Pollen's profiling:automated classification pollen

- **Pollen's are the tiny,egg,shape,round,square,rectangle Male cell(organ) of flowering plants**
- **Average pollen particle size is less than the width of human hairs**
- There are types of pollen

Type 1:Anemphilous:small size 15-45 diameter,light,non adhesive &relatively smooth

□ Ex:-trees,grasses,weeds

Type 2:Entomophilous:large size 200 diamenter,heavier, somewhat spring

□ Ex:-honey suckle&rose

Overview:

The steps for our process are as follows:

□ Image acquisition and particle segmentation

□ Feature extraction and

□ Classification.

2) Our process begins with scanning glass slides of the various pollen species

1. with a digital microscope, then segmenting these images to gather samples ofindividual pollen particle.
2. These images are then further segmented to identify the pollen boundary
3. The area within this boundary is used for features extraction 18 shape features, texture features including the Fast Fourier Transform, Local Binary Patterns, Histogram of Oriented Gradients, and Haralick
4. features, as well as aperture features are used.
5. These features are then trained using supervised learning to build a model for the 5 pollen species sampled.
6. The model is then tested with ten-fold cross validation. The process is illustrated in figure

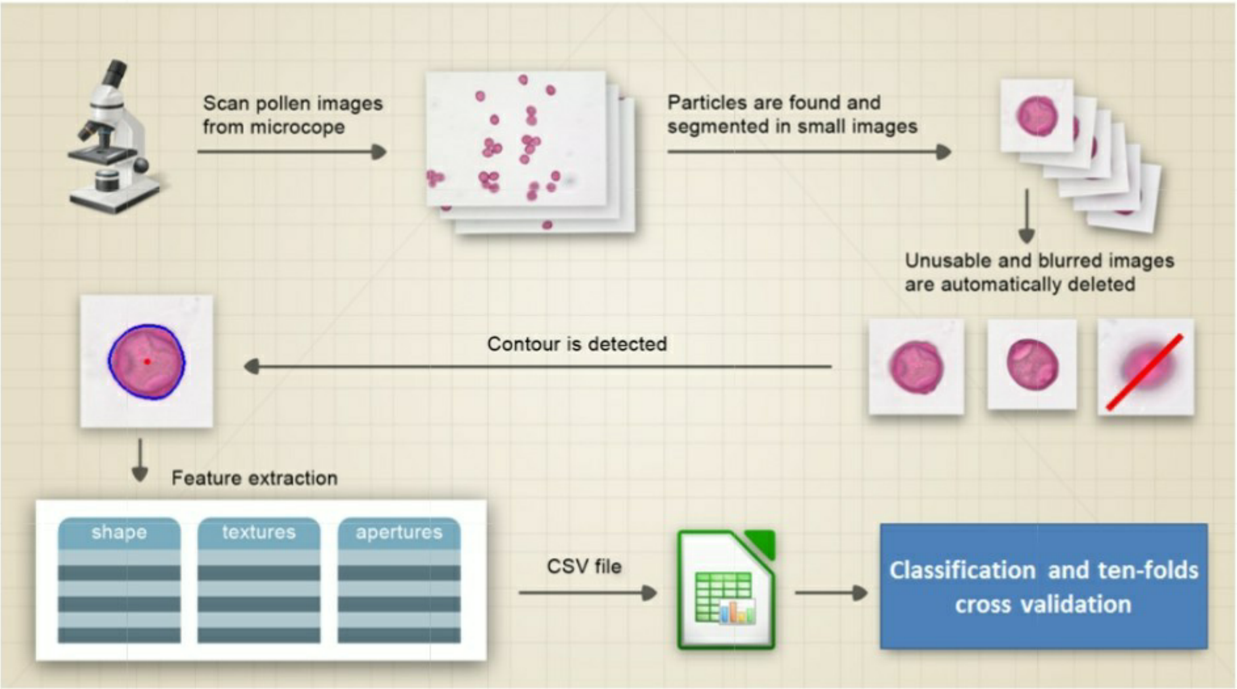

Scan pollen images from microcope
Particles are found and segmented in small images
Unusable and blurred images are automatically deleted
Contour is detected
Feature extraction
shape
textures
apertures
CSV file
Classification and ten-folds cross validation

Fig. 1. Pollen image acquisition and classification process

3) Image Acquisition and Particle Segmentation Five different species (Alder, Birch, Hazel, Mugwort, and Sweet Grass) have been stained and prepared on glass slides for use with a common digital bright￾field microscope. In order to build a robust model, all species had sample images derived from three distinct laboratory slides (using a total of 600 sample images obtained from 15 different slides).

