Automatic Extraction of Leaf Characters from Herbarium Specimens

David P.A. Corney[1], Jonathan Y. Clark[1], H. Lilian Tang[1] and Paul Wilkin[2]

[1] *Department of Computing, University of Surrey, Guildford, GU2 7XH, UK*
[2] *Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK*
Author for correspondence: *David Corney, d.corney@surrey.ac.uk*

Abstract Herbarium specimens are a vital resource in botanical taxonomy. Many herbaria are undertaking large-scale digitization projects to improve access and to preserve delicate specimens, and in doing so are creating large sets of images. Leaf characters are important for describing taxa and distinguishing between them and they can be measured from herbarium specimens. Here, we demonstrate that herbarium images can be analysed using suitable software and that leaf characters can be extracted automatically. We describe a low-cost digitization process that we use to create a set of 1,895 images of *Tilia* L. specimens, and novel botanical image processing software. The output of the software is a set of leaf characters. As a demonstration of this approach, we extract the length and width of a large number of leaves automatically from images of whole herbarium specimens. We show that the lengths and widths that we extract are very strongly correlated with values in a published account of cultivated species, but are also consistently smaller. We discuss some particular features of herbarium specimens that may affect the results of this form of analysis, and consider further applications to extract characters such as leaf shape and margin characters.

Keywords: automation; digitization; herbarium specimens; leaf morphology; *Tilia*.

Introduction

The world's major herbaria contain over 350 million specimens between them (Thiers, 2011), collected from all parts of the world over several centuries. Each specimen consists of a number of leaves, stems, flowers and other plant organs, mounted on thick paper and labelled. Labels typically include information about the collector, collection date, location and the taxon determination. Many specimens are of great botanical significance, are irreplaceable, and moreover, are extremely delicate. As imaging equipment, such as scanners and digital cameras, has improved in quality while also becoming cheaper, many herbaria have embarked on large-scale digitization projects (Lughadha & Miller, 2009; Yesson & al., 2007; Vollmar & al., 2010). These allow images of specimens to be shared widely over the Internet with no risk of loss or damage to the specimens themselves. It also provides a digital archive which to some extent duplicates the herbarium itself but with no risk of damage from insects or fungi.

We argue that these image sets form a significant new resource, but one that is not being used to its full potential, beyond the advantages to communication and sharing. Digital imaging also means that image processing methods can be used to automatically analyse images of specimens. This allows thorough comparison and modelling of phenotypic measurements of the specimens and the taxa to which they belong, with the advantages of high-throughput automation.

Here, we will show that we can automatically extract the leaf characters that are used by botanists in taxonomic descriptions, ecological studies, morphological studies, etc. This automation allows many specimens to be analysed, and large character data sets to be generated, easily and with minimal expert time. This latter point is especially important given the ongoing shortage of skilled taxonomists, the "taxonomic impediment" (e.g. Secretariat of the Convention on Biological Diversity, 2007). By taking measurements from thousands of leaves

across a single taxon we can derive not only descriptions of "typical" specimens or the average value of characters, but also descriptions of the full statistical distribution of each character. Otherwise, by reducing a set of values to their mean, or to a pair of values defining a range, we may discard valuable information regarding the variance or skew of the values (Gould, 1996; Jardine & Sibson 1970).

Furthermore, image processing opens up the possibility of extracting not only the linear measurements typical of botanical descriptions (leaf length, leaf width, petal length, etc.), but also more sophisticated and precise descriptions such as mathematical models of leaf shape. Our work will allow such methods to be applied directly to images of whole herbarium specimens for the first time.

Many applications become possible once the extraction of leaf characters from herbarium specimens is automated. These include the large-scale analysis of specimens to discover (or verify) taxonomic groupings; analysis to identify boundaries between taxonomic groups; the automatic discovery and indication of any specimen that appears to be mislabelled or placed in the wrong folder or cupboard in a herbarium; identification of the species of specimens that currently have only a generic determination; modelling relationships between leaf characters and climate (Royer & al., 2005); among many others.

In this paper, we limit our analysis to images of herbarium specimens of *Tilia* L., though we believe that similar results would be obtained for other broadleaved taxa. In a recent taxonomic treatment of the genus *Tilia*, Pigott (1997) states that "many of the species are based on study of herbarium material alone and little progress can be made until analyses of natural populations are completed". While such a full population-based study is clearly desirable, it is also expensive and difficult to carry out. However, we believe that analysing a large set of herbarium specimens automatically will provide new and useful information in the absence of such a study.

Manual leaf character extraction. — For centuries, botanical taxonomists have made detailed observations of leaves, flowers, fruits and other plant organs by combining simple linear measurements (such as leaf length and width) with precise descriptions using controlled vocabularies (e.g. Stearn, 1973; Ellis & al., 2009). Such measurements and descriptions form the basis of formal taxonomic descriptions and identification keys, such as those found in floras and monographs, and are also used in subsequent statistical or computational modelling. They aim to reduce subjectivity in formal descriptions, but some inevitably remains in any linguistic description. For many species, fruits and flowers provide useful characters, but leaves are more readily available so we will limit our study to leaf analysis in common with most of papers discussed below.

As an example of a formal description that we will return to later, Pigott (1997) describes the shapes of leaves for each species of *Tilia*. For example, "*T. platyphyllos*... leaves variable in size, 8−−15 × 7−−13 cm, circular to ovate, with a short point at apex". Wilkin (1999) describes the leaf shape of *Dioscorea quartiniana* A. Rich. in terms of the length of their leaflets and four distinct width measurements of each leaflet, all measured manually. Similarly, Clark (2000, 2004, 2007) used manual measurements, such as blade length, petiole length, and the presence or absence of hairs as the inputs to an artificial neural network to identify cultivated *Tilia* species.

One variation of this approach is to digitize the coordinates of the landmarks of a leaf or to manually trace the outline (Dickinson et al., 1987). From such a set of coordinates, leaf characters such as length and width can easily be derived, as can more sophisticated shape features, as we discuss below. Such work can be done using fresh leaves or using herbarium specimens, but in either case it requires considerable time and specialist skills.

Leaf shape analysis. — In recent years, there has been a rapid growth in the use of software to study leaf shape. In several studies, leaves have been chosen and photographed specifically for analysis via image processing. Fresh leaf specimens are typically selected that are free from insect damage, mechanical damage and so on and then photographed in isolation on a plain background. Using a plain background makes it easy for software to identify the leaf boundary precisely, such as by a simple threshold method. All pixels darker (or lighter) than a given value are identified as being the background; the remaining pixels are the leaf. By simplifying the task of finding the leaf, subsequent feature extraction can be carried out

automatically and reliably to generate high-quality data sets.

In contrast, the leaves found in typical herbarium specimens (Fig. 1) are often damaged before, during or after collection and mounting. They may also be very old, leading to fragility and potential further damage. The leaves typically overlap other leaves and the specimen sheets also have fruits, flowers, stems and other items mounted on them. One of the challenges of this work is to find the leaf boundaries from the complex images of whole herbarium specimens, something that until now has been largely avoided. The variety of colours that dried, mounted leaves take on, and the irregular placement of specimens on mounting sheets, makes even the usually robust threshold segmentation methods struggle.

Examples of the single-leaf approach include Hearn (2009), Bylesjö & al. (2008) and the "LeafSnap" electronic field guide (Belhumeur & al., 2008; http://leafsnap.com). LeafSnap runs on a smart phone and attempts to identify the species of a plant given a single leaf. It uses pixel-clustering to separate the leaf from the (plain) background and then models the shape using the distances and angles between many points on the boundary. The "LAMINA" software (Bylesjö & al., 2008) uses a threshold method with a region-growing algorithm to separate the leaf from the background, and then estimates the leaf length, area, circularity etc. LeafProcessor (Backhaus & al. 2010) uses a simple binary threshold to find the edge of the leaf on a plain background, which is then improved via the Canny edge detector and an active contour model, both of which we use in this work. LeafAnalyser (Weight, Parnham & Waites 2008) also starts with a binary threshold to separate a leaf from the background, and allows the user to interactively adjust the threshold paramters if required. Other applications such as SHAPE (http://lbm.ab.a.u-tokyo.ac.jp/~iwata/shape/) and the earlier MorphoSys have been used in a variety of botanical projects (e.g. McLellan & Endler, 1998); both also work in part by using the contrast between the plain background and the leaf to find the leaf boundary. Some of these applications allow the user to correct mistakes made by the automated outline detecting system, but this inevitably limits the work rate.

