Buds morphometrics – How to distinguish and predict tree species with images of buds

Felix Nößler

March 13, 2021

Bioimage Analysis and Extended Phenotyping, Dr. Christian Kappel

1 Introduction

Digital morphometrics of plants is a tool to aid the identification of species [1, 2]. It is nowadays easy to get a high amount of pictures in short time. There is a lack of species experts [3, 4], therefore species identification with images could be a valuable tool for describing biodiversity and quantifying biodiversity change. In this study it is examined: firstly which morphometric characteristics are important to describe buds of different species and secondly how these measured properties can be used to predict the right species.

2 Material and methods

The workflow can be separated in collection of material, image acquisition, image segmentation, description of objects and the following analysis. A flowchart of the workflow is shown in the appendix (see Fig. 10).

Collection of material

Branches of trees were collected around Potsdam (Germany) at the end of January and in February of 2021. At this time all buds were closed and stopped growing. Branches from different tree individuals were chosen to account for intraspecific variability, however branches were collected only from one location per species. Eight different species were sampled: *Acer pseudoplatanus* L., *Aesculus hippocastanum* L., *Alnus glutinosa* (L.) Gaertn., *Carpinus betulus* L., *Fagus sylvatica* L., *Populus* × *canadensis* Moench, *Quercus petraea* (Mattuschka) Liebl. and *Tilia platyphyllos* Scop.. In total 1422 buds were collected. A complete list of locations and species is listed in the online appendix.

Image aquisition

The buds were broken off and placed on a sheet of paper. The white background made it easy to segment the buds. A scale was used to convert the pixel length and area to centimeters and square centimeters. The same light source was used for all of the pictures (see Fig. 2). However, no colour palette was used as a reference. A common smartphone was used to take the in total 46 pictures.

Figure 2: Setup for taking images

Image Segmentation

The scale was segmented with Otsus Method implemented in van der Walt *et al.* [5] and detected with the approximate size of the scale. Most of

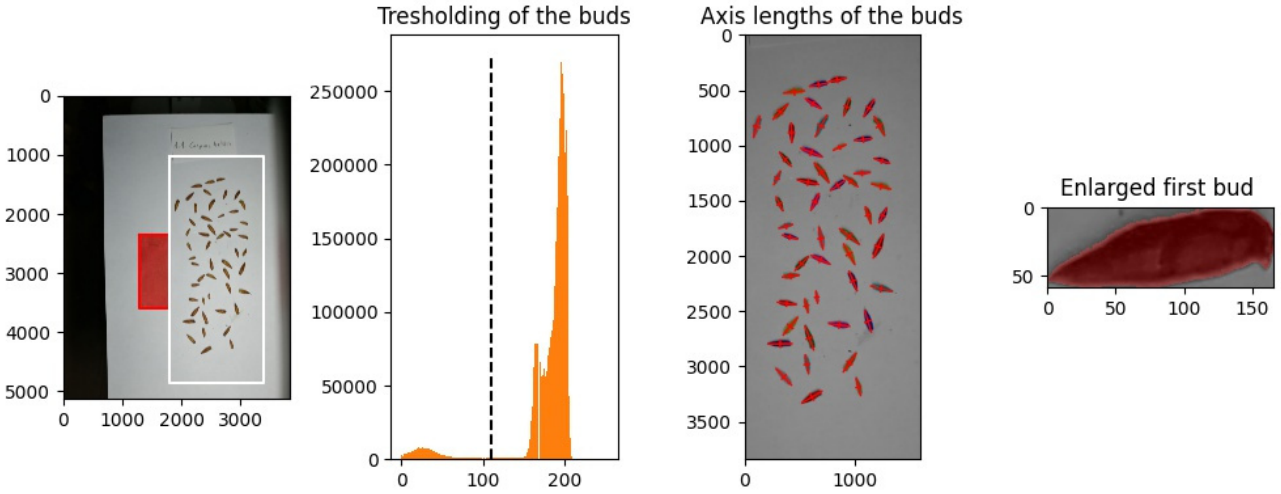

Figure 1: Control image of *Carpinus betulus* buds

pictures were taken from a similar distance, therefore the approximate size helped well to choose the scale. The position oft he scale was used to select the part oft he image with the buds. The buds were also segmented with Otsus Method. A problem for large buds was shadow around the buds and therefore larger segmented areas. Too small objects were removed. All holes were filled. For further processing all the segmented buds were labelled. The labelled images was saved to a binary file with NumPy [6]. Control images were made for all images to check that the segmentation worked correctly (see Fig. 1 for an example).