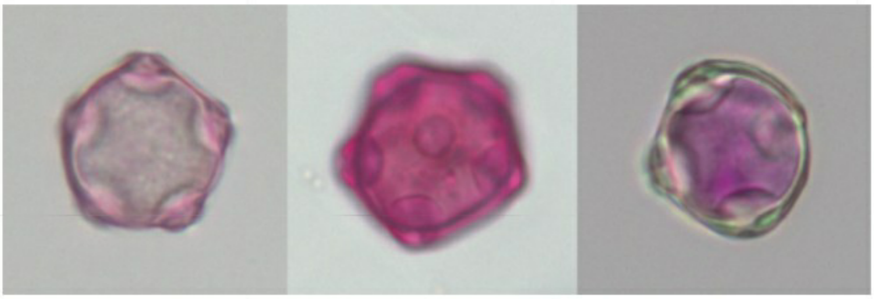


Fig. 2. Diverse image types, all example data used for training the Alder class of pollen. Access to the complete dataset can be found at http://dx.doi.org/10.5072/dans-zpr-rjm6.

1. For particle segmentation, each digital image is processed in order to locate and segment out a confining square surrounding a pollen particle. First, a me￾dian blur and Gauassian blur are applied to a negative of the image in order to remove smaller particles that are background noise (often dirt or imperfections on the background). Next, a threshold is applied to the image, using the OTSU algorithm to automatically detect the histogram peak. The returned image is an optimized binary image. A second set of filters is then applied using morpho￾logical operators (iterations of erosions and dilations) to fill in the particle area.Finally, the image is converted to have a white background in preparation for further processing steps.
2. A blob detection algorithm is now applied in order to extract a small image surrounding each particle. This algorithm is based on four attributes – Area, Circularity, Convexity and Inertia Ratio, with parameters for “minimum” and“maximum” values for each. By setting the parameters for the expected charac￾teristics of pollen grains, the smaller images are then found and extracted.
3. The last filter used on the resulting images of particles is depicted in Figure3. Because the pollen grains settle into the slide adhesive at different depths, some particles will be out of focus. These blurry images will provide insufficient data especially concerning texture features, therefore we remove them from our analysis. A blur detection algorithm was developed and applied to each image: a Laplacian filter set to a manually determined threshold value determines whichimages are too blurry and removed from further processing steps.
4. Lastly, the contour surrounding each pollen particle is identified, using OpenCV’s findContours() method.

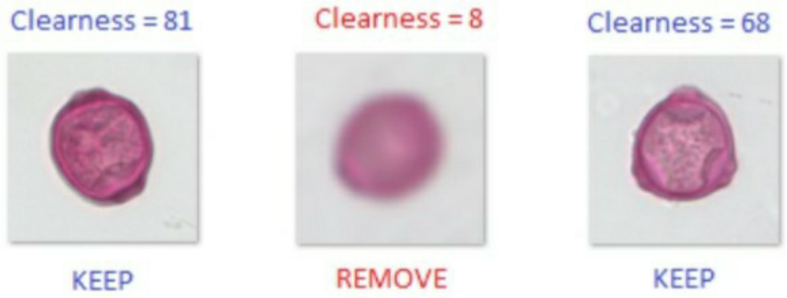


Fig. 3. Blur detection example

4.1) Feature Extraction:

1. Shape features :We have used 18 shape features already identified to be useful through previous iterations of our research [5]. The 18 selected were based on the research of developing an identification process for the Urticaceae family of pollen [11], aswell as research into developing universal shape descriptors [1].
2. Shape features used: Perimeter (P) Length of contour given by OpenCV's arcLength() function Area (A) Number of pixels contained inside the contour Roundness (R)$4\pi A/P$ 2 Compactness 1/R.
3. Roundness/Circularity Ratio (RC) Another measure of roundness, see [9] $P -\sqrt{P}$ $2-4\pi A/P +\sqrt{P}$ $2-4\pi A$
4. Mean Distance (S¯) Average of the distance between the center of gravity and the contour
5. Minimum Distance (Smin) Smallest distance between the center of gravity and the contour
6. Maximum Distance (Smax) Longest distance between the center of gravity and the contour
7. Ratio1 (R1) Ratio of maximum distance to minimal distance Smax/Smin
8. Ratio2 (R2) Ratio of maximum distance to mean distance Smax/S¯
9. Ratio3 (R3) Ratio of minimum distance to mean distance Smin/S¯
10. Diameter (D) Longest distance between any two points along the contour
11. Radius Dispersion (RD) Standard deviation of the distances between the center of gravity and the contour
12. Holes (H) Sum of differences between the Maximum Distance and the distance between center of gravity and the contour
13. Euclidean Norm (EN2) Second Euclidean Norm
14. RMS Mean RMS mean size
15. Mean Distance to Boundary Average distance between every point within the area and the contour
16. Complexity (F) Shape complexity measure based on the ratio of the area and the mean distance to boundary

4.2) Texture feature extraction:A variety of texture features were selected due to their performance in prior research [11,10,7,8]. The texture features extracted included: Gabor Filters (GF),the Fast Fourier Transform (FFT), the Local Binary Pattern (LBP), the Histogram of Oriented Gradients (HOG), and Haralick features.

1. Gabor Filters Gabor filters have been proven useful in image segmentation and texture analysis [12]. The Gabor Filter function consists of the application of 5 different size masks and 8 orientation masks (See Figure 4) in order to produce output images. For each of the 40

resulting images, we calculate the local energy over the entire image (the sum of the square of the gray-level pixel intensity), and the mean amplitude (the sum of the amplitudes divided by the total number of images). In addition to these 80 values, we also store the total local energy for each of the 8 directions as well as the direction where the local

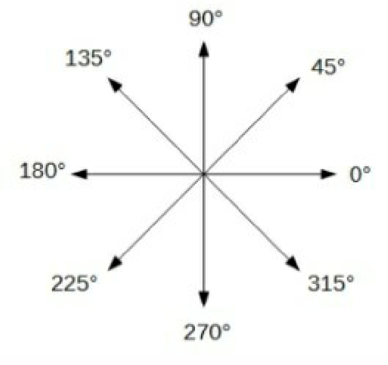
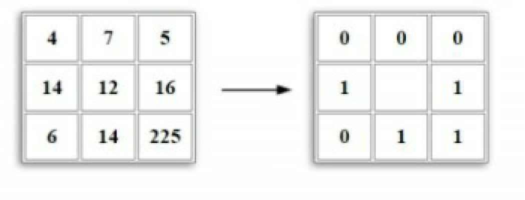


Fig. 4. The 8 directions of the mask for the Gabor Filters

1. Fourier Transform Fourier Transforms translate an image from the spatial domain into the frequency domain, and are useful because lower frequencies represent an area of an image with consistent intensity (relatively featureless areas) and higher frequencies represent areas of change [2]. Just as in spatial analysis,we cannot compare images directly, but first need to extract features. In the frequency domain, we likewise extract useful information through analysis of frequency peaks. Here, we apply a Fast Fourier Transform to the image, applya logarithmic transformation, and create a graph of the resulting frequency domain. After taking the highest 10 frequency peaks, we compute the differences between the peaks and store these values, as well as the mean of the differencesand the variance of the differences.Haralick Features.
2. Haralick features [3] are determined by computations overthe GLCM (Grey-Level Co-Occurence Matrix). Here, we use: the angular second moment, contrast, correlation, sum of squares: variance, inverse differencemoment, sum average, sum variance, sum entropy, entropy, difference variance,difference entropy, measure of correlation 1, and measure of correlation 2. These are 13 out of the 14 original features developed by Haralick: the 14th is typically left out of computations due to uncertainty in the metric’s stability.
3. Histogram oriented gradient (HOG) The Histogram of Oriented Gradientsis calculated by first determining gradient values over a 3 by 3 Sobel mask. Next,bins are created for the creation of cell histograms; here, 10 bins were used. The gradient angles are divided into these bins, and the gradient magnitudes of the pixel values are used to determine orientation. After normalization, the values are flattened into one feature vector.
4. Local Binary Pattern (LBP) To obtain local binary patterns, a 3 by 3 pixel window is moved over the image, and the value of the central pixel is compared to the value of its neighbors. In the case that the neighbor is of lower value,it is assigned a zero, and in the case of a higher value, a one. This string of eight numbers (”00011101” for instance) is the determined local pattern. The frequency of the occurrence of each pattern is used as the texture description.