Whether an image is simple enough to find the leaf outline via a simple threshold method or whether it is more complex and requires more sophisticated methods, the resulting leaf boundary can then be represented and analysed in many ways. Geometric morphometric methods use the angles and distances between "landmarks" to represent shapes (Adams, Rohlf & Slice, 2004). According to Jensen (1990), the only landmarks found unambiguously in (nearly) all simple leaves are the apex of the blade and the petiole insertion point (i.e. the juncture of the blade and petiole), limiting the general applicability of these methods in leaf shape analysis. We use these two landmarks to define leaf length in our work. Jensen & al. (2002) studied *Acer* L. using the angles and distances between the manually located lobe apices and sinus bases to find the phenotypic relationship between two species and their hybrid. They compared this method with elliptic Fourier analysis, single-parameter outline descriptors and relative warp analysis, and found similar patterns in each case.

One common technique is elliptic Fourier analysis (EFA; White & al. 1988). McLellan & Endler (1998) compared EFA with several other methods for describing leaf shape. They demonstrate that it can discriminate successfully between various leaf groups and argue that it is an appropriate method as leaves typically lack the distinctive landmarks that many other methods require. Hearn (2009) used a combination of EFA and Procrustes analysis (a combination of rotation, scaling and translation) to perform species identification using a set of nearly 2500 specially-photographed leaves. Other approaches that have been applied to leaf morphology include the centroid-contour distance (Meade & Parnell, 2003; Ye & Keogh, 2009), which we use in our work; fractal dimension analysis (Plotze & al., 2005); and curvature scale space methods (Mokhtarian & Abbasi, 2004). Principal components analysis (PCA) is often applied to landmarks or to outlines, such as the LeafAnalyser and LeafProcessor applications mentioned above. The work we describe here is focussed on extracting the length and width of blades, but only after the full outline has been obtained; therefore any of these methods could be used to analyse the outlines.

Although these methods have been shown to be useful, they have not been applied to complete herbarium specimen images as far as we are aware, and are not by themselves suitable for such an application. Thus full use has not been made of this important resource.

Materials and Methods

In this section we summarize our methods of image capture; our algorithms for image processing and character extraction; and our data analysis.

Specimens and species. — We photographed every mounted specimen of *Tilia* available in the herbarium of the Royal Botanic Gardens, Kew (K), including type and non-type specimens, creating a total of 1,895 images. In particular, we did not ignore specimens that contained mostly damaged leaves or mostly overlapping leaves, despite the fact that automatic processing of these would be problematic (Fig. 1). One of our aims is to determine the difficulty of automatically processing herbarium specimens in general, rather than only specimens that have been manually selected as being likely to present fewer problems.

We assigned each specimen to a single species within *Tilia* to allow comparison with published descriptions on a species-by-species basis. To do this, we assumed that each specimen was filed in the "correct" folder in the herbarium, and that each folder was labelled with the "correct" species. In *Tilia* studies, several taxonomic groupings have been defined over many years; here, we used the taxonomy defined by Pigott (1997) with some exceptions detailed in Appendix 1.

We photographed and analysed 1,895 specimens but for the detailed analysis below, we ignored any specimen labelled as "hybrid" or "cultivated" (around 35%) and used only wild specimens as these can usually be identified with greater confidence. Our main image set therefore contains 1,127 wild-collected specimens from a total of 18 species (Table 1). Due to the vagaries of historic specimen collection, some species are relatively under-represented. In the results section we identify these, and acknowledge that any results drawn from very small samples will be unreliable.

Image capture. — We used a standard digital SLR to capture all the images, though we believe similar results could be obtained from scanned images. This provided sufficiently high resolution images (c.15Mp) within a limited budget, and was portable enough to use in a crowded herbarium. Leaves are rarely perfectly flat, which can cause shadows to be cast around their margins, potentially reducing the information available in the images. To minimise this problem, we used a diffuser screen and a flash. Basic information was also collected for each specimen and stored in a spreadsheet for later reference. This included the determined taxon, the collector's name, collection date, location and so on, and was collected from specimen labels and from the herbarium folder labels. De la Cerda & Beach (2010) describe a large-scale and efficient method for digitizing herbarium specimens and collecting such data. All of the images used here are available under licence for research purposes.

Algorithms to locate leaves. — Our software finds leaves in three phases. First, it identifies a set of all of the regions in the image that might be leaves; then it repeatedly refines and improves the initial estimates; and finally it discards poor-quality candidates until it has identified accurately a set of leaves in the image, ready for subsequent character extraction.

All software was developed in Matlab v.7.10 (Mathworks, Mass., USA) including the Mathworks Image Processing Toolbox v.7.0, on a standard desktop PC (3.1GHz CPU, 4Gb RAM). The software ("Herbarium Leaf Finder") is available for research purposes from www.computing.surrey.ac.uk/morphidas.

Stage 1: identify set of "candidate leaves"

The aim of this stage is find any and all regions of the image that could be a leaf. We define each such region as a "candidate leaf" and represent it with a continuous boundary. The goal is to quickly find approximate boundaries of many candidate leaves, leaving later stages of the algorithm to refine the boundaries or to reject candidates entirely if they are not in fact leaves. For example, this stage may falsely identify bracts, leaf fragments, stems or other "background clutter" as leaves, but these mistakes can be easily corrected a later stage. We found that separating these stages of "candidate generation" and "candidate selection" made the task much more tractable.

We use a deformable templates approach (Jain & al., 1996) optimized using a simple evolutionary algorithm (Fogel, 2006). Deformable templates have been used in a variety of object localization and retrieval problems where the targets vary in their exact form. Examples include medical image processing, tracking moving people, and handwriting recognition (Jain

& al., 1998). In each case, an initial prototype template is created from the outline of a “typical” object. This is then repeatedly moved and deformed until its boundary largely coincides with the edges found in the image under consideration (Fig. 2). We used evolutionary computing methods to search for a (near) optimal fit for the template to the edges of the image (found using the Canny edge detector; Canny, 1986). We repeated this several times to find several leaves in each image. Further details are given in Appendix 1.

Given an initial template we followed the pattern of trigonometric deformations defined by Jain & al. (1996). This produces a series of deformed versions of the template such that the bounding contour always forms a continuous, closed loop, as shown in Fig. 2a. While such deformations are theoretically powerful enough to match any shape, in practice it takes a very large number of iterations to produce a perfect match but only relatively few iterations to find an approximate match. We therefore defer the attempt to find the exact boundary of the leaves in the image to the next stage and use the deformable templates merely to find likely locations of leaves (Fig. 3).

Stage 2: refine candidate boundaries

In this stage, we take each candidate leaf produced in the previous stage and refine it so that its bounding contour lies close to the edge of the objects (e.g. leaves) on the herbarium mounting sheet.

We use a level set method (Malladi & al., 1995) to iteratively adjust the candidate leaf boundary until it corresponds closely to the high-contrast edges in the image. The level set method is closely related to active contour models, such as the snake edge detector (Kass & al., 1988), and it makes no *a priori* assumptions about the exact shape of the object being modelled. It works by maintaining a series of points (forming a “front”) on the current estimated boundary and moving them until the boundary line closely matches the edge of the region that it starts in. By initializing this process with the candidate leaves found in the previous stage we can find the leaf boundaries precisely.

The end result is a set of well-defined objects (Fig. 3), although the set will still contain a number of non-leaf objects which we filter out in the next stage.

Stage 3: filter candidate leaf set to remove non-leaf objects

The aim of this stage is to remove from our set of refined candidate objects those that are not in fact leaves. To do this we need some way of automatically distinguishing “leaf” and “non-leaf” objects. We do not want a system that is restricted to a single genus (or any other taxon) but we recognize that the diversity of leaves makes it hard to define in advance absolutely what constitutes a leaf from a purely visual perspective. To solve this, we present the

user with a very simple task: they are shown a number of candidate leaves, as produced from the previous stages, and simply have to click on a few leaves whose outlines have been correctly identified. The user does *not* have to identify the taxa nor carry out other time-consuming tasks such as drawing round the outline of a leaf. In our case, a set of 120 leaves were chosen by one of the authors in a process that took less than ten minutes. This is the only part of the system that requires user interaction.

