



In the Introduction (Chap. 1) an account was given of the rationale for using the morphology of leaves as the source of raw material for automatic systems for the identification of plant species. In this chapter we focus on how morphological features have been used in statistical analysis to define taxa like species, test pre-existing species classifications and carry out identification. Our aim here is to survey briefly the progress and development of the statistical analysis of morphological variation. This general area of scientific research is called “morphometrics”, a term which is a compound of the classical Greek words *morphé*: “form”, and *metron*: “that by which anything is measured” (http://www.lexilogos.com/english/greek_ancient_dictionary.htm). “Morphometrics” can thus be construed as “the measurement of form”. “Form”, in turn, can be understood in a narrow sense as referring only to shape, or more broadly as including shape, structure (which includes size, architecture and internal anatomy) and other aspects of external appearance. It is in the broad sense that we will use it here.

There are a variety of “schools” of morphometrics which are “characterized by what aspects of biological ‘form’ they are concerned with, what they choose to measure, and what kinds of biostatistical questions they ask of the measurements once they are made.” (Slice 2005). The term “morphometrics” seems to have entered the biological vocabulary in an eponymous article by Blackith (1965), but has been used widely only from the 1970s onwards, following the publication of Blackith and Reyment’s book “Multivariate Morphometrics” (Blackith and Reyment 1971). Under this title these authors emphasized the application of multivariate statistics to data consisting mainly of linear measurements and hence constituting relative and absolute measures of size, from which shapes of morphological structures could be estimated indirectly.

In the 1980’s and 1990’s the field of morphometrics underwent a major upheaval with the development of geometric morphometrics (Rohlf and Marcus 1993; Adams et al. 2004), which focusses on the analysis of shape variation in organisms and

their component parts. The impact of geometric morphometrics has been such that the term “morphometrics” is today widely understood to mean the study of shape variation alone (Zelditch et al. 2012).

The increasing use of automated methods of data capture from the morphology of organisms, either directly or from images, brings into focus other aspects of form which do not seem satisfactorily understood as either shape or size features, e.g. textures (Lexer et al. 2009), patterns visible on the surface such as venation (Plotze et al. 2005) and others mentioned in the following chapters. The analysis of such features and the use of their patterns for identification can also be treated under a concept of morphometrics that is broad enough to include the algorithmic analysis of any aspect of the morphology of organisms (Gaston and O'Neill 2004; O'Neill 2007).

2.1 Historical Background to Morphometrics

In this section the development of morphometrics is briefly outlined, from the biometricians of the nineteenth century to the use of multivariate statistics for the quantitative analysis of morphological features of organisms, an area of study usually known as “Multivariate Morphometrics” (Marhold 2011) or “Traditional Morphometrics” (Marcus 1990), and finally to the rise of geometric morphometrics.

2.1.1 Phase 1: Classical Taxonomy, Evolutionary Theory and Biometry: The Background for Morphometrics

The recognition of species of plants and animals by their appearance (morphology) and behaviour goes back to humanity's own origins and certainly even before that (Atran 1990). Many religions account for the diversity we perceive by means of creation myths and implicit in these is the notion of stasis – that once created, species remain the same, since offspring almost always resemble their parents closely. Contrary to what has sometimes been argued (e.g. Hull 1965a, b), the dependence of the morphological similarity of the individuals of a species on heredity has always been recognized by taxonomists since at least Aristotle (Richards 2010; Wilkins 2009, 2010) and from time immemorial by ordinary people. Nothing is more obvious than that like breeds like.

Prior to the widespread acceptance of evolutionary theory following Darwin's *Origin of Species* (Darwin 1859), taxonomists attempted to delimit distinct units of biodiversity (as species) using qualitative morphological characters, inferring at the same time that the individuals composing these units would represent parent-offspring lineages (Hennig 1966, tokogenesis; Wilkins 2009), thus accounting for their phenotypic similarity. Individual variation, according to this generally accepted view (de Candolle 1813), was just an observed but trivial fact, again intuitively obvious to us all from our own family experience. “Good” species could only be accepted

as such if delimited qualitatively; for those organisms that proved more intractable taxonomically, there was an assumption that the necessary characters for achieving satisfactory groups must exist, but are simply hidden from current scientific view. This search still motivates most taxonomists and with good reason, since such species delimitations better serve the practical purpose of dissecting the bewildering diversity of nature into units which can serve as raw material for biological science. Here lies a fundamental tension of taxonomy with the evolutionary worldview: while evolution can be shown to produce groups of phenotypically distinct individuals (species), there is no logical justification for expecting qualitative distinctions between them to necessarily exist, given the theory of evolution – qualitative taxonomic distinctions are sufficient, but not necessary, for the evolutionary delimitation of species. On the other hand, all sciences need units and in biology the fundamental units above the level of individuals are species. Mankind hates the continuum and it has been and remains the taxonomists' lot to discretize biological diversity. In order to do this, taxonomists have always used morphological information and in the process have had to find ways of dealing with the observed fact that morphology varies from one individual to another.

Darwin's theory of natural selection changed the explanatory context of biodiversity in fundamental ways. In particular, he focussed on the significance of phenotypic variation in species and argued that it was variation itself which provided the raw material for the evolution of contemporary species from ancestral ones. An immediate consequence of this was that gradual phenotypic change must occur as a species is transformed by selection into another and that therefore there could be no fundamental distinction between taxonomic varieties and taxonomic species. Continuous variation in morphological characters now achieved new prominence since this would be an expected feature of the evolution of species.

The measurement and the quantitative relations of patterns of phenotypic (mostly morphological) variation and heredity became objects of interest and the detailed study of quantitative variation within and between species and infraspecific groups gave rise to the science of biometrics (statistical analysis of biological data), among whose pioneers were Francis Galton (e.g. Galton 1889), W.F. Raphael Weldon (e.g. Weldon 1889) and Karl Pearson (e.g. Pearson 1901; Pearson et al. 1901). Galton invented the correlation coefficient to help him investigate the principles of inheritance (mainly in people) in the pre-genetics era. These scientists discovered in biological organisms the common occurrence of such patterns as the Normal (or Gaussian) frequency distribution in many variables, both continuous and discontinuous. In order to analyse the data they collected, they made fundamental advances in mathematical statistics, developing a range of new mathematical methods, such as regression, which soon had wide application both within and beyond biology (Briggs and Walters 1997; Porter 1986).