4.



4	7	5
14	12	16
6	14	225

→

0	0	0
1		1
0	1	1

Fig. 5. Example output for Local Binary Patterns

Aperture Detection: The number and type of apertures present on the pollen surface is a typical feature used by palynologists in order to determine the pollen species. Therefore, it seems useful to also build an automatic aperture detection function in order to identify and count apertures as an addition feature set.Preliminary work identifying apertures [4] has shown potential for this analysis.First, a moving window segments the pollen image into smaller areas. Each smaller image is manually labeled as an aperture or not an aperture. Texture features are extracted from these smaller images, including those through a Fast Fourier Transform (FFT), Gabor Filters (GF), Local Binary Pattern (LBP),Histogram of Oriented Gradients (HOG), and Haralick features. A supervised learning process (through the use of support vector machines) then creates a model for each of the four species expected to include apertures on the surface.Once an unlabeled pollen image is given to be classified, the system again uses a moving window to break up the image into subsections. These smaller sections are then loaded into the generated model, and four values are returned for each detected aperture, corresponding to the probability that the aperture is of typeAlder, Birch, Hazel, and Mugwort.

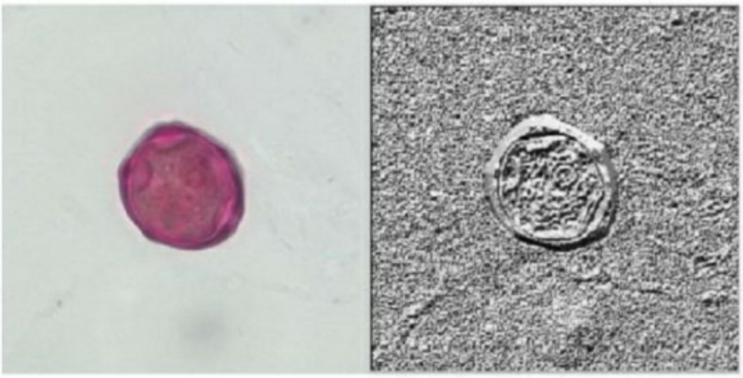


Fig. 6. Local Binary Pattern function applied to pollen image

5) Classification:Once the shape, texture, and aperture features have been calculated, they are added together into a csv file. A data set of 5 species with 40 sample pollen images from 3 separate sample slides led to a total of 600 samples, each with 252 extracted features. A supervised learning process used this data for model creation, which was then tested using ten-fold cross validation. Both support vector machines and a random forest classifier showed promising (and very sim□ilar results); for the results reported here a random forest classifier

was used due to faster processing on the larger data sets. The n-estimators parameter for this method was set to a typical size of 100 (increasing this number did lead to slightly improved results yet also dramatically increased processing times).

6)Results:Using a random forest classifier on a total of 600 samples (120 each for each species) and 252 features, a model was generated with an accuracy of 87% ±2%.