This produces a hand-labelled subset of candidates that are known to represent leaves and can be used as a “ground truth". Any object that is found to be substantially different from these will be discarded and assumed to be incorrectly extracted objects (e.g. bracts, flowers, fruits etc.); badly damaged leaves; or the result of multiple overlapping leaves being extracted as if they were a single leaf. In this way, we correct the “mistakes” made by the earlier stages of the algorithm, which in turn makes the earlier object-locating stages less critical and easier to develop.

To compare each candidate object with the hand-labelled ground-truth leaves we convert both the outline of the hand-labelled leaves and the outline of the unknown candidate object to a centroid-contour distance trace, similar in appearance to a time-series (Ye & Keogh, 2009). The algorithm traces round the boundary from an arbitrary starting point and measures the distance to the centroid of the object at each point, as shown in Fig. 4. The distances are normalized to make the comparisons scale-invariant because for this comparison we are not interested in the absolute size of each leaf but only the shape (although clearly for the later character extraction we do use the absolute size). The candidates that have been extracted by the earlier stages can then be compared to the set of ground-truth leaves by comparing the centroid-contour

distance traces. Two objects are compared by sliding one trace past the other (equivalent to rotating one object) until the Euclidean distance between them is minimized. We are not greatly concerned with computational efficiency here, although savings could be made (Keogh & Ratanamahatana, 2005).

We assume that all healthy, undamaged leaves being examined will share broadly similar shape characteristics whereas damaged or incorrectly extracted leaves will differ in arbitrary ways. We define a simple distance metric between outlines as the Euclidean distance between their centroid-contour distance representations. We then use a threshold to reject any object that is greater than a given distance from the nearest ground-truth leaf shape. Different thresholds can be chosen, if necessary, to produce relatively "optimistic" or "pessimistic" results when rejecting invalid shapes.

Extracting leaf characters. — Published floras typically describe many characters of leaves including blade length and width. These are usually expressed as a representative range. The characters are used to identify taxa and are chosen to be easy for a botanist to measure. Whether these characters are extracted manually or automatically they must be defined precisely and this may be taxon-specific. Here, we define leaf length and leaf width as explicitly and unambiguously as we can.

We define the length of a *Tilia* leaf blade as being the straight-line distance from the petiole insertion point to the apex (Fig. 5). This is equivalent to the "midvein length" (Ellis & al., 2009, p.10) except that we explicitly use the straight-line distance and don't follow any curvature of the midvein, a subtle distinction that is often left ambiguous. Bylesjö & al. (2008) also explicitly use a straight-line distance, while acknowledging that for some leaf forms this is far from ideal. Once a length axis has been determined (on which both the apex and insertion point lie) we can then determine the *width* of the leaf. This we define as the greatest straight-line distance that is perpendicular to the length axis across the leaf blade (Fig. 5). Note that the length defined this way may not be the longest possible line across the leaf blade. Some leaves are effectively circular while others may be significantly wider than they are long and others, especially in *Tilia*, may have pronounced lobes either side of the insertion point. This contrasts with the "major axis length" described by Royer at al. (2005), which is defined as the "longest measurable line across the leaf blade", and the "minor axis length" that is perpendicular to the major axis.

To determine the length, we must therefore locate both landmarks (apex and insertion point) and the "axis" running through them. We combine several methods as we found that no single method worked across all the leaves being analysed. We assume that the midvein runs from the petiole insertion point to the apex and also that the leaf is approximately symmetric along this vein. We used the Hough transform (Ballard, 1981) to identify the midvein and combine this with a leaf-symmetry test and local morphology to verify that we have found the primary axis.

Note that in many herbarium specimens the petiole is missing completely or is folded under the leaf (and may be protruding from the "wrong" location) or else is simply hard to find against background clutter. We do not therefore rely on the location of the petiole but use the midvein and (where present) the leaf-base sinus to indicate the insertion point.

One specific problem that arises when identifying a leaf perimeter on a herbarium sheet is that many leaves are held in place by means of paper strips (Fig. 6a). These are thin white strips that are glued to the herbarium sheet at both ends and held flat across the leaf. In most cases, they are used towards the apex of the leaf to avoid directly gluing the leaf to the sheet. Any segmentation or edge-detection algorithm that is guided by colour (or intensity) will tend to treat these paper strips as part of the background sheet and so will treat any part of the leaf apex protruding beyond the strip as being a separate object. (This includes the level set method we use, as described earlier.)

We developed a specific algorithm to address this problem. Given the primary axis of the leaf (i.e. line AB in Fig. 5) defined by the midvein, we extrapolate this line beyond the main body of the leaf blade and measure the intensity profile (Fig. 6). If the apex is partially obscured by a paper strip then we expect this profile to clearly show the leaf apex beyond any paper strip. Specifically, as one moves from the blade towards the apex and beyond, the image intensity will rise, corresponding to the strip, drop again for the apex, and rise again for the (relatively pale) herbarium paper. If such an apex is found then we adjust our estimate of the primary axis of the leaf, and therefore the length, to reflect the improved estimate of the apex location.

In practice, we found that this "leaf apex finder" found a mounting strip on around 25% of the extracted leaves. In those cases, the extracted leaf lengths were extended by up to 3.0 cm (with an average of 1.2 cm). The results below use these improved length estimates. The manual and published values we discuss below are measurements taken by expert botanists, who we assume will have ignored any paper strips and measured the total blade length. Our extracted measurements can therefore be directly compared with those measurements.

Manual measurements. — Clark (2000) used a ruler to measure several leaves from each of 3 specimens of 19 species, giving a total of 222 measurements of both length and width. These formed part of the character set for a neural network-based identification study of the genus *Tilia*. These data are independent of the current study in that they were already available as raw data and were used in earlier studies (Clark 2004, 2007, 2009), and it was not envisaged that they would be used for comparison with results from a later image-processing based study. A list of the herbarium specimens from which the measurements were taken is provided by Clark (2009).

Results

To evaluate the quality of the extracted lengths and widths, we compare them with the two other sources of data: comprehensive descriptions of the species, which include leaf dimensions (Pigott, 1997); and an independent set of manual measurements of leaves from each species (Clark, 2000). We compared the lengths (Fig. 7) and widths (Fig. 8) for all the leaves found by our software (Table 1) and calculated the correlations between these three sets of measurements. We calculate the Pearson sample correlation coefficient, r and test this correlation against a null hypothesis (i.e. no correlation, where $r = 0$) and give the corresponding p-value as an indication of significance, along with the sample size n, which here is the number of species being compared.

Comparison with published values. — We consider the correlation between the extracted measurements and the measurements given by Pigott (1997). Pigott describes leaves of each species with a range of lengths and range of widths such as "8--15 × 7--13 cm" for *T*. *platyphyllos*. In common with similar publications, other descriptive statistics such as mean and variance (if the distribution is Gaussian) or percentile ranges are *not* given (*contra* Jardine & Sibson, 1970). We will treat these lower and upper bounds separately and treat length and width separately.

First, we consider the correlation between the upper range of length and the longest leaf extracted for each species. Here, the correlation is highly significant ($r = 0.8119$, $p \approx 0$, $n = 18$), suggesting that our method is correctly extracting leaf lengths from the images. However, comparing the lower range of each length and the shortest leaf extracted shows no correlation ($r = -0.0913$, $p = 0.7187$, $n = 18$). On closer examination, we found that our software tends to produce consistently *smaller* size estimates. We consider this in more detail in the discussion section but note here that herbarium specimens typically include a mixture of mature and immature leaves, whereas published descriptions tend to describe only the mature leaves. Rather than attempt the challenging task of distinguishing between mature and immature leaves, for now we simply remove the smaller leaves from the analysis. (Immature leaves are typically distinguished by size, colour and blade thickness; the first of these is the problem here while the latter two features are hard to identify reliably from images alone.) To form a more robust estimate of leaf size we must remove these smaller leaves from consideration. If we apply a strong filter and remove 50% of the smallest leaves found within each species then the correlation between the published lower range and the shortest (remaining) leaf for each species is now significantly positive ($r = 0.6324$, $p = 0.0049$, $n = 18$). Similar figures are obtained with other cut-off points. (E.g. if we remove 40% or 60% of the smallest leaves, the correlation is still significant and positive.)

Note that some of the extracted measures are based on very small samples, such as those of *T*. *kiusiana* and *T*. *mongolica* (Table 1). In such cases, our estimates of the range of sizes are clearly less certain than for species with many specimens available, and are likely to be biased towards the mean for that species.