Object descriptors

All object descriptors were calculated with a combination of NumPy [6], scikit-image [5] and SciPy [7]. The shape of of an ellipse was estimated for all buds. The major and minor axis length of the ellipse was recorded. The ratio between the minor and the major axis length was calculated. Moreover, the area and the perimeter of the buds were saved. The roundness was computed with the area and the perimeter of a bud.

To check the accurateness of the major and minor axis length of the ellipse the width and height was also determined from the shape of the buds. The width at half of the height of the bud was calculated (blue in Fig. 3). The length was not perfectly in a 90 degree angle, because there is not a point exactly on the other side of the contour (it was chosen the closes point). Therefore the pythagoras theorem was used to calculated the length. The maximal x- and y-values were chosen as an approximation of the maximal width and height (green and red in Fig. 3). The ratio between the width and the height was also again calculated. The position of the maximal width along the major axis was determined.

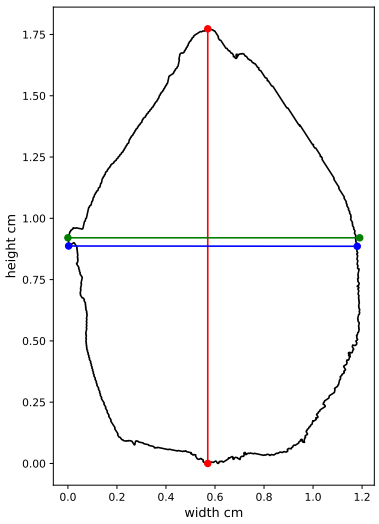

Figure 3: Width (green/blue) and height (red) of a bud.

The RGB colour model was converted to the HSV colour model. Mean, standard deviation and

skew of hue-, saturation- and value-distributions were recorded. Furthermore from the blue RGB channel the mean, standard deviation and skew was computed. An overview of all descriptors is shown in Table 2 in the appendix.

Elliptic fourier analysis

An elliptic fourier analysis [8] was applied to the contours of the buds to statistically compare the shape of the buds. Firstly, the contours of the buds were rotated (see Fig. 4), scaled to the actual size and rooted (minimum x and y-vales are 0). For most of the species it was possible to automatically rotate the buds. However for the buds of *Aesculus hippocastanum* the rotation had to be manually specified. Secondly, the coefficients up to the 50^{th} order of the elliptic fourier transformation were calculated with Blidh [9]. It was not possible to use the normalization step of the fourier coefficients, because the results showed artefacts (upside-down shapes in principal component analysis).

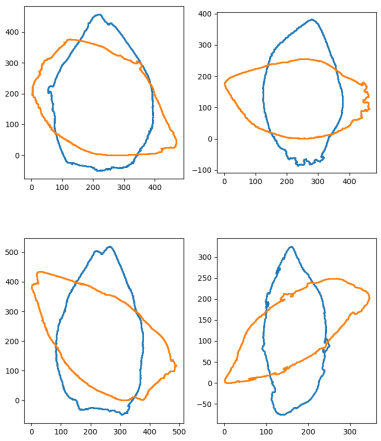

Figure 4: Rotation of the contours of *Aesculus hippocastanum* buds, (orange original, blue rotated)

The fourier coefficients were analysed with a principal component analysis (PCA). For each of the four squares in the PCA-space the mean of the coefficients were calculated. With these average coefficients the contour was reconstructed and plotted for visual inspection.

Prediction of species

Logistic Regression, Decision Tree and Random Forest were used to predict the species. All three methods are implemented in scikit-learn [10]. The maximal depth for the Decision Tree method was set to 10. For the logistic Regression maximal 200 iterations were used. 15 estimators were calculated with the Random Forest classifier. The dataset were splitted into train and test dataset (testsize: 30 %). The train dataset was used to fit all three models. The performance oft he models was evaluated with the test dataset. Moreover, the importance of the image descriptors was evaluated with the Decision Tree method.