2.1.2 Phase 2: Genetics and Statistical Methods in Evolution, Agronomy and Biosystematics

With the birth of genetics in 1900 following the rediscovery of Gregor Mendel's work, a more complete and satisfactory theory of heredity became available. Mendel (Mendel 1866) discovered his laws of inheritance using a statistical treatment of seven categorical (qualitative) morphological characters of the pea plant (*Pisum sativum*). But major difficulties remained in applying the particulate Mendelian theory to explain the continuous character patterns studied by the biometricians. Important insights were provided by the quantitative work of e.g. Johannsen (1909), Nilsson-Ehle (1909) and East (1915), which clearly distinguished genotype patterns from those of the expressed phenotype (Mather and Jinks 1971). It was Ronald A. Fisher's seminal paper of 1918 (Fisher 1918) that showed that continuous variation of characters was compatible with Mendelian particulate inheritance and he introduced the key concepts of variance and its analysis by partition — Analysis of Variance or ANOVA.

The development of biological statistics underwent rapid and continuous growth in succeeding decades, driven in particular by the application of genetic theory to agronomy and evolutionary theory, resulting in the sciences of quantitative genetics (e.g. Falconer 1989) and population genetics (e.g. Fisher 1999; Hamilton 2009), using data which in the case of plant studies was derived from phenotypic and mostly morphological features of plants. The three mathematical biologists who contributed most significantly to this development, especially between the 1920s and 1960s, were John B.S. Haldane, Ronald A. Fisher and Sewall G. Wright, whose work greatly expanded the scope and power of statistics as a toolkit for analysis of biological variation, both genetic and phenotypic. Reyment (1996) gives an illuminating account of some of the key biometric personalities in the period from Galton to Fisher.

The first half of the twentieth century was an interesting time in the prehistory of morphometrics. It was when genetics was still young and taxonomy still fairly dominant in botanists' thinking. Geneticists' discoveries of the depth and complexity of genetically influenced phenotypic variation within taxonomists' species had a major impact — an enormous amount of quantitative morphological work was done in this period. Turrill's discussion of the ecotype concept (Turrill 1946) gives a flavour of how the influence of genetics on taxonomy was seen at the time, from a botanist steeped in taxonomy but highly sympathetic to the "new systematics" that arose in response to genetic and evolutionary studies (Huxley 1971). This was when the science of biosystematics was born — essentially the quantitative study of infraspecific diversity, leading to the formulation of theories of microevolutionary processes — dubbed the Modern Synthesis (Briggs and Walters 1997).

Stebbins (Stebbins 1950, pages 13–21) gives an excellent survey of the use of quantitative methods in descriptive systematics in the period 1920–1950 (though mostly restricted to Anglophone authors). These studies include Woodson Jr. (1947) on leaf morphometrics of *Asclepias tuberosa*, Gregor et al. (1936) on experimental gardens for biosystematics, Lewis (1947) on leaf variation in *Delphinium variegatum*, Erickson (1943, 1945) on *Clematis fremontii* var. *riehlii*, Fassett (1941) on

Rubus; Fassett (1942) on *Diervilla*, McClintock and C. (1946) on *Teucrium*, Anderson and Whitaker (1934) on *Uvularia*, Anderson (1936a) on *Iris*, Clausen, Keck and Hiesey (Clausen et al. 1940, 1945) on *Potentilla*, Fassett (1943) on *Juniperus*, Epling (1944) on *Lepechinia*, Anderson (1946, 1949) on introgression, Fisher (1936) on *Iris*, Anderson and Abbe (1934) on Betulaceae, Davidson (1947) on polygonal graphs for simultaneous presentation of several variables, Clausen (1922) on *Viola* and Anderson (1936b) on the hybrid index.

Edgar Anderson (Stebbins 1978; Heiser 1995) was a plant geneticist whose work well reflects the impact that genetic understanding had on classical plant taxonomy, e.g. in his studies of introgression (Anderson 1949). He advocated more detailed quantitative morphological studies of plant populations (Anderson and Turrill 1935; Anderson 1941) on the basis that classical taxonomy provided only a very sketchy idea of the true nature of morphological variation within species. He was also innovative in inventing techniques for tracking the occurrence in natural populations of several genetically determined morphological characters simultaneously (Anderson 1949), e.g. his much-copied pictorialized scatter diagrams. This was a way of graphically labelling the points in a scatter plot so that the patterns of more than two characters could be seen simultaneously. These were purely graphical ways of handling multivariate information. His hybrid index was another non-statistical but multivariate technique for handling a larger number of characters with the aim of making rough estimates of the frequency of hybrids in natural populations by a rapid inspection of several characters simultaneously.

2.1.3 Phase 3: Multivariate Statistics and Morphometrics

The problem of clustering individuals (classification *sensu* Gordon 1999) using several to many variables in a more formal (algorithmic) way did not attract wide attention from mathematically-minded biologists until the 1950s, with the birth of numerical taxonomy and the invention of electronic computers. This phase led directly to the development of morphometrics as we currently understand it.

Until then, statistical methods had been used to investigate quantitatively the patterns of discrete and continuous characters mainly in connection with genetic studies. In these the focus was on the hereditary transmission of features, whether for the purpose of breeding and improvement of domesticated plants and animals, or for the purpose of understanding evolutionary processes at the fine scale. The taxonomic context of species as the key units was for the most part taken as given, although the new biosystematic studies often provided evidence of complexity which could confound or at least blur the species delimitations of traditional taxonomic revisions. The important point here is that the focus was on the genetic transmission of individual traits rather than attempting the discrimination of groups of individuals described by several or many characters simultaneously – i.e. multivariate analysis aimed at taxonomic delimitation.

Nevertheless, progress in genetics and the realization that a continuous phenotypic trait was very often influenced by more than one gene, led to methods like multiple

regression in which the variation of a single dependent variable is analysed in relation to several or many independent ones. In the 1930s other important methodological foundations were laid for multivariate analysis. Fisher (1936) invented the discriminant function using four different continuous measured flower characters from E. Anderson's genetical work on North American irises (Anderson 1936a), now one of the world's most widely used test data sets in statistics. Hotelling (1931) provided a multivariate test of the difference between groups by generalizing Student's T as Hotelling's T^2 and developed canonical correlation analysis for comparing two data sets (Hotelling 1935, 1936). Mahalanobis (e.g. Mahalanobis 1936) developed the D^2 generalized distance metric, an important measure of multivariate distance which takes into account covariation between the variables (Dasgupta 1995). For Reyment et al. (1984), Mahalanobis's distance "provides the only realistic measure of multivariate distance."