We now repeat the analysis for leaf widths (Fig. 8). For maximum widths, we find the correlation between published and extracted widths is very strong ($r = 0.8642$, $p \approx 0$, $n = 18$),

and again we see a significant positive correlation between the minimum extracted leaf width and published leaf width ($r = 0.5113$, $p = 0.0301$, $n = 18$), after we filter out the same objects as before (i.e. the shortest 50%).

Comparison with manual measurements. — Having compared the automatically extracted leaf measurements with the published descriptions we now compare them with a sample of manually measured leaves taken from herbarium specimens (Clark, 2000).

The correlation between the manually measured lengths and the extracted lengths is significant both for the maximum lengths found for each species ($r = 0.5160$, $p = 0.0284$, $n = 18$), and for the minimum lengths found ($r = 0.7066$, $p = 0.0010$, $n = 18$). Similarly, the correlation between manually measured widths and extracted widths is significantly positive for the maximum values ($r = 0.5400$, $p = 0.0207$, $n = 18$) and for the minimum values ($r = 0.6136$, $p = 0.0068$, $n = 18$). These correlations are calculated after we remove the smallest extracted leaves, as before.

For completeness we also calculated the correlation between the manually measured characters and those published by Pigott (1997). The maximum and minimum values for both length and width had significant positive correlations ($p < 0.021$). The correlation scores for the maximum values are included in Table 2. Finally, as expected we also find a very strong correlation between leaf length and leaf width for our extracted estimates for each species in turn ($r = 0.8591$, $p \approx 0$, $n = 18$).

Discussion

We have demonstrated that we can automatically extract leaf characters from digitized images of whole herbarium specimens. To the best of our knowledge, this has not been previously achieved. This is an encouraging result and we believe that these large, readily-available image sets should be further exploited in future work.

We have shown that the extracted leaf lengths (and widths) of different species are very strongly correlated with the leaf lengths (and widths) published primarily for identification purposes by Pigott (1997) and those measured by Clark (2000). Although Pigott will have considered the specimens at Kew in the preparation of his account, undoubtedly specimens in other herbaria, together with living trees both cultivated and in habitat, formed a large part of his experience and so influenced his descriptions (see also Pigott, 2012).

As noted earlier, we found that the extracted leaf lengths were typically shorter than those published, which is why we introduced a filter to remove very small leaves from the analysis. We now discuss several possible reasons for this bias.

1. Leaves mounted on herbarium sheets will be, on average, smaller than those found on living plants if specimen collectors tend to choose smaller leaves for some reason. For example, collectors may choose stems that contain flowers and/or fruits as well as leaves; for some taxa, leaves growing on reproductive shoots tend to be smaller than leaves on purely vegetative, non-reproductive shoots, and may also vary in shape and complexity. Examples include the common holly (*Ilex aquifolium*; Obeso, 1997) and the stinging nettle (*Urtica dioica*; Oñate & Munné-Bosch, 2009), both of which show reductions in leaf area on reproductive shoots. In contrast, a botanist making observations in the field may concentrate on vegetative leaves on non-reproductive shoots leading to a larger measure of leaf size.

2. A botanist writing a formal description of a plant will typically describe the size and shape of representative adult vegetative leaves (e.g. Pigott, 1997) or the “observed natural range” (Jardine & Sibson, 1970) and will ignore immature leaves, shoot leaves, leaves on reproductive shoots, damaged leaves etc. The software described here does not attempt to make such distinctions but rather measures all the leaves it can find, which will skew the results somewhat towards the smaller end. As noted earlier, immature leaves can be recognised by size, colour and thickness but blade colour and thickness are hard to identify reliably from images of herbarium specimens alone.

3. Small leaves may be easier for our algorithm to find than large leaves. For example, by simple geometry, the larger a leaf is, the more likely it is to partially overlap another leaf, which makes it harder to distinguish the leaf boundary. Larger leaves, having larger perimeters, may

also be more prone to damage, again making automatic extraction more challenging.

4. The ranges of lengths and widths given in descriptions, such as in a flora or monograph, may not accurately reflect the distribution of leaf sizes. For example, if a range is given as "8--15 cm", we are not told whether, say, 10% of the leaves are around 15cm long or 1%. In such cases, a small sample of measurements may not contain any at the upper end of the range, skewing the measurements downwards. Without including a great deal of more information in formal descriptions it is hard to be certain. For this reason, Jardine & Sibson (1970) recommend providing a set of percentile scores describing a full distribution, but in practice this advice seems to be rarely followed. In general, if we make the common assumption that the underlying distribution of sizes is approximately Gaussian, then a small sample will inevitably be biased towards the mean rather than the extremes. This will lead to an underestimate of the maximum leaf lengths (even if the sample mean is an unbiased estimate).

5. The presence of paper mounting strips was mentioned earlier and we described a method for identifying and ignoring them. However, this may not work perfectly, meaning that the extremes of some blades may have been ignored when calculating the lengths.

While some of these issues may be specific to *Tilia*, others are likely to apply to other taxa. It is not clear how much this would affect other morphology studies based on herbarium specimens alone, but such issues should certainly not be ignored.

Turning to other aspects of this work, in Figs. 7--8, the ranges for *T. mongolica* are very small; Pigott (1997) gives the leaf size as "*c*. 4 x 4 cm" with no wider range. Only two herbarium specimens of this species were available for our analysis and only one leaf was extracted by the software. Therefore, the extracted results also take on a single value rather than a range. *T. kiusiana* similarly only had very few specimens (five) and extracted leaves (six), so the range extracted is likely to be an underestimate. All other species considered had 20--514 leaves, giving a better indication of the range of sizes and more confidence in our results.

One of our goals was to consider the full distribution of leaf sizes, as compared to only the upper and lower ranges of length and width. We use a box-and-whisker plot to help visualise the difference between the species leaf sizes. Fig. 9 shows the distributions of blade lengths of the 10 species that have the most examples in our set. Although the ranges overlap to a large extent, these plots show how the distributions vary. Many of the leaves are close to the median, with relatively few extremes, making the range (as usually published) a less helpful statistic.

We should sound a further note of caution here. Although the extracted measurements are strongly correlated with the manual and published measurements, we do not claim that our algorithms are flawless. Inspecting the extracted leaf objects, for example, reveals that even at the end of the filtering process, some objects remain that are clearly not leaves; similarly, when finding the length and width the algorithm sometimes fails to find the apex and/or petiole insertion point correctly. In both cases the values extracted will be incorrect. While this adds noise to the data the success of the correlation analysis suggests that such issues do not invalidate the approach we suggest, although it does leave room for improvement.

Having demonstrated the feasibility of automatic character extraction from herbarium specimens we are currently extending this work to extract shape information, which we expect will be more powerful at discriminating between taxa. The current approach can be seen as a working prototype that would need to be extended and modified to be able to analyse compound or highly-dissected leaves or other substantially different leaf forms; for compound leaves, it would probably be more straightforward to measure individual leaflets. The work could also be extended from leaves to find and analyse flowers, fruits, seeds etc. The software described here already locates and extracts the boundaries of leaves and given this set of coordinates it is straightforward to calculate characters such as area, perimeter and compactness, as well as more complex descriptors such as the elliptic Fourier coefficients (White & al., 1988; Zhang & Lu, 2005), the inner-distance shape context (Agarwal & al., 2006), wavelet descriptors (Gu & al., 2005), fractal dimension analysis (Plotze & al., 2005) and so on. When comparing the shape of two leaves we can use a scale-independent representation, such as elliptic Fourier analysis or the centroid-contour distance approach described earlier, making the issues of bias in size discussed above less important. For specific taxa, it may also be possible to automatically identify a range of landmarks, as we have identified the blade apex and insertion point, allowing geometric morphometric methods to be applied (Adams, Rohlf &

Slice, 2004). We are also in the process of applying and extending the methods described here to a second broadleaved genus, namely the predominantly climbing monocotyledon *Dioscorea* L., with encouraging preliminary results.

Acknowledgments

This work is kindly funded by the Leverhulme Trust (grant F/00242/H), as part of MORPHIDAS: the Morphological Herbarium Image Data Analysis research project. We wish to thank the Board of Trustees of the Royal Botanic Gardens, Kew for providing access to the herbarium specimens and for allowing reproduction of the images in this work. All of the images used are available under licence for research purposes.

