_{x0} \) being the \( k_x \) value of the illuminating field, plotted for \( M/\Gamma x \) = (from top to bottom) 8, 4, 2 and 1. Lines are vertically offset for clarity.

features this is quite valid, but where precision is needed in the absolute measurement, such as for some types of sensor, we must add such factors as the precision with which the angle of incidence and other factors are known and stable.

An important final consideration is the effective multiplicity of possible mechanisms for variation of the plasmon signal, and how this affects interpretation of measurements. We have considered the sensitivity to movement of the resonance curve caused by a change in index of the bulk dielectric and by the addition of a thin additional layer. Where we are effectively only measuring one parameter, we cannot distinguish between these two mechanisms, or others such as addition of multiple layers, addition of a lossy layer, etc. Even if we measure the complete resonance curve, we have seen already that the two mechanisms analysed do not significantly alter the resonance shape, and so not much advantage is gained by measuring it.

Great improvement in the possibility of extracting parameters is obtained if the plasmon curve is made to be intrinsically multi-parametered, in particular by introducing a number of modes. This can be done by mixing plasmon effects with dielectric slab waveguide configurations and has been used to great advantage (Welford, 1991) to characterise the spatial variation of liquid crystal orientation in the direction perpendicular to adjacent metal surfaces. To combine such techniques with lateral imaging will require development of algorithms and software for image reconstruction, and work in this area is beginning to be reported (Morgan & Taylor, 1994).

ACKNOWLEDGEMENTS

The helpful comments of Professor Samir El-Bader on the manuscript are gratefully acknowledged. I am also indebted to Janusz Sadowski and Jukka Lekkala of VTT for inviting me to the ABI meeting, and hence in stimulating my re-analysis of the matters reported here.

REFERENCES