Although originally presented mathematically by Pearson et al. (1901), principal component analysis (PCA) also began to be developed at this time (Reyment et al. 1984), impelled by a focus on the study of allometry, on which Huxley (1932) had published a pioneering monograph. Multivariate algebra using linear dimensions as variables dominated efforts to analyse shape quantitatively. Teissier (1938) was a key early explorer in using PCA for a multivariate approach to allometric changes and this was later developed further by Quenouille (1952), Jolicoeur and Mosimann (1960) and Pearce (1965), among others (see discussions in Sprent 1972; Reyment et al. 1984, 1985, 1991, 1996).

A key publication in the emergence of morphometrics as a distinct discipline was the book *Multivariate Morphometrics* (Blackith and Reyment 1971; second edition Reyment et al. 1984). During the period covered by the two editions, numerical taxonomy was an important area of development and debate in taxonomy and to some extent this is evident in these texts, where morphometrics was seen as just one element of radical changes taking place in the methodology of taxonomy at that time. In the 1960s the numerical taxonomy movement (e.g. Sneath 1961; Sokal and Sneath 1963) had an almost revolutionary effect on taxonomists and during this period, as in the 1970s and 1980s when the cladistics movement became ascendant, almost every facet of taxonomic theory and practice came under scrutiny, e.g. Cain and Harrison (1960), Sneath (1961), Sokal (1961), Cain (1962), Sneath (1962), Sokal (1962), Camin and Sokal (1965), Mayr (1965), Farris (1967), Rohlf (1967), Gabriel and Sokal (1969).

Morphometrics can thus be viewed, at least in its earlier phase, as a subset of a much broader enterprise to quantify the procedures of taxonomy. Both numerical taxonomy and cladistics aimed at a general theory and praxis for taxonomy to replace traditional procedures and to take advantage of the calculating power of (electronic) computers which had then only recently started to become available to researchers. They were concerned with procedures to generate taxonomies at all levels and most discussion and controversy at this period centred on the use of discrete morphological character information for this purpose.

Morphometrics — a term introduced by Blackith (1965) and further popularized by Blackith and Reyment (1971) and Pimentel (1979) — has always been concerned

with applying multivariate statistical methodology to the analysis of morphological data sets, using such techniques as principal component analysis, factor analysis, principal coordinate analysis, discriminant function analysis, canonical variate analysis, canonical correlation analysis, non-metric multidimensional scaling, correspondence analysis, etc. Until the advent of geometric morphometrics, the morphological data usually consisted of linear and sometimes angular measurements, i.e. continuous (real number) variables. Numerical taxonomy and morphometrics both arose out of a need to quantify differences between groups of organisms described with multivariate data, i.e. vectors of values of many characters for each organism. Measures of resemblance were developed such as association coefficients and distance coefficients, which involved the transformation of a set of original data, be they continuous or discrete or even non-numeric, into real-numbered scalar values representing the degree of resemblance between every possible pair of organisms in the study. Gower (1971) (see also Gower 2008) developed methods for combining discrete and continuous variables to produce distance matrices which could be analysed by scaling methods, especially principal coordinate analysis (Gower 1966) and this provided a bridge between studies using discrete data and those dependent on continuous data.

Subsequent developments, including the rise of cladistics, phylogenetic taxonomy and molecular systematics, overtook numerical taxonomy as a methodology for constructing the taxonomic hierarchy. Multivariate morphometrics (often termed “traditional” morphometrics, Marcus 1990) has always been more relevant for studies of the lower taxonomic categories and is most used by taxonomists for exploring complex patterns of variation at and below species level. However, taxonomy is by no means the only, nor even the most important area of application of morphometrics. Reyment et al. (1984) discuss a wide range of examples of the application of multivariate morphometrics to biological problems and the computer package PAST and associated literature (Hammer et al. 2001; Hammer and Harper 2008; Hammer 2012) illustrate a diversity of practical applications. A major use of morphometrics has always been the correlation of morphometric variation patterns with those provided by other, non-morphological data, e.g. in ecological and palaeontological studies, such as soil characteristics, chemical composition of rocks, time and space, etc.

2.1.4 Phase 4: Geometric Morphometrics

“Whether broadly ... or narrowly ... construed, morphometrics clearly has something to do with the assignment of quantities to biologic shapes” — thus begins an important review of the field in the early days of geometric morphometrics (Bookstein 1982), a major transformation of morphometrics that took place in the late 1970s and 1980s which resulted in a new field of research (Mitteroecker and Gunz 2009). A recent text is Zelditch et al. (2012) and MacLeod’s (MacLeod 2012a, b) beautiful and explicative PalaeoMath webpage is an essential source. The most important feature of this “revolution” (Rohlf and Marcus 1993) was to use configurations of landmark coordinates as the basic data, rather than linear measurements. Although many tradi-

tionally used measurements were made between points on the organism that could be considered biologically homologous by the accepted criteria of homology (e.g. Rieppel 1988), once gathered this data no longer preserved the geometric relationships of those points nor their covariation. Landmarks became the new name for selected homologous points on the organisms or biological structures under study and the data consisted of the configurations of the set of landmark coordinates recorded for each study object, each configuration consisting of the same number of landmarks.

A crucial advance that paved the way for geometric morphometrics was a clearer definition of shape: “Shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object” (Kendall 1977; Dryden and Mardia 1998; Kendall et al. 1999). This resolved the problem of differentiating size and shape parameters that had plagued traditional morphometrics and with the general acceptance of centroid size as the most suitable measure of size, the way was open to define geometric morphometrics as the quantitative study of biological shape and its variation.

As recounted by Bookstein (1993), the solution to the problem of how to place the statistical analysis of landmark coordinates on a firm mathematical foundation was the result of independent theoretical work by C.R. Goodall, F.L. Bookstein and D.G. Kendall, synthesized in Bookstein (1986) and Bookstein (1991). Fundamental to this was the development of appropriate ways of “filtering” location, size and rotation from the landmark coordinate data so as to produce data matrices containing information only on shape variation. The solution was general Procrustes analysis (see Zelditch et al. 2012 for a recent account). Then Bookstein (1989), in an inspired move, introduced the thin-plate spline (TPS) metaphor from physics to obtain a mathematical interpolation for the space between the landmarks. Thus, in a stroke, the problem of operationalizing D’Arcy Wentworth Thompson’s famous morphological transformations was solved (Thompson 1917, 1942). Bookstein derived parameters — principal warps, partial warps and later relative warps — which provided the tools for statistical analysis of the “bending energy” that measures the deformations represented by the study objects when these were considered as shape deviations from the mean shape of the study data set (Bookstein 1991). One reason why the TPS approach has been so influential is because it communicates visually the results of morphometric analysis by reconstructing shapes derived from analysis, e.g. mean shapes of populations, or trends in shape along principal component axes — this was something that multivariate morphometrics had never achieved.