Literature Cited

Adams, D., Rohlf, F. J., & Slice, D. 2004. Geometric morphometrics: Ten years of progress following the "revolution." *Ital. J. Zool.*, 71(1):5--16.

Agarwal, G., Belhumeur, P., Feiner, S., Jacobs, D., Kress, W. J., Ramamoorthi, R., Bourg, N. A., Dixit, N., Ling, H. & Mahajan, D. 2006. First steps toward an electronic field guide for plants. *Taxon*, 55(3):597--610.

Backhaus, A., Kuwabara, A., Bauch, M., Monk, N., Sanguinetti, G. & Fleming, A. 2010. LeafProcessor: a new leaf phenotyping tool using contour bending energy and shape cluster analysis. *New Phytol.*, 187(1):251--261.

Ballard, D. H. 1981. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recogn.*, 13(2):111--122.

Belhumeur, P., Chen, D., Feiner, S., Jacobs, D., Kress, W., Ling, H., Lopez, I., Ramamoorthi, R., Sheorey, S., White, S. & Zhang, L. 2008. Searching the world's herbaria: A system for visual identification of plant species. Pp. 116--129 in *European Conf. Computer Vision,* Springer.

Bylesjö, M., Segura, V., Soolanayakanahally, R., Rae, A., Trygg, J., Gustafsson, P., Jansson, S. & Street, N. 2008. LAMINA: A tool for rapid quantification of leaf size and shape parameters. *BMC Plant Biol*, *8*:82.

Canny, J. F. 1986. A computational approach to edge detection*. IEEE Trans. Pattern Anal.* 8(6):679--698

Clark, J.Y. 2000. Botanical identification and classification using artificial neural networks. PhD thesis, University of Reading, UK.

Clark, J. Y. 2004. Identification of botanical specimens using artificial neural networks. Pp. 87--94 in *Proc. IEEE Symposium CIBCB.*

Clark, J. Y. 2007. Plant identification from characters and measurements using artificial neural networks. Pp. 207--224 in: MacLeod, N. (ed.), *Automated taxon identification in systematics: theory, approaches and applications*, CRC.

Clark, J. Y. 2009. Neural networks and cluster analysis for unsupervised classification of cultivated species of *Tilia* (Malvaceae). *Bot. J. Linn. Soc.*, 159(2):300--314.

de la Cerda, I. G. & Beach, J. H. 2010. Semi-automated workflows for acquiring specimen data from label images in herbarium collections. *Taxon*, 59(6):1830--1842.

Dickinson, T. A., Parker, W. H., & Strauss, R. E. 1987. Another approach to leaf shape comparisons. *Taxon*, 36(1):1--20.

Ellis, B., Daly, D. C., Hickey, L., Johnson, K. R., Mitchell, J. D., Wilf, P. & Wing, S. L. 2009. *Manual of leaf architecture*. New York, USA:Cornell University Press.

Fogel, D. B. 2006. *Evolutionary Computation*: *Toward a new philosophy of machine intelligence*. Wiley-IEEE Press, 3rd edition.

Gould, S. J. 1996. *Life's grandeur/Full house*. London:Jonathan Cape.

Gu, X., Du, J. X. & Wang, X. F. 2005. Leaf recognition based on the combination of wavelet transform and Gaussian interpolation. Pp. 253--262 in Huang, D-S., Zhang, X-P. & Huang, G-B. (Eds.) *Advances In Intelligent Computing*, *LNCS* vol. 3645, Springer.

Hearn, D. J. 2009. Shape analysis for the automated identification of plants from images of leaves. *Taxon*, 58:934--954.

Jain, A. K., Zhong, Y. & Dubuisson-Jolly, M. 1998. Deformable template models: A

review. *Signal Process.*, 71(2):109--129.

Jain, A. K., Zhong, Y. & Lakshmanan, S. 1996. Object matching using deformable templates. *IEEE Trans. Pattern Anal.*, 18(3):267--278.

Jardine, N., & Sibson, R. 1970. Quantitative attributes in taxonomic descriptions. *Taxon*, 19(6):862--870.

Jensen, R. J. 1990. Detecting shape variation in oak leaf morphology: a comparison of rotational-fit methods. *Amer. J. Bot.*, 77(10):1279--1293.

Jensen, R. J., Ciofani, K. M. & Miramontes, L. C. 2002. Lines, outlines, and landmarks: Morphometric analyses of leaves of *Acer rubrum*, *Acer saccharinum* (Aceraceae) and their hybrid. *Taxon*, 51(3):475--492.

Kass, M., Witkin, A. & Terzopoulos, D. 1988. Snakes: Active contour models. *Int. J. Comp. Vis.*, 1(4):321--331.

Keogh, E. & Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. *Knowl. Inf. Sys.*, 7(3):358--386.

Lughadha, E. N. & Miller, C. 2009. Accelerating global access to plant diversity information. *Trends Plant Sci.*, 14(11):622--628.

Malladi, R., Sethian, J. A. & Vemuri, B. C. 1995. Shape modeling with front propagation: A level set approach. *IEEE Trans. Pattern Anal.,* 17(2):158--175.

McLellan, T. & Endler, J. A. 1998. The relative success of some methods for measuring and describing the shape of complex objects. *Syst. Biol.*, 47(2):264--281.

Meade, C. & Parnell, J. 2003. Multivariate analysis of leaf shape patterns in Asian species of the *Uvaria* group (Annonaceae). *Bot. J. Linn. Soc.*, 143(3):231--242.

Mokhtarian, F. & Abbasi, S. 2004. Matching shapes with self-intersection: application to leaf classification. *IEEE Trans. Image Process.*, 13(5):653--661.

Obeso, J. R. 1997. Costs of reproduction in *Ilex aquifolium*: Effects at tree, branch and leaf levels. *J. Ecol.*, 85(2):159--166.

Oñate, M. and Munné-Bosch, S. 2009. Influence of plant maturity, shoot reproduction and sex on vegetative growth in the dioecious plant *Urtica dioica*. *Ann. Bot.* 104(5):945--956.

Pham, D. L., Xu, C. & Prince, J. L. 2000. Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng.*, 2:315–337.

Pigott, C. 1997. *Tilia*. In Cullen, J., Knees, S.G. & Cubey S.H. (eds.), *The European Garden Flora*, Volume V, pp. 205--212. Cambridge University Press, Cambridge.

Pigott, C. 2012. *Lime Trees and Basswoods: Biology, Ecology and Distribution of the Genus* Tilia. Cambridge University Press. (To appear).

Plotze R.O., Maurício, F., Pádua J.G., Bernacci, L.C., Vieira, M.L.C., Oliveira, G.C.X. & Bruno, O.M. 2005. Leaf shape analysis using the multiscale Minkowski fractal dimension, a new morphometric method: a study with *Passiflora* (Passifloraceae). *Can. J. Bot.*, 83(3):287--301.

Royer, D. L., Wilf, P., Janesko, D. A., Kowalski, E. A. & Dilcher, D. L. 2005. Correlations of climate and plant ecology to leaf size and shape: potential proxies for the fossil record. *Amer. J. Bot.*, 92(7):1141--1151.

Secretariat of the Convention on Biological Diversity, 2007. Guide to the Global Taxonomy Initiative. *CBD Technical Series No. 30*.

Stearn, W. T. 1973**.** *Botanical Latin*, 2nd edition, Newton Abbot, UK:David and Charles.

Thiers, B. 2011. *Index Herbariorum: A global directory of public herbaria and associated staff*. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/ih/. Accessed June 10th 2011.

Vollmar, A., Macklin, J. A. & Ford, L. 2010. Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics*, 7(2)93--112.

Weight, C., Parnham, D. & Waites, R. 2008. LeafAnalyser: a computational method for rapid and large-scale analyses of leaf shape variation. *Plant J.*, 53(3):578--586.

White, R. J., Prentice, H. C. & Verwijst, T. 1988. Automated image acquisition and morphometric description. *Can. J. Bot.*, 66(3):450--459.

Wilkin, P. 1999. A morphometric study of *Dioscorea quartiniana* A. Rich. (Dioscoreaceae). *Kew Bull.*, 54(1):1--18.

Ye, L. & Keogh, E. 2009. Time series shapelets: A new primitive for data mining. Pp.