3 Results

All the 1422 buds were described with the help of image descriptors and the coefficients of elliptic fourier analysis. Through the result section it is used the same colour palette for all images (see Fig. 5).

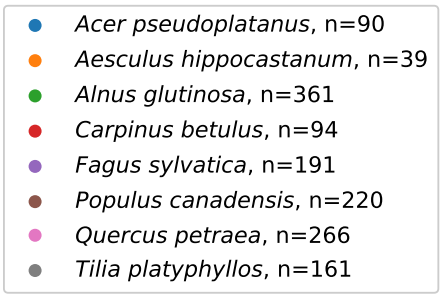

Figure 5: Colours used in all folowing images

The median major axis length differs significantly between species (Kruskal-Wallis $H = 1162$, $p < 0.001$). However, the median major axis length of the three largest species (*Aesculus hippocastanum*, *Fagus sylvaltica* and *Populus canadensis*) is not significantly different (pairwise Dune-test with correction after Benjamini/Hochberg $p > 0.05$). The shortest species are *Quercus petraea* and *Tilia platyphyllos*. There are differences between the median bud area of species (Kruskal-Wallis $H = 1016$, $p < 0.0.1$). The two biggest species are *Aesculus hippocastanum* and *Populus canadensis*, but there is no significant difference between these two species (pairwise Dune-test with correction after

Benjamini/Hochberg $p = 0.21$). The smallest three species are *Carpinus betulus*, *Quercus petraea* and *Tilia platyphyllos*. The relationship between the area and the major axis length is shown in Figure 7.

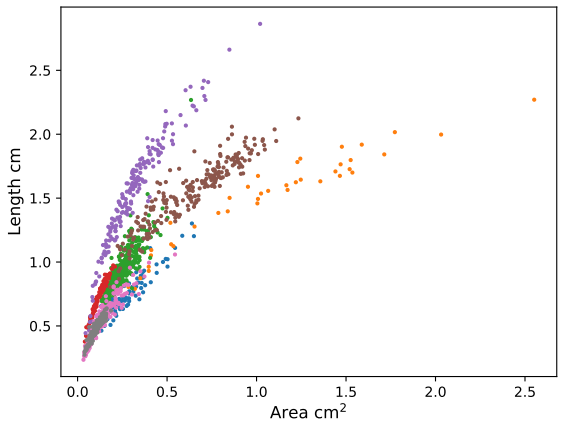

Figure 7: Area and major axis length of estimated ellipse of all buds

Principal component analysis

The result of a principal component analysis is shown in Fig. 6. The first three axes explain 67 % of the total variance. The first axes is correlated with the major axis length of the ellipse, the height of the contour, the area and the perimeter of the buds. The second axes represents a gradient in colour values. Points in upper direction tend to have a large standard deviation of the hue distribution. In contrast, points in the lower part of the PCA-space have a large mean saturation value and a high standard deviation of the saturation distribution. The third axes can be best described by the ratio of minor axis and major axis length of the ellipse and the ratio of height and width of the contour. Most of the species can be separated along these three axes. However, points of *Quercus petraea* and *Tilia platyphyllos* are largely overlapping.

Elliptic fourier analysis

A PCA of all fourier coefficients is shown in Fig. 8. The first two axes describe 92 % of the total variance, of which the first axis describes by far the largest portion. On the right side of the diagram tend to be larger buds. The second axis can be described as a gradient from thin to bigger buds. Most of the species are well separated along the first two axes, though points of *Quercus petraea* and *Tilia platyphyllos* are again clumped together.