The solid theoretical foundation of geometric morphometrics using landmarks has made it the preferred approach for morphometric analysis in studies where the organisms and their component parts offer sets of biologically meaningful landmarks. This has made the extension of such studies into three dimensions relatively easy especially in anthropological research (e.g. Weber and Bookstein 2010). Bookstein (1991, page 63) gives a detailed discussion of landmark types.

Landmark geometric morphometrics is limited by the need for biologically meaningful homologous points on the structures of interest, but these are often few or lacking in leaves, still the focus of most botanical studies. An alternative approach that developed alongside landmark studies was the analysis of outline shapes using

various kinds of functions, including Fourier's harmonic series. Although less well-founded theoretically, outline analysis has nevertheless been important in the absence of alternatives (Lestrel 1997). Recently, however, semilandmarks have come into prominence as a means of capturing the shapes of homologous structures which exhibit digitizable contours between homologous landmarks (Silva et al. 2012; Zelditch et al. 2012; MacLeod 2013; Gunz and Mitteroecker 2013). They have the advantage of making it possible to use the landmark approach, with all its mathematical advantages, to capture the shapes of contours. Zelditch et al. (2012) note that semilandmarks have less degrees of freedom than landmarks and this needs to be borne in mind in analysis. Rohlf's tpsRelw software for analysing partial and relative warps (Rohlf 2010b) performs a sliding procedure which optimizes the position of semilandmarks in relation to the bending energy of the deformations (Bookstein 1997).

Elliptic Fourier analysis (EFA) and Eigenshape analysis are two commonly used techniques of outline analysis (Rohlf 1986; MacLeod 2012a, b). In these methods points (usually a large number) are sampled along the outline but without assigning any homology meaning to them. The coordinates of these points are then used to fit a function, resulting usually in a vector of coefficients of the function. Each such vector mathematically describes its corresponding leaf outline. A matrix of such vectors thus constitutes a dataset expressing the shape variation of a population of outlines, after some appropriate standardization to remove the effects of position, scale and rotation. Outline morphometrics, like landmark techniques, can be used to produce reconstructed shapes, i.e. visually comprehensible results from the analysis of the dataset. Among the various computer implementations of Elliptic Fourier analysis (EFA) are software tools developed by Rohlf (2005), Slice (2008) and Bonhomme et al. (2014).

Other considerations have also played a key part in the success of geometric morphometrics. F.J. Rohlf developed the TPS series of free software tools (e.g. Rohlf 2010a, b, c) which have made geometric morphometric analysis easily available to anyone. His NTSYSpc multivariate analysis package handles these forms of analysis. Rohlf (2014) also created the morphometrics website at the State University of New York, Stony Brook, which remains a key factor in enabling the spread of the new techniques and analytical tools around the world; among much other useful information it includes a glossary of terms used in morphometrics (Slice et al. 2015). Meetings and courses on morphometrics are regularly advertised on the Stony Brook website and there is a comprehensive list of books which map the development of the field and provide in-depth reviews of theoretical issues as well as practical applications.

2.2 Morphometric Analysis of Leaves

Detailed and well-illustrated treatments of leaf descriptors used by taxonomists are available in several recent publications (e.g. Hawthorne and Jongkind 2006; Ellis et al. 2009; Gonçalves and Lorenzi 2011). Typically, leaves vary considerably, even



Fig. 2.1 Variation in leaves taken from a single specimen of *Quercus nigra*

within a single plant, and sampling must take this into account (Fig. 2.1). Hearn (2009), in a study of 2,420 leaves from 151 species, found that an accurate estimate of leaf shape in a species required a minimum of 10 leaves. Blade shape alone is often insufficient to diagnose a species and other leaf features may be needed. In addition, methods of analysis have to be selected according to the problem to be tackled, e.g. when distinguishing leaves of rather similar overall shape, landmark methods are most suitable, but comparison of leaves of very different shape is likely to be more successful using qualitative characters.

2.2.1 Analysis of Conventional Botanical Descriptors

There is a very large literature on the multivariate analysis of taxonomic leaf descriptors, usually in combination with characters from other organs or data types. A good example is the study by Joly and Bruneau (2007) on North American *Rosa* (Rosaceae), which combined morphometric data with cytological and molecular information. Henderson (2011) based his species delimitations of the genus *Geonoma* (Arecaceae) on morphometric analysis of a wide range of taxonomic descriptors, including leaf characters, in an unusually complete implementation of morphometrics as the basis for the species taxonomy in a major monograph.

In most published studies the data was gathered by manual methods, but there are some publications reporting work in which semi-automatic methods of data capture were deployed. West and Noble (1984), White et al. (1988) and McLellan (2000) combined manually guided electronic digitisation of leaf outlines with computerised measurement of conventional leaf descriptors such as length and breadth. Corney et al. (2012) extended this approach by investigating the automatic extraction of taxonomic characters from specimen images, necessitating image processing methods to automatically recognize leaves within images.

2.2.2 Analysis of Leaf Outline Shape

Although variation in leaf outline shape (Fig. 2.2) is traditionally described using a large set of qualitative botanical terms (see e.g. Stearn 1992), many attempts to quantify leaf shape have been made — Melville (1937) and Melville (1951), for example, presented an early manual method using grid coordinates. The development of powerful personal computers and new techniques has made outline analysis more easily available to modern researchers.

Elliptic Fourier analysis (EFA) is the most frequently used technique for quantifying leaf shape (e.g. Jensen et al. 2002; Andrade et al. 2008, 2010; Hearn 2009; Lexer et al. 2009; Chitwood et al. 2014). The digitized outlines of the sampled leaves are decomposed by Fourier analysis into a finite series of ellipses, each represented by four coefficients. Typically, 20–40 harmonics are used, which results in a mathematical description of each leaf as a vector of 80 to 160 numbers (Fig. 2.3). Principal component analysis (PCA) is then usually used to reduce the number of variables and demonstrate the major trends of shape variation in the data set. Extreme shapes along principal components and mean shapes can easily be visualized by reconstruction of shapes from the computed Fourier coefficients (EFDs).