947--956 in *IEEE Int. Conf. Knowledge Discovery and Data Mining,* ACM.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A. & Culham, A. 2007. How global is the global biodiversity information facility? *PLoS ONE,* 2(11):e1124.

Zhang, D. & Lu, G. 2005. Study and evaluation of different Fourier methods for image retrieval. *Image Vision. Comput.*, 23(1):33--49.

Tables

Species	Number of herbarium specimens	Number of extracted leaves
T. cordata Mill.	86	135
T. amurensis Rupr.	31	58
T. japonica (Miq.) Simonk.	47	92
T. kiusiana Makino & Shiras.	5	6
T. mongolica Maxim.	2	2
T. platyphyllos Scop.	298	514
T. dasystyla Steven	40	57
T. americana L.	198	156
T. heterophylla Vent.	66	71
T. caroliniana Mill.	19	32
T. tomentosa Moench	116	144
T. chinensis Maxim.	59	113
T. tuan Szyszył.	31	41
T. oliveri Szyszył.	23	60
T. miqueliana Maxim.	20	38
T. mandschurica Maxim.	41	42
T. maximowicziana Shiras.	21	20
T. henryana Szyszył.	24	64
Total:	1127	1645

Table 1 Total number of leaves found for each species. We exclude all specimens labelled in the herbarium as “cultivated” or “hybrid”.

		Extracted	Published
Lengths	Published	0.8119	-
	Manual	0.5160	0.6366
Widths	Published	0.8642	-
	Manual	0.5400	0.6363

Table 2 Summary of correlations found. The top panel shows the correlations for maximum leaf length and the bottom panel for maximum leaf width. All associated *p*-values are less than 0.03. See text for details.

Figure Legends

Fig.1. Detail from a typical herbarium specimen. Leaf A is isolated from other objects and (almost) undamaged making it an ideal leaf for image processing. Leaf B is somewhat damaged and lacks an apex. Leaf C seems undamaged but parts of it are obscured by bracts making the exact outline hard to obtain. Leaf D overlaps another leaf of a similar colour making it hard to obtain the exact outline of either leaf. In some specimens, leaves may be re-arranged before imaging to simplify the layout but this is often impossible, such as when the leaves are glued down.

Fig. 2. Examples of a) a deformable template; and b) the edges found in an image. The top-left outline in Part a) shows the original, "undeformed" template, being the outline of a single leaf, along with three randomly deformed variants. Part b) shows the result of applying the Canny edge detector to a single image. This algorithm correctly identifies the edges of many of the leaves but also identifies the edges of other objects (including flowers and labels) and the interior structure of objects (including leaf veins). By fitting the template (a) to the edge map (b) we can locate leaf outlines successfully.

Fig. 3 Examples of candidate leaves being generated, refined and selected (or rejected). Parts A & B show the initial boundary estimate of two candidates by deformable templates. Parts C & D shows the refined boundary estimate of the same two candidates after applying the level set method. The candidate on the left (C) will be kept while the candidate on the right (D) will be rejected as the boundary is not sufficiently similar to previously identified leaves and so is rejected from further analysis.

Fig. 4. Representing a leaf boundary as using centroid-contour distances. a): a single leaf; b) the outline of a leaf, with four boundary points marked and lines converging on the leaf centroid; c) a "time series" trace showing the same four points and the distances from the centroid.

Fig. 5 Blade length and width. We define the blade length as the straight-line distance from the insertion point (A) to the blade apex (B). We define the width as the greatest distance perpendicular to this line (AB) that crosses both margins (at C and D).

Fig. 6. Finding the apex of a leaf when it is partially obscured by a mounting strip. The intensity profile indicates the presence of a white paper strip (B) across the darker leaf blade (A,C) relative to the pale mounting paper (D).

Fig. 7. Range of lengths. Each bar extends from the minimum to the maximum length found for a single species in the extracted values, the published ranges, or the manual measurements.

Fig. 8 Range of widths. Each bar extends from the minimum to the maximum width found for a single species in the extracted values, the published ranges or the manual measurements.

Fig. 9 Box plot comparing the leaf length distribution of ten species. Each box shows the 25^{th}-75^{th} percentile range with the horizontal bar denoting the median. The whiskers extend to the most extreme points that are not outliers; we define outliers as points that are more than twice the interquartile range from the lower and upper quartiles, and show these with separate crosses. Each species shown has at least 58 leaves identified by the software (Table 1). While the ranges of blade lengths for these species largely overlap, the distributions are clearly distinct.

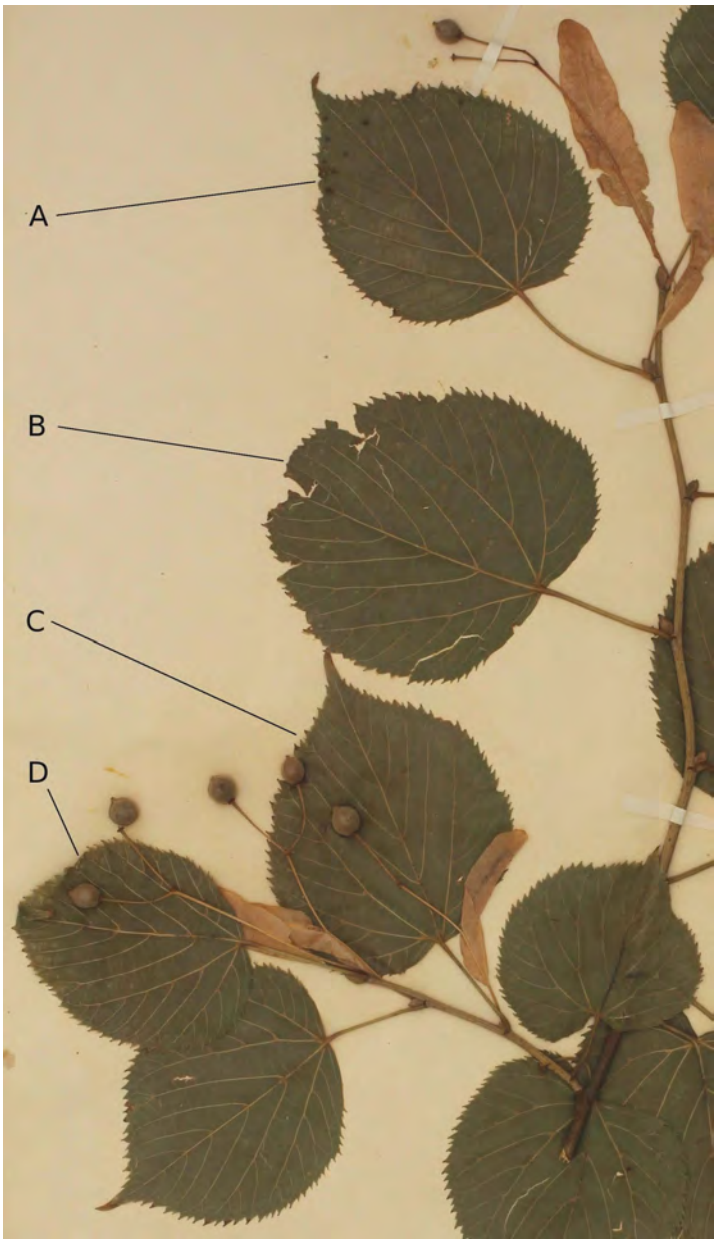
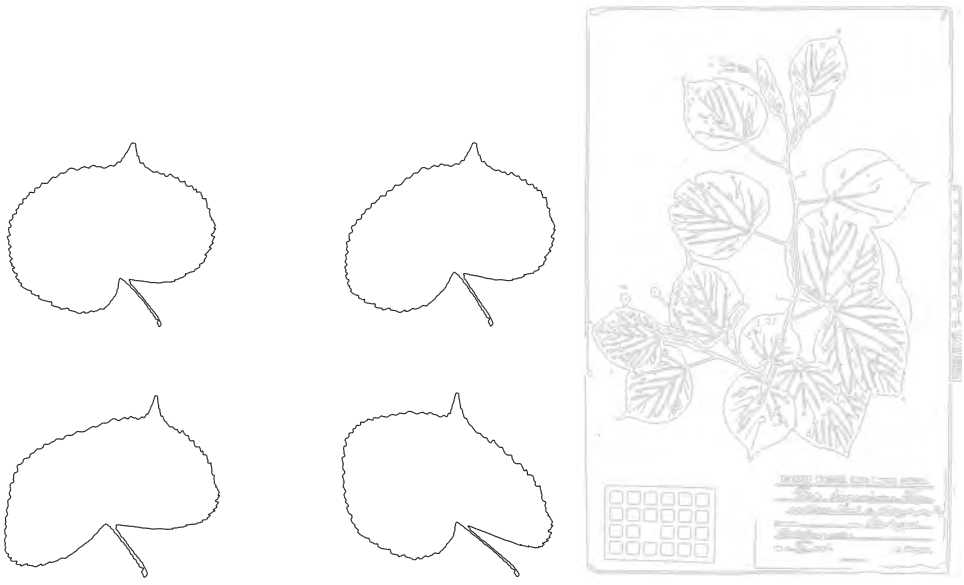