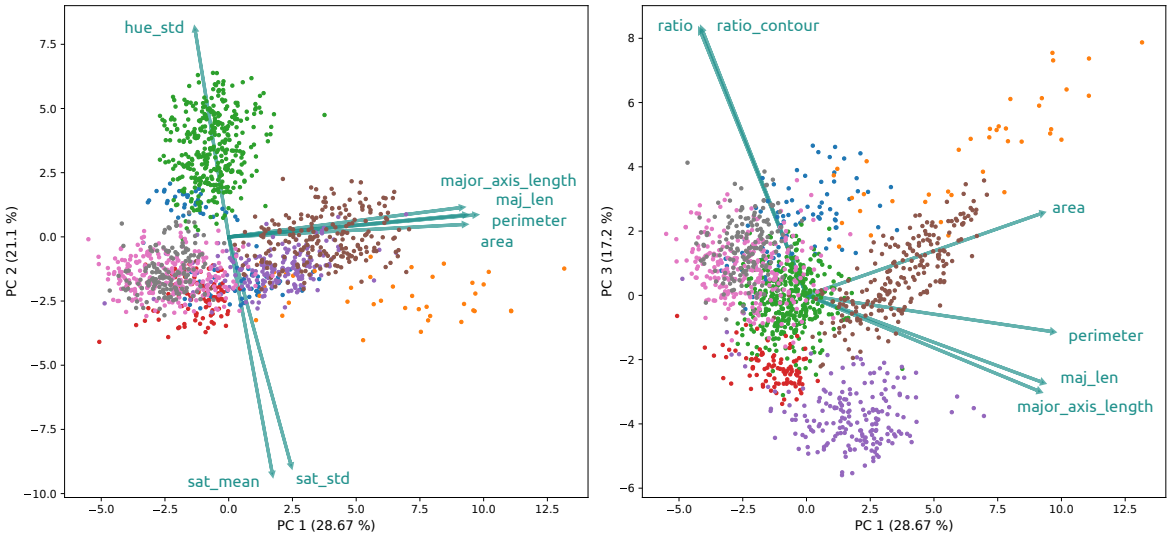

Figure 6: Principal component analysis of the complete dataset (but without elliptic fourier coefficients), factors (image descriptors, arrows) were chosen with a absolute loading larger than 0.8 and were multiplied by 10 for better visual inspection, the left figure shows the principal components one and two, the right figure the components one and three

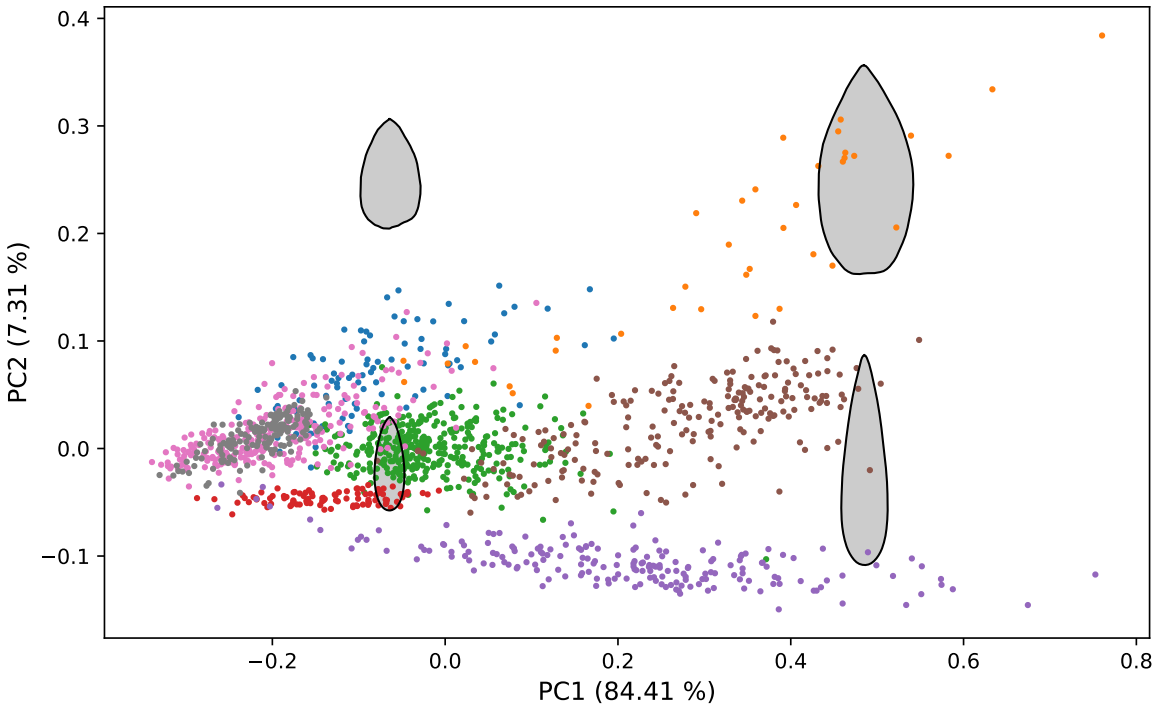

Figure 8: Principal component analysis of the elliptic fourier coefficients of the shape of buds