Eigenshape analysis (Ray 1992; MacLeod 1999) is a related technique which uses data obtained from the sequence of angular deviations that occur when progressing along the leaf contour from one digitized point to the next. Principal component variables are then obtained from the data using singular value decomposition. Start and end points of the contours are usually represented by landmarks.

Also related are Contour Signatures, which are sequences of values calculated at successive points along the outline. The centroid-contour distance (CCD) consists of

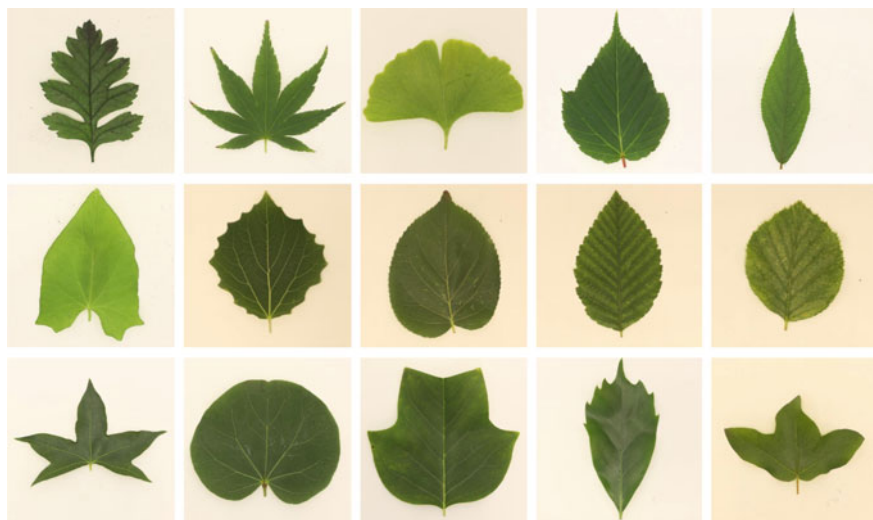


Fig. 2.2 Examples of leaf shapes

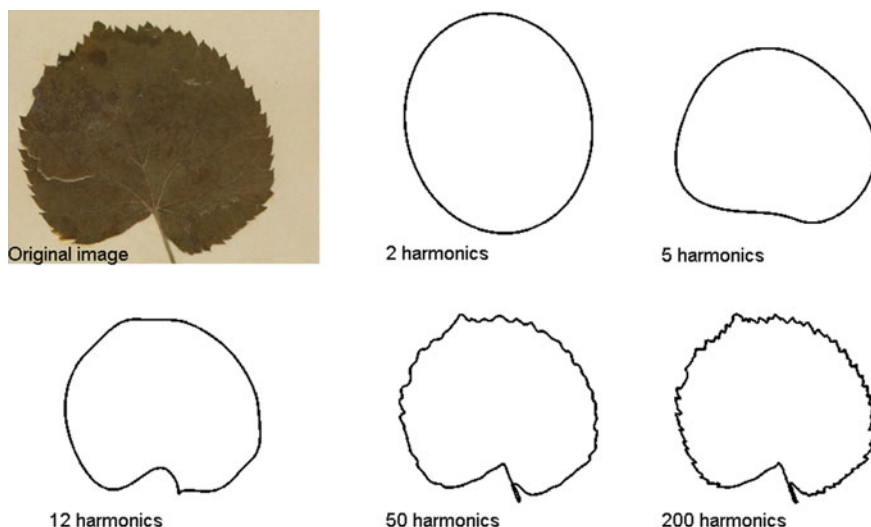


Fig. 2.3 An example of elliptic Fourier analysis. As more harmonics are used to reconstruct the original outline, more detail is preserved

the sequence of distances from the shape's centre to the points along the outline. Other signatures include the sequences of centroid angles and tangents to the contour. The aim is to represent the shape as a numerical vector, independent of the orientation and position of the object within the measuring frame. Various authors have developed refinements of the contour signature approach (Meade and Parnell 2003; Wang et al. 2000, 2003; Ye and Keogh 2009).

Overlap (self-intersection) of leaf parts, especially common in strongly lobed leaves, poses a significant problem for automating the capture of leaf contours. Mokhtarian and Abbasi (2004) and Mokhtarian (1995) developed a method for resolving this difficulty which involved identification of overlap areas using back-lighting and curvature scale space to compare outlines.

Single Parameter Shape Features are quantitative shape descriptors that are intuitive, easy to calculate and of general application. Examples are aspect ratio of leaf blade, ratio of petiole and blade length (Ashby 1948), rectangularity, circularity and perimeter-to-area ratio (McLellan and Endler 1998; McLellan 2005), and invariant moments (Hu 1962; Teague 1980). Methods of analysing such data include the "move median centres" hypersphere classifier (Du et al. 2007; Wang et al. 2008) and neural networks (Wu et al. 2007). Reconstruction of shapes from single parameter shape features is usually not possible and in general the comparison of leaf shapes using only a few such parameters is often unsatisfactory, due to confounding factors such as high correlation between them (McLellan and Endler 1998).



Fig. 2.4 Examples of leaf margins

2.2.3 Analysis of Leaf Margin Patterns

Leaf margins (Fig. 2.4) exhibit a wide morphological variation much used by taxonomic botanists for diagnosing species and genera due to the diverse forms and venation types of the marginal “teeth” (Ellis et al. 2009). As leaf teeth features such as shape, size and number are correlated with climate and growth patterns, they are important for palaeobotanists making inferences about prehistoric climates of fossil species (Royer and Wilf 2006). Hitherto, little use of these features has been made in automated analysis. McLellan and Endler (1998) created an “incision index” as a measure of roughness and Wang et al. (2003) used angular measurements to study marginal variation.

2.2.4 Analysis of Homologous Landmarks and Semi-landmark Configurations

Despite the importance of landmark-based techniques in modern morphometrics in biological research studies (Slice 2005; Weber and Bookstein 2010; Cardini and Loy 2013), relatively few botanical applications have been published and this is probably because most interest has focussed on leaf shapes, which have few good landmarks (Jensen et al. 2002; Volkova and Shipunov 2007; Magrini and Scoppola 2010; Viscosi and Cardini 2011; Klingenberg et al. 2012; Silva et al. 2012; Duminil et al. 2012).

Although landmark methodology has a very well-founded mathematical framework, as mentioned earlier, there are some limitations. Recognition of landmarks from images normally requires manual intervention and this restricts scaling up such studies by automatic extraction. Landmark approaches are also restricted to comparisons between similar leaf shapes, because the technique depends on projecting configurations onto a planar tangent space from non-linear Kendall shape space (Zelditch et al. 2012). Most multivariate statistical methods applied to analysis of landmark configurations use linear combinations of variables, and are only valid for the central part of the tangent projection plane where distortion is at a minimum. Leaf

shapes can vary greatly within a single genus (e.g. *Philodendron* in the Araceae) and for these kinds of comparison, landmark methods are unsuitable.