Figure 1: Detail from a typical herbarium specimen. Leaf A is isolated from other objects and (almost) undamaged making it an ideal leaf for image processing. Leaf B is somewhat damaged and lacks an apex. Leaf C seems undamaged but parts of it are obscured by bracts making the exact outline hard to obtain. Leaf D overlaps another leaf of a similar colour making it hard to obtain the exact outline of either leaf. In some specimens, leaves may be re-arranged before imaging to simplify the layout but this is often impossible, such as when the leaves are glued down.



(a) (b)

Figure 2: Examples of a) a deformable template; and b) the edges found in an image. The top-left outline in Part a) shows the original, "undeformed" template, being the outline of a single leaf, along with three randomly deformed variants. Part b) shows the result of applying the Canny edge detector to a single image. This algorithm correctly identifies the edges of many of the leaves but also identifies the edges of other objects (including flowers and labels) and the interior structure of objects (including leaf veins). By fitting the template (a) to the edge map (b) we can locate leaf outlines successfully.

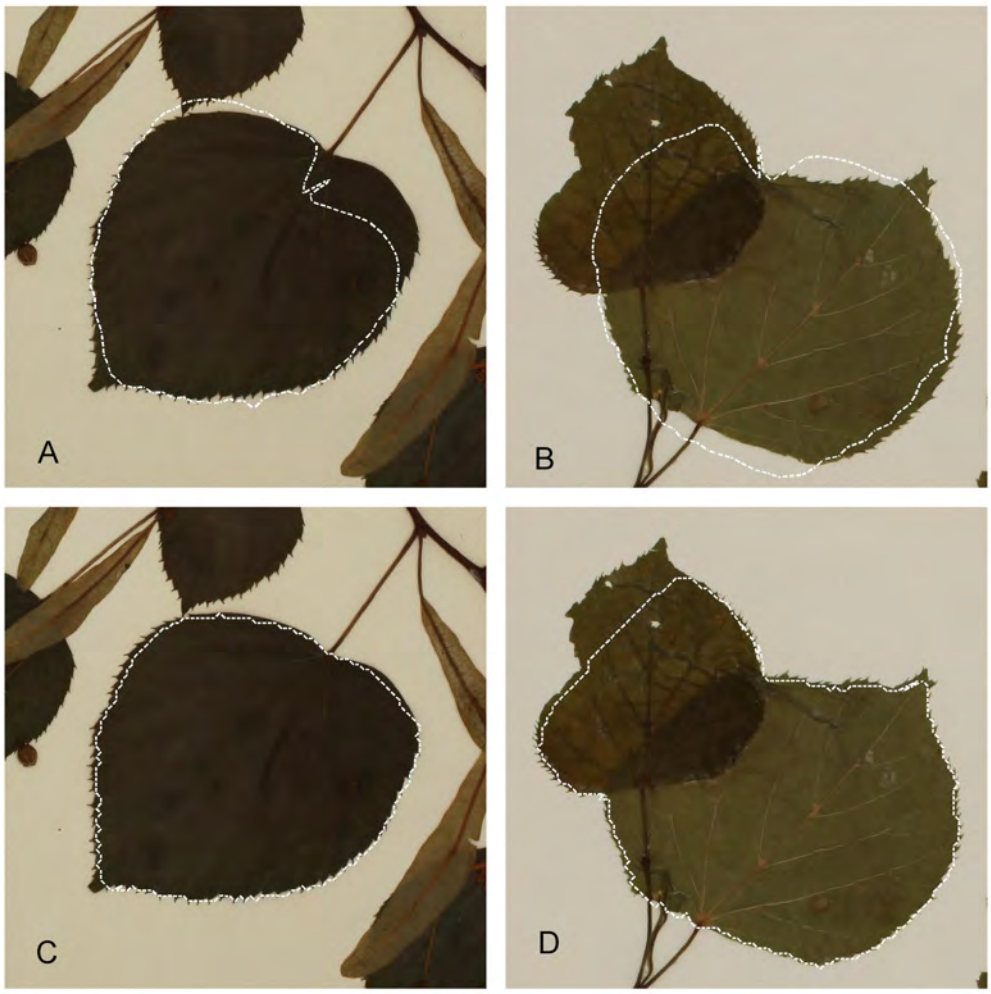
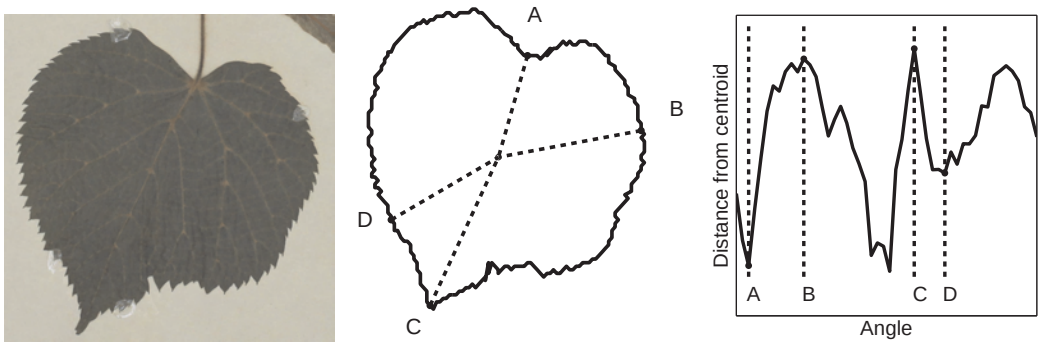


Figure 3: Examples of candidate leaves being generated, refined and selected (or rejected). Parts A & B show the initial boundary estimate of two candidates by deformable templates. Parts C & D shows the refined boundary estimate of the same two candidates after applying the level set method. The candidate on the left (C) will be kept while the candidate on the right (D) will be rejected as the boundary is not sufficiently similar to previously identified leaves and so is rejected from further analysis.



(a) (b) (c)

Figure 4: Representing a leaf boundary as using centroid-contour distances. a): a single leaf; b) the outline of a leaf, with four boundary points marked and lines converging on the leaf centroid; c) a "time series" trace showing the same four points and the distances from the centroid.