Prediction of species

The accurateness of all three methods were tested on train and test data (see Tab. 1). Random Forest showed the highest prediction score for the test score, followed by the Logistic regression. Decision tree had a high score in the train data, but the score for the test data was the lowest (92.8 %).

Table 1: Result of 200 times predicting the species with the three methods, numbers for the correct assignment are given in percent ± standard deviation

Method	Train data	Test data
Logistic regression	97.4 ± 0.3	95.9 ± 0.8
Decision Tree	99.3 ± 0.3	92.8 ± 1.3
Random Forest	99.9 ± 0.1	96.2 ± 0.9

The importance of the image descriptors for predicting the species can be extracted from the fitted Decision tree (see Fig 9). The three most important features chosen by the method were already elaborated with the PCA (see Fig. 6).

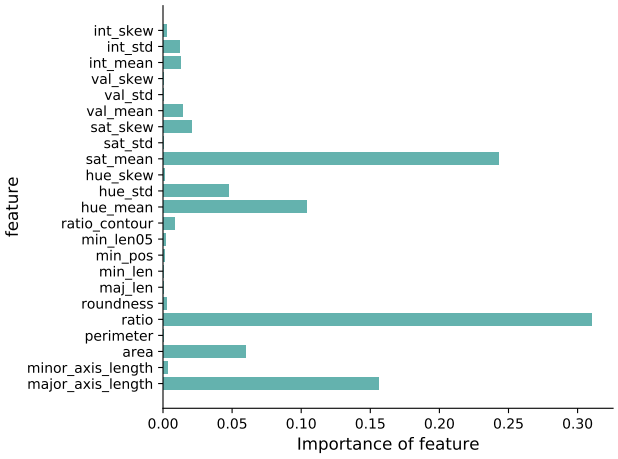

Figure 9: Importance of image descriptors for the decision tree

Comparing measurements of the estimated ellipse and of the shape

Major and minor axis length of the ellipse and the ratio of minor and major axes length showed a very high correlation with the height and the width of the contour and the ratio of width and height re-

spectively (for all three: Spearman $\rho = 0.99^{***}$). The decision tree method chose the values from the estimated ellipse for the prediction of species (see Fig. 9) and not the calculated values from the contour.

4 Discussion

The segmentation of the buds from the white paper and also the prediction of the species worked well. The next step would be to segments buds directly from a real environment and to use more species. Simple tresholding will not work in a real environment, because there is no clear colour change around the buds. The prediction score will also be lower if we include more species. There are already projects on working with automated plant species identification from images in the nature. An example is the FloraIncognita project [11]. The app asks the user to take multiple images from different organs of the plant, afterwards on a server these images are processed in a convolutional neural network. In the winter also images of buds are used to help to identify woody plants.

Branches with buds were only collected from one location. Therefore it is possible that the intraspecific variability of bud characteristics will be higher if the buds were sampled from many locations. Consequently the predictive power could also be lower with more locations.

Colour values were used to describe the buds and also to predict the correct species. It is not absolutely clear that the difference in colour value derives from a real difference or is just an artefact. All images of one species were taken directly one after the other and no colour key was used. It would be better to take alternate pictures of different species or use a colour key. However, it seems very likely that the difference is not just an artefact.

There appears to be no difference between values derived from an estimated ellipse and values derived directly from the contour. Furthermore, the values from the ellipse have a slighter higher power for separating different species.

5 Conclusion

The extraction of data from the images worked well. The most important features are overall size (major axis length, area and perimeter), colour values (mean saturation, standard deviation of saturation and standard deviation of hue) and the ratio between the minor and the major axis length. The accuracy of the prediction is for all methods higher than 90 %. However this presented approach will not work for images taken outside directly on the tree.