2.2.5 Fractal Dimensions and Polygon Fitting

Fractal dimensions have been used in a few studies for quantitative comparison of leaf shapes (Borkowski 1999; McLellan and Endler 1998; Plotze et al. 2005; Backes and Bruno 2009). Plotze et al.'s study is especially interesting because they compared very different leaf shapes within a single genus, *Passiflora*. However, in general fractal dimension measures alone are probably insufficient to capture and explain enough of the shape variation. Modelling leaves by fitting and comparing polygons has been investigated by Du et al. (2006) and Im et al. (1999).

2.2.6 Analysis of Leaf Venation Patterns

Leaf venation — the pattern of veins at larger and smaller scales — is almost as important in orthodox plant taxonomy as shape for characterizing plant species and genera (Ellis et al. 2009) and of particular importance for leaf fossil identification (Fig. 2.5).

Various methods have been used to capture venation patterns from leaves using image processing methods. Clarke et al. (2006) used scale-space, smoothing and edge detection algorithms. Li et al. (2006) used Independent Component Analysis (ICA, Comon 1994). Mullen et al. (2008) used artificial ants as an edge detection method. Fu and Chi (2006) combined thresholding and neural networks and achieved a good result. Kirchgessner et al. (2002) used b-splines to represent veins and Plotze et al. (2005) used a Fourier high-pass filter combined with a morphological Laplacian operator.

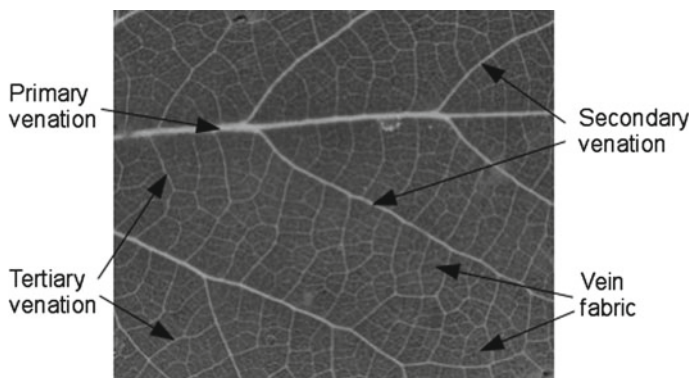


Fig. 2.5 Leaf vein structure



Fig. 2.6 Examples of vein structures

Classification of the extracted vein pattern data has been attempted by Park et al. (2008) using end points and branch points of veins and by Chitwood et al. (2014) using landmark geometrics.

Extraction and comparative analysis of leaf venation patterns continues to be a significant challenge, especially as a technique for large scale implementation. Automated methods are still a long way from realizing the full potential of these features for identification (Fig. 2.6).

2.2.7 Analysis of Leaf Texture

Various studies to capture and compare leaf texture for identification purposes have been published and appear to be most useful when combined with analysis of leaf outline shape. Backes et al. (2009, 2010) used multi-scale fractal dimensions and tourist walks, and Casanova et al. (2009) used Gabor filters. Liu et al. (2009) applied wavelet transforms and support vector machines while Lexer et al. (2009) estimated reflectance from mean greyscale values of leaf images.

2.2.8 Analysis of Other Features of the Leaf Blade

Some preliminary attempts to use 3D imaging and modelling in leaves have been made. Ma et al. (2008) reconstructed leaves and branches from 3D data captured by scanner. Teng et al. (2009) modelled 3D structure and segmentation by combining 2D images, and classified the results into major shape classes. Song et al. (2007) used stereo matching and a self-organizing map to model plant surfaces.

2.3 Morphometrics of Flowers and Other Plant Organs

A number of methods have been proposed to identify plants from digital images of their flowers. Although colour is a more common distinguishing feature here, many methods used to analyze leaf shape can also be deployed. Nilsback and Zisserman

(2007) combined a generic shape model of petals and flowers with a colour-based segmentation algorithm. The end result was a good segmentation of the image, with species identification left for future work.

Das et al. (1999) demonstrated the use of colour alone to identify a range of flowers in a database related to patents covering novel flower hybrids. Their method allows the database to be searched by colour name or by example image, although no shape information is extracted or used. A colour-histogram segmentation method was used by Hong et al. (2004) and then combined with the centroid contour distance (CCD; see Sect. 2.2.2) and angle code histograms to form a classifier. They demonstrated that this method works better than using colour information alone to identify a set of fourteen species. This again suggests that outline shape is an important character to consider, especially in combination with other features.

Elliptical Fourier descriptors (Sect. 2.2.2) were used by Cannon and Manos (2001) on *Lithocarpus* fruits and by Iwata et al. (2004) on vegetative storage organs of *Raphanus*. Yoshioka et al. (2004) and Yoshioka et al. (2007) used EFA to study the shape of the petals of *Primula sieboldii*. Wilkin (1999) used linear measurements of floral organs, seeds and fruits as well as leaves, and principal component methods to investigate whether a closely related group of African *Dioscorea* species were morphologically distinct or not. He discovered that they formed a single morphological entity and hence all belonged to one species. Gage and Wilkin (2008) used Elliptic Fourier analysis on the outlines of tepals (petal-like flower parts) of three closely related species of *Sternbergia* to investigate whether they really formed distinct morphological entities. Clark (2009) used linear measurements of bracts (specialized leaf-like organs associated with the inflorescence) in a study of *Tilia*, and Huang et al. (2006) analyzed bark texture using Gabor filters and radial basis probabilistic neural networks.

Shipunov and Bateman (2005) used analysis of partial warps in a study of the shape of floral labella in *Dactylorhiza*. Gómez et al. (2006) used relative warps analysis to study the evolution of zygomorphy in *Erysimum* flowers. Van der Niet et al. (2010) presented a landmark analysis of zygomorphic floral shape in three dimensions using species of the genus *Satyrium*.