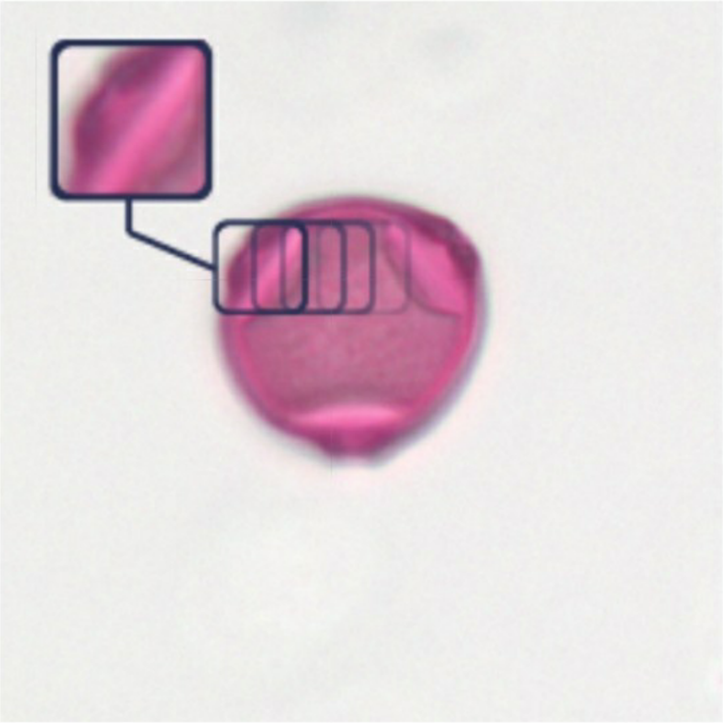


Fig. 7. Window moving all over the pollen

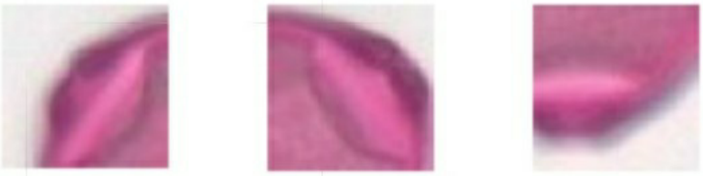
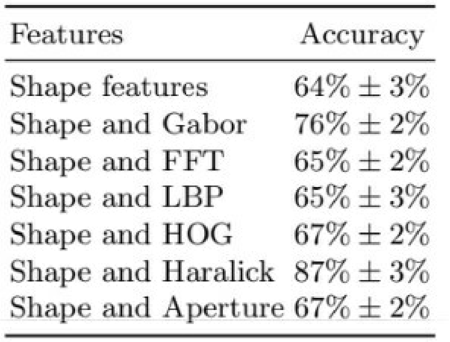


Fig. 8. Apertures detected by the program on a Birch pollen

Considering that the samples were intentionally selected for variability in their

1. appearance and background (See Figure 2), this is an indication of a robust,reliable model that shows promise for expansion in the future to also include datasets collected from an outdoor environment.
2. The dataset was further modified into different versions in order to test theresults using only subsets of the features available.
3. The above table shows the accuracies of the trained models. Using only the 18 shape features, an accuracy of 64%±3% was achieved, and adding texture information either through Gabor Filters or Haralick features substantially improved the result

3.



Features	Accuracy
Shape features	64% ± 3%
Shape and Gabor	76% ± 2%
Shape and FFT	65% ± 2%
Shape and LBP	65% ± 3%
Shape and HOG	67% ± 2%
Shape and Haralick	87% ± 3%
Shape and Aperture	67% ± 2%

7)Conclusion:Through this research, we have tested an expanded sample set of 5 species of pollen particles and used shape, texture and aperture features for use in classification. Use of all features led to an accuracy of 87% ± 2%. Through testing of individual texture features in combination with shape features, it was found that using only the shape and Haralick features resulted in an accuracy of 87% ±3%.Gabor Filters also proved to be a useful feature as seen through the improved accuracy compared to using just the shape features alone. Surprisingly, the other texture features as well as the aperture features did not result in significant accuracy gains. One next step of research would be to investigate under which exact conditions certain texture features prove useful. In the case of the aperture features, one known limitation is that the aperture types were trained on a more limited dataset. Because the aperture detection process technique developed did have positive results in determining correct aperture positions, it would be interesting to retrain the aperture type on a wider dataset and see if this results in a more useful set of extracted features. Furthermore, extending the dataset not only beyond 600 images but especially to include more than three microscope slides per species would test against possible overfitting to particular slide conditions. Future research would also include application of this process to data collected outside of a laboratory environment, as well as expansion to include more pollen species.