Data and code availability

All images and the Python code are available at https://github.com/FelixNoessler/Buds-morphometrics.

References

1. Cope, J. S., Corney, D., Clark, J. Y., Remagnino, P. & Wilkin, P. Plant species identification using digital morphometrics: A review. *Expert Systems with Applications* **39,** 7562–7573. `https://doi.org/10.1016/j.eswa.2012.01.073` (2012).
2. Gehan, M. A. *et al.* PlantCV v2: Image analysis software for high-throughput plant phenotyping. *PeerJ* **5,** e4088. `https://doi.org/10.7717/peerj.4088` (2017).
3. Brummitt, N., Bachman, S. & Moat, J. Applications of the IUCN Red List: towards a global barometer for plant diversity. *Endangered Species Research* **6,** 127–135. `https://doi.org/10.3354/esr00135` (2008).
4. Wäldchen, J., Rzanny, M., Seeland, M. & Mäder, P. Automated plant species identification—Trends and future directions. *PLOS Computational Biology* **14,** 1–19. `https://doi.org/10.1371/journal.pcbi.1005993` (2018).
5. Van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2,** e453. `https://doi.org/10.7717/peerj.453` (2014).

6. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. https://doi.org/10.1038/s41586-020-2649-2 (2020).

7. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272 (2020).

8. Kuhl, F. P. & Giardina, C. R. Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* **18,** 236–258. https://doi.org/10.1016/0146-664X(82)90034-X (1982).

9. Blidh, H. PyEFD. https://github.com/hbldh/pyefd (2020).

10. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html (2011).

11. Rzanny, M., Mäder, P., Deggelmann, A., Chen, M. & Wäldchen, J. Flowers, leaves or both? How to obtain suitable images for automated plant identification. *Plant Methods* **15.** https://doi.org/10.1186/s13007-019-0462-4 (2019).