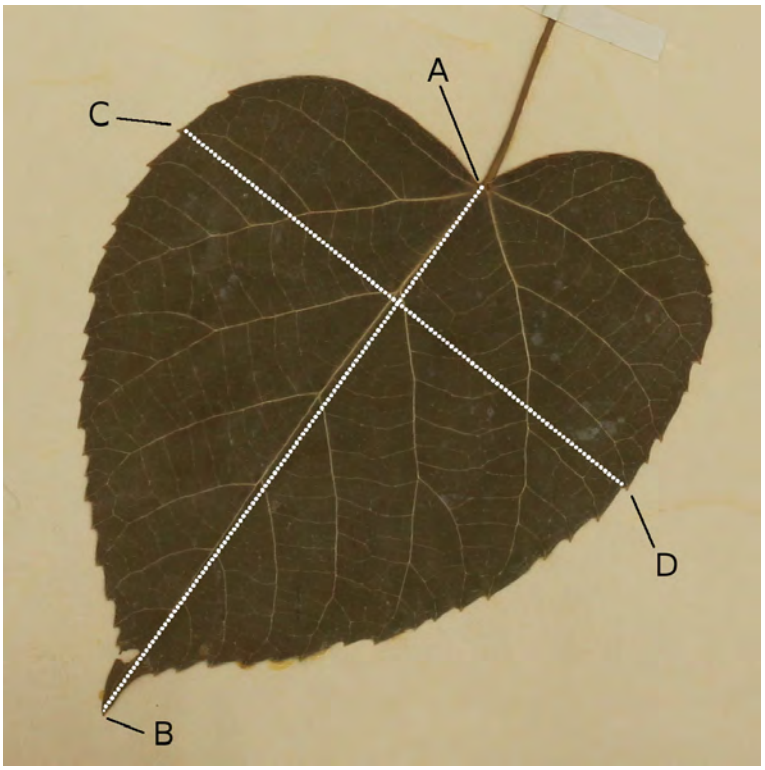
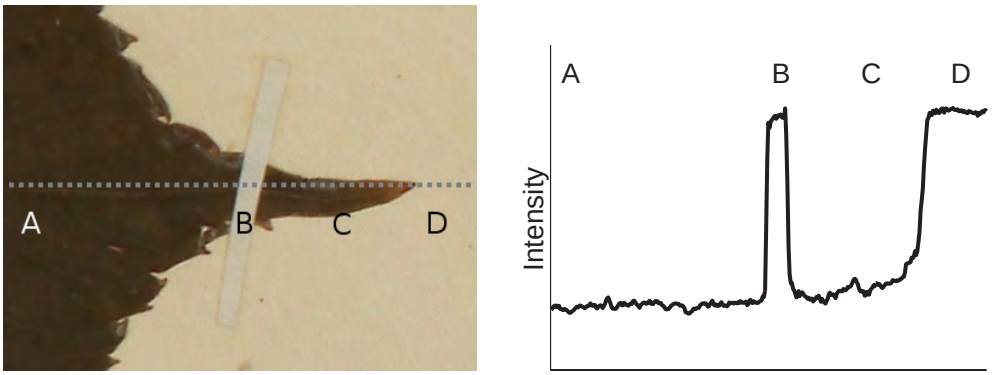


Figure 5: Blade length and width. We define the blade length as the straight-line distance from the insertion point (A) to the blade apex (B). We define the width as the greatest distance perpendicular to this line (AB) that crosses both margins (at C and D).



(a) (b)

Figure 6: Finding the apex of a leaf when it is partially obscured by a mounting strip. The intensity profile indicates the presence of a white paper strip (B) across the darker leaf blade (A,C) relative to the pale mounting paper (D).

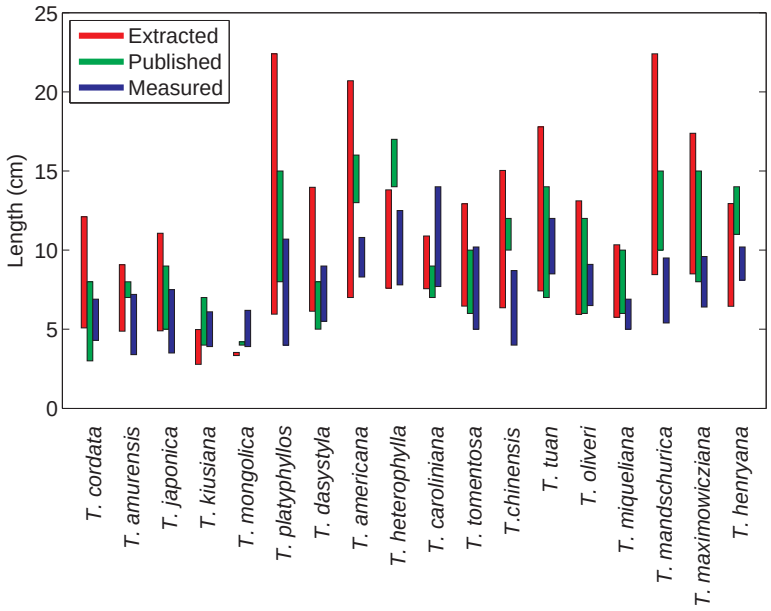


Figure 7: Range of lengths. Each bar extends from the minimum to the maximum length found for a single species in the extracted values, the published ranges, or the manual measurements.

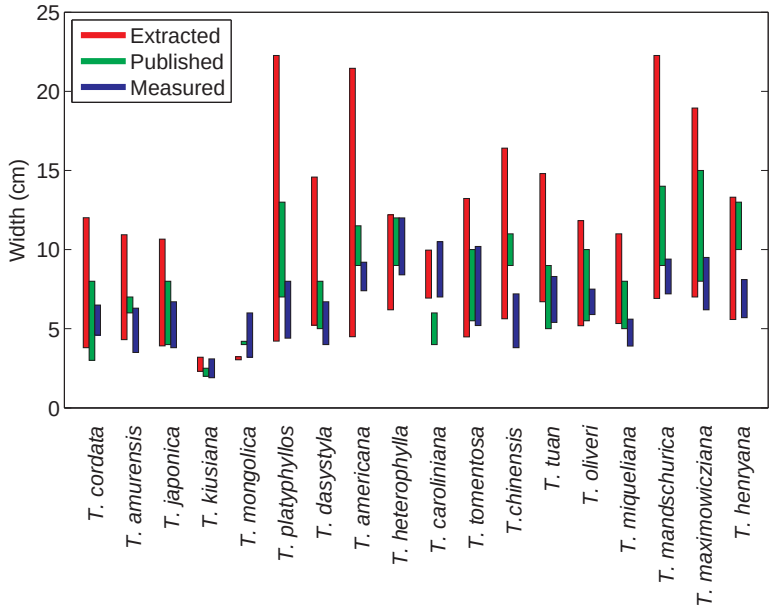


Figure 8: Range of widths. Each bar extends from the minimum to the maximum width found for a single species in the extracted values, the published ranges or the manual measurements.

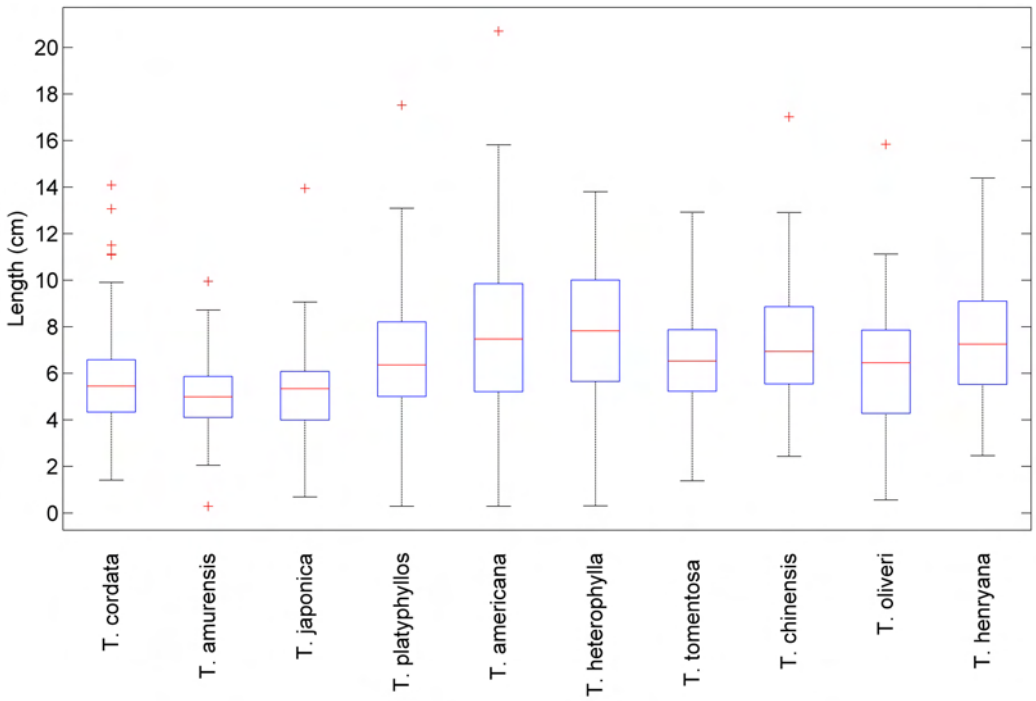


Figure 9: Box plot comparing the leaf length distribution of ten species. Each box shows the 25th-75th percentile range with the horizontal bar denoting the median. The whiskers extend to the most extreme points that are not outliers; we define outliers as points that are more than twice the interquartile range from the lower and upper quartiles, and show these with separate crosses. Each species shown has at least 58 leaves identified by the software (Table 1). While the ranges of blade lengths for these species largely overlap, the distributions are clearly distinct.