At a smaller scale, the growth of individual grains of barley has been modelled by 3D reconstruction from multiple 2D microscopic images (Gubatz et al. 2007). This allowed both “virtual dissecting” of the grains as an educational tool and also visualization of gene expression via mRNA localization. At a smaller scale still, Oakley and Falcon-Lang used a scanning electron microscope to analyze the vessels found in fossilized wood tissue (Oakley and Falcon-Lang 2009). They used principal component analysis to identify two distinct morphotypes, which correspond to one known and one novel plant species that grew in Europe about 95 million years ago. Butterworth et al. (2009) used sliding semilandmarks to compare the shapes of developing fibres in wild and cultivated cotton. Savriama et al. (2010) studied symmetry and allometry in cells of the alga *Micrasterias* using C.P. Klingenberg’s MorphoJ package for geometric morphometric analysis.

A number of studies have used image processing techniques to analyze root structures in the “rhizosphere” (the region in which roots grow, including the soil, soil microorganisms and the roots themselves). For example, Huang et al. (1992) used

digital images of roots captured by placing a small camera inside a transparent tube located beneath growing plants. They then used expert knowledge of root shapes and structures (e.g. their elongated shape and symmetrical edges) to combine multiple sources of information and fit polynomial curves to the roots, using a graph theoretic model to describe them. More recently, Zeng et al. (2008) used image intensity to distinguish root pixels from soil pixels and then deployed a point process to combine and connect segments to efficiently identify complete root systems.

These studies show that while the clear majority of botanical morphometrics research publications have focussed on leaves due to their ready availability and usefulness for discriminating taxa, other plant organs, when available, should not be ignored.

2.4 Applications

In this section we move beyond specific algorithms in isolation and methods designed for the laboratory to consider some complete systems and prototypes designed for practical use in the field. In order to have wider impact, it is important to demonstrate that an algorithm can be applied in practice and can be scaled up from a few idealised examples to tackle larger and more complex problems. Here we review systems designed to identify species from plant images, several agricultural applications, scientific research tools for studying species variation and distribution and applications aimed at correlating morphological and climatic variation.

2.4.1 General-Purpose Species Identification

Plant identification is particularly important at the present time because of concerns about climate change and the resultant modifications in the geographic distributions and abundance of species. Development of new crops often depends on the incorporation of genes from wild relatives of existing crops, so it is important to map accurately the distribution of all plant taxa and keep track of their ranges over time. Automated identification of plant species from digital images has become increasingly attractive because of the need to survey the dwindling biodiversity of the world much more rapidly than is possible using traditional taxonomic techniques. Satellite imaging and other such remote surveying technologies can provide useful imagery, but the plant parts thus visualized, e.g. leaves and whole plant shape diagnostics, require novel classification techniques to facilitate accurate identification, given that its ultimate foundation continues to be the orthodox botanical taxonomic system.

As previously discussed in the Introduction, the species to which a plant belongs is an index of critical practical importance. Accurately identifying an organism to its species provides access to existing data linked to the specific name. This information includes such emergent properties as geographical distribution, uses and social or economic value, phytochemistry, breeding potential, functional role in the ecosys-

tems in which it occurs, etc. The binomial botanical name is the label which unlocks the knowledge about a species that lies scattered throughout the scientific literature. A robust automated species identification system, when based on expertly identified training sets, would also encourage and facilitate the participation of a much wider range of interested people, including those with only limited botanical training and expertise, to contribute to the survey of the world's biodiversity.

A number of systems have been developed that aim to recognize plant species from the shapes of their leaves, based on algorithms such as those in Sect. 2.2.2. One such plant identification system is described by Du et al. (2006). They argue that any method based on the *global* leaf shape is likely to perform poorly on images of damaged or overlapping leaves because parts of the leaf perimeter are missing or obscured. Instead, they suggest that methods based on localized shape features are more robust for this type of task. Their system matches leaves from images by fitting polygons to the contour and using a modified Fourier descriptor with dynamic programming to perform the matching. It aims to be robust with regard to damaged or overlapping leaves, as well as blurred or noisy images. They claim 92% accuracy for their method on one sample of over 2000 "clean" images, representing 25 different species, compared with 75–92% for other methods and that their method is more robust than others for incomplete leaves or blurred leaf images.

The increasing power and availability of cheap hand-held computers, including personal digital assistants (PDAs) and smart phones, has led to a number of prototype applications. A goal widely seen as highly desirable is a portable automated system that professional botanists and interested amateurs can use to identify plant species in the field. Although this is a challenging objective, not least because of the very large number of plant species that may be encountered, there is no doubt as to its potential importance.

One major ongoing project aims to produce an "electronic field guide" to plants in the USA (Agarwal et al. 2006). The user photographs a single leaf and the system will then display images of twenty plant species that have the closest matching shape according to their Inner-Distance Shape Context algorithm, an approach which extends the shape context work of Belongie et al. (2002). A related prototype from the same project includes an "augmented reality" feature (Belhumeur et al. 2008) and provides a visual display of a herbarium specimen for side-by-side comparison with the plant in question (White et al. 2006).

The CLOVER system (Nam et al. 2005) allows users to provide a sketch or a photograph of a leaf using a hand-held computer, which then accesses a remote server. The server retrieves possible matches based on leaf shape, using several shape matching methods including an enhanced version of the minimum perimeter polygons algorithm, and returns the matches to the device to display to the user. The prototype described is demonstrated to work effectively at recognizing plants from leaves, using over 1000 images from the Korean flora, with the inevitable trade-off between recall and precision.

A similar system uses fuzzy logic and the centroid-contour distance to identify plant species from Taiwan (Cheng et al. 2007). However, this requires the user to

select various characteristics of the plant from a series of menu options, rather than using morphometric analysis directly.

Each of these general-purpose prototypes has been demonstrated to work successfully on at least a small number of species, under more or less stringent conditions. Currently, there is no such system available for everyday use, although interest remains high (Lipske 2008; Kumar et al. 2012).

2.4.2 Agriculture

Rather than trying to identify a plant as belonging to one or other species, it is sometimes sufficient to carry out a binary classification (for example, as healthy or not healthy), without needing to be concerned about the exact taxon to which it belongs. One goal of automated or “precision” agriculture (Burgos-Artizzu et al. 2010) is to allow targeted administration of weed killer, fertilizer or water, as appropriate, from an autonomous robotic tractor, not least to minimize the negative impact on the environment of large scale agriculture. To do this, the system must obviously identify plants as belonging to one category or the other, such as “weed” versus “crop”.

As is often the case with machine vision systems, variable lighting conditions can make image processing very difficult. One proposed solution is to control the lighting by building a light-proof “tent” that can be carried on wheels behind a tractor and which contains lamps inside it along with a camera. Such a system successfully distinguished between crop plants (cabbages and carrots) and weed plants (anything else) growing in field conditions (Hemming and Rath 2001). It is questionable whether carrying around such a bulky tent is feasible on a larger scale.