6 Appendix

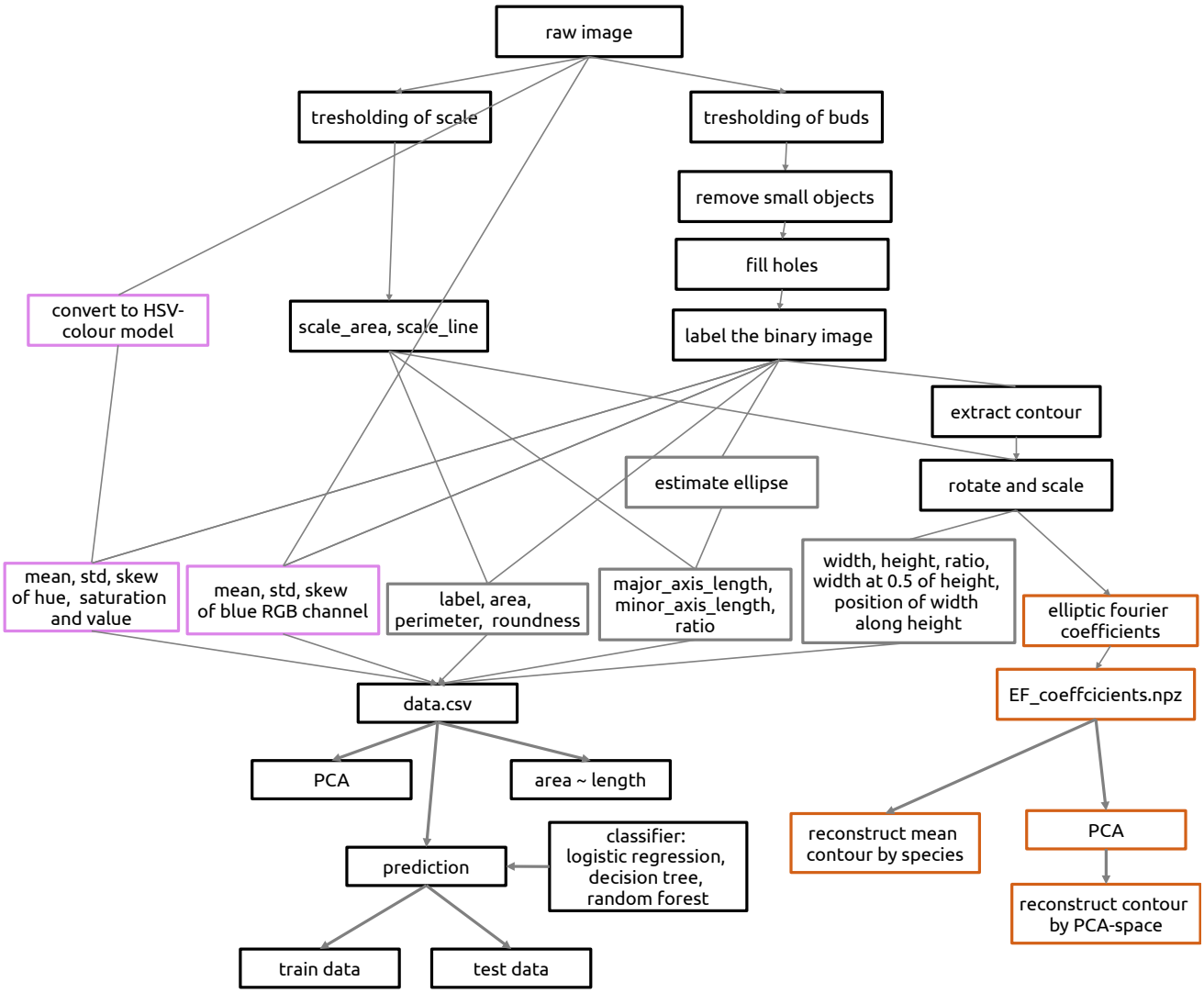

Figure 10: Processing and analysing of the raw images - workflow

Table 2: Overview of the used image descriptors

Descriptor name	Description
major_axis_length	height of estimated ellipse
minor_axis_length	width of estimated ellipse
area	area of bud
perimeter	perimeter of shape of bud
ratio	minor_axis_length / major_axis_length
roundness	roundness of the shape, 4 * Pi * area / perimeter**2
maj_len	height of shape
min_len	width of shape
min_pos	position of min_len along maj_len, starts from bottom of bud
min_len05	width of shape at 50 % of maj_len
ratio_contour	min_len / maj_len
hue_mean	
hue_std	information of the distribution of hue of the HSV-colour model
hue_skew	
sat_mean	
sat_std	information of the distribution of saturation of the HSV-colour model
sat_skew	
val_mean	
val_std	information of the distribution of value of the HSV-colour model
val_skew	
int_mean	mean intensity value of blue channel
int_std	standard deviation of blue channel
int_skew	skew of blue channel

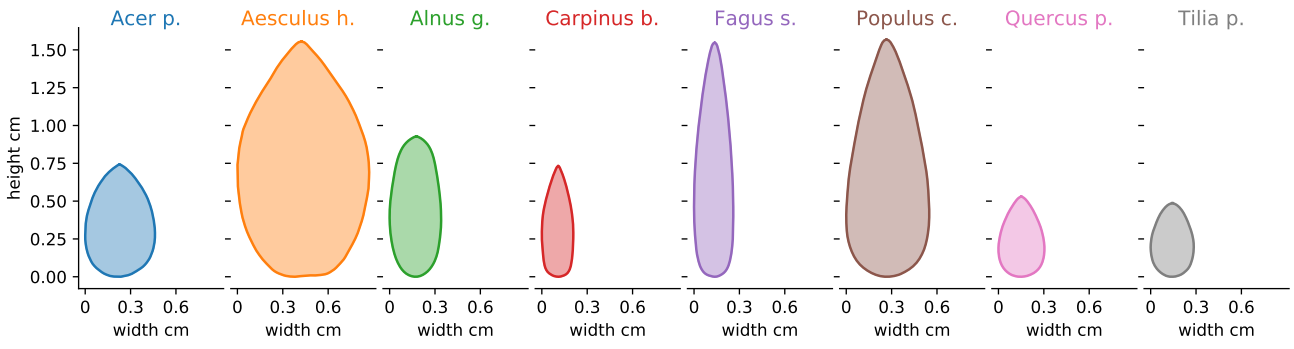

Figure 11: Mean contours of all species reconstructed from elliptic fourier coefficients