Analysis of Relevant Features for Pollen Classification

Gildardo Lozano-Vega[1,2], Yannick Benezeth[2], Franck Marzani[2], and Frank Boochs[1]

[1] i3mainz, Fachhochschule Mainz, Lucy-Hillebrand-Strasse 2, 55128 Mainz, Germany
gildardo.lozano@fh-mainz.de
[2] Le2i, Université de Bourgogne, B.P. 47870, 21078 Dijon Cedex, France

Abstract. The correct classification of airborne pollen is relevant for medical treatment of allergies, and the regular manual process is costly and time consuming. Aiming at automatic processing, we propose a set of relevant image-based features for the recognition of top allergenic pollen taxa. The foundation of our proposal is the testing and evaluation of features that can properly describe pollen in terms of shape, texture, size and apertures. In this regard, a new flexible aperture detector is incorporated to the tests. The selected set is demonstrated to overcome the intra-class variance and inter-class similarity in a SVM classification scheme with a performance comparable to the state of the art procedures.

Keywords: pattern recognition, feature extraction, feature evaluation, apertures, palynology.

1 Introduction

The correct estimation of airborne pollen concentration is important for the prevention and treatment of allergies. Traditional methods require manual and specialized labor, which is expensive, time consuming and susceptible to inconsistency. Commonly, collected airborne particles are analyzed manually under a brightfield microscope in order to count the frequency of different pollen taxa. With the introduction of computer vision techniques, pollen counting aspires to become automatic, faster and more accurate, which would enable a more frequent and broader analysis of samples.

Common strategies for image-based pollen classification follow a typical image classification process: image digitization, preprocessing, segmentation, pollen description, and pattern recognition. In all the cases, pollen description is based on different types of metrics and representations of the pollen image, known in the literature as *features*. Previous strategies are differentiated mainly by the type of employed features. Chen *et al.* found a relevant combination of seven general shape features, four aperture-colpus features, and a statistical gray-level feature to recognize *Birch, Grass* and *Mugwort* pollen with 97.2% of accuracy with a Linear Normal Classifier (LNC) and forward feature selection [1]. Unfortunately,

L. Iliadis et al. (Eds.): AIAI 2014, IFIP AICT 436, pp. 395–404, 2014.

the aperture and colpus detectors are type-specific and are not easily usable for other taxa.

Additionally to general shape and aperture features, Boucher *et al.* used color features for the classification of 30 pollen taxa [2]. With the same deficiency as Chen *et al.* for the aperture detection, they achieved an accuracy of 77%. The average of 11.6 images/taxon in the dataset looks small for capturing all the variations. Interestingly, they used a knowledge-based classification instead of typical statistical methods. A good performance just employing five general shape features was achieved by Rodríguez Damián *et al.* in the classification of three types of *Urticaceae* pollen with 85.6% of accuracy with a minimum distance classifier [3]. Employing only Haralick measures, Li *et al.* classified four pollen taxa with 100% of accuracy using the Fisher's linear discriminant method [4]. However, the tested taxa does not appear to have strong inter-class similarity which leaves the performance unknown on typical allergenic pollen. The aforementioned results suggest that the synergic contribution of features from different descriptive foundations could be the key for a robust and accurate classification.

Atypically, Ranzato *et al.* used local invariant features without segmentation with accuracy of 78.2% on eight pollen taxa, and employed a Bayesian classifier with a Gaussian mixture probability density function [5]. Similarly, Ronnenberger *et al.* developed 3D features from the Haar integration framework, achieving 98.5% of precision with a Support Vector Machine (SVM) and the Radial Basis Function (RBF) as the transformation kernel, for 33 taxa after the rejection of uncertain results [6]. Although very impressive, unfortunately the method requires 3D volumetric scans of the particle, which looks impractical for real-time applications.

Our proposed solution takes advantage of the strength of diverse feature groups in order to provide robustness and accuracy. The rest of the paper is organized as follows: in section 2, the proposed strategy for description and classification is explained in detail. In section 3, experiments and results are presented. Finally, in section 4, conclusions of the proposed method and the following steps are stated.

2 Strategy for Pollen Description and Classification

2.1 Outline of the Proposed Solution

We propose the application of selected features for the classification of the five most important allergenic pollen taxa in Germany: *Alder, Birch, Hazel, Mugwort* and *Grass*. Starting with microscopic 2D images, the solution involves the segmentation of the pollen from the background and the description of the particle by different types of features. Our contribution is a robust set of descriptive features, which range from general to more specific, and it is able to overcome the most important problems in pollen recognition: the inter-class similarity and the intra-class variation. Robustness is provided by employing features related to different pollen characteristics such as shape, size, texture and apertures. Moreover, an unified flexible aperture detector is employed for first time in pollen classification instead of multiple fixed algorithms. Finally, a relevant set of features is

systematically selected and applied to a support vector machine process for classification of the taxa. This classifier is chosen due to its well known performance in different tasks, particularly when high dimensional data are involved [7].

2.2 Segmentation of the Pollen Particles

Color images of magenta-dyed pollen are manually digitized from microscope slides and labelled by palynologists. The isolation of the particle from the background by a bounding contour is a necessary step to compute the set of features. For this purpose, the application of the active contour method is often found in previous works, for example in [3][6]. Although effective in filling gaps, this method requires a modelling stage for a good performance with the risk of loosing flexibility. In contrast, Chen *et al.* successfully used a simple two-step automatic thresholding and hole-filling for segmentation of three different pollen taxa [1].

Following this simple idea, our automatic method employs Otsu's thresholding because it is nonparametric, unsupervised and fast [8]. This method searches the threshold that minimizes the intra-class variance of the background and the particle pixels on gray level images. It was found experimentally that the method can determine the correct threshold even under variation of light intensity due to the scanning process, as long as a certain intensity difference between the background and the particle exists (*cf.* Fig. 1). After the binarization, it is easy to trace the contour using Suzuki's method [9]. Due to its high contrast, debris could affect the correct estimation of the contour. If not stuck to the pollen, debris is detected and rejected by its size and irregular shape.