A similar system uses rails to guide a vehicle carrying a camera along carefully laid out plots (Gebhardt et al. 2006). Rather than carrying its own lights, the system is only used under standardized illumination conditions (e.g. bright but overcast). This system extracts shape features such as leaf circularity and area and uses a maximum likelihood estimator to identify leaves that are weeds (specifically dock leaves, *Rumex obtusifolius*) in grassland, with around 85–90% accuracy. A different system to identify dock leaves is described by Šeatović (2008) and uses a scanning laser mounted on a wheeled vehicle to generate 3D point clouds. These are then segmented to separate out leaves from their background and a few simple rules based on leaf size are used to distinguish the dock leaves from other leaves in the meadow.

A related study to distinguish weeds, crops and soil in field conditions uses morphological image processing (Soille 2000). This attempts to identify the centre of each leaf by using colour threshold segmentation and locating the leaf veins. The system finds the veins using a combination of morphological opening and hierarchical clustering. The final classification makes use of a priori knowledge about features of the target plant species. A similar system, combining morphological processing with an artificial neural network classifier, has also been suggested (Pan and He 2008). A combination of colour segmentation and morphological programming has

been used to aid the development of a robotic cucumber harvester (Qi et al. 2009). A variety of methods to distinguish various crops from weeds and soil are discussed by Burgos-Artizzu et al. (2010), including colour segmentation and morphological processing. This paper also provides a useful overview of research into “precision agriculture”, which aims to use modern technology to optimize crop production, allowing for local variation in soil, landscape, nutrients and so on.

2.4.3 Intraspecific Variation, Geographical Distribution, Climate, Phylogeny

It has long been known that the climate in which a plant grows affects the shape of its leaves (Royer et al. 2005). Recent work has extended this by using digital image analysis to enhance botanical and climatic measurements. Huff et al. (2003) analysed leaves collected from temperate and tropical woodlands and measured the shape factor, finding a correlation with mean annual temperature. The study was then extended to a wider range of environments (seventeen in total) in North America (Royer et al. 2005). A variety of simple digital image analysis methods were used to measure semi-automatically features such as leaf blade area, tooth area and number, major and minor axis lengths. These were then compared to climatic measurements from the different field locations. Finally, correlations between leaf shape and climate were measured, confirming previous findings that plants growing in colder environments tend to have more teeth and larger tooth areas than similar plants growing in warmer environments. One of the goals of this body of work was to support analysis of leaf fossils with the aim of estimating paleoclimatic conditions. By verifying correlations between the shapes of leaves in extant plant species and their environments, it is hoped that fossil leaf shapes can be used to hypothesize how the Earth’s climate has changed in the past, at both global and local scales.

An early study by Dickinson et al. (1987) used manual digitization with a tablet to identify landmarks on leaf cross-sections and principal component analysis to analyse the resulting data. They identified geographically correlated variation between collection sites and identified intermediate forms among the specimens, suggesting the occurrence of hybridization. Work by Wilkin (1999) and Gage and Wilkin (2008) used morphometric analysis to determine species boundaries, a goal just as important as identifying the species to which a particular specimen belongs.

One of the most significant developments in comparative biology in the last 30 years has been the development of phylogenetic reconstruction methods. Phylogenetic or cladistic analysis, as this research field is called, is today dominated by the use of nucleic acid or protein molecular data, but analyses using morphology remain important, both for comparison with molecular results and because of the need to include fossils in classifications.

Phylogenetic analysis has been used mostly for establishing evolutionary relationships between taxa above the species level — higher taxonomy — while morphometric approaches have proved to be most useful at species level and below. This is reflected in the fact that phylogenetic systematics tends to use discrete (usually qual-

itative) characters, while morphometric studies predominantly focus on quantitative (real number-valued) variables.

Another difference is that cladistic analysis uses the tree paradigm and models the phylogeny of the organisms by searching the data set for optimal nested arrangements of discrete character values on the internal branches (synapomorphies). The resulting cladogram is an hypothesis of the phylogenetic history of the study organisms and depends only on the characters which have been selected by the analysis as synapomorphies. In contrast, the methods of morphometrics are more appropriate for dealing with character variation — a major confounding factor in identification of species and infraspecific taxa. Tree methods such as hierarchical cluster analysis and classification and regression trees are used, but ordination is also important. The character data is often synthesized into new variables such as principal components or distance parameters which can obscure the relative importance of different variables in the mathematical criterion ultimately used to distinguish groups of individuals. Morphometrics freely transforms original variables into a wide variety of derived mathematical forms (e.g. z-values, logarithms, square roots, etc.), while phylogenetic studies tend to use techniques which keep the characters in their original form, or merely transform their values into probabilities.

2.5 Summary

In this chapter, various morphometric methods used hitherto in botanical studies have been mentioned and the historical development of the subject has been briefly surveyed. Because of the extensive range of problems to which morphometrics has been applied, a diversity of analytical techniques has become available over the years. Even within the restricted area of using morphometrics for plant identification — the subject of this book — appropriate methods must be chosen for the task at hand. Plants are extremely diverse in structure, shape, size and colour. A method that works well in one plant group may rely on features that are absent in another, e.g. landmarks may be readily definable for species with distinctively lobed leaves, but not for those with simple unlobed leaf blades.

The very large number of plant species and the morphological variation usually found within any one of them, means that the development of automated identification systems is ultimately a large scale enterprise, requiring processing of high numbers of digital images. Automation is therefore essential since any system that requires significant manual effort, for example in tracing leaf outlines or accurately locating landmarks on images, is unlikely to be practical when scaled up to thousands of specimens. However, there will be contexts in which the user needs to be directly involved at some critical stage in the identification process: if an electronic field-guide provides e.g. ten predictions of species, rather than just one, the user will need to make the final choice (Agarwal et al. 2006). A related issue is how fast processing

speed needs to be. The user of a hand-held electronic field-guide is likely to require responses interactively and so (near) instantaneously, whereas for an identification tool to be used on a large set of images in a herbarium or laboratory, it may be acceptable to wait overnight for a comprehensive response — assuming no human interaction is needed in this case.

Computational Botany

Methods for Automated Species Identification

Remagnino, P.; Mayo, S.J.; Wilkin, P.; Cope, J.; Kirkup, D.

2017, VIII, 114 p. 38 illus., 20 illus. in color., Hardcover

ISBN: 978-3-662-53743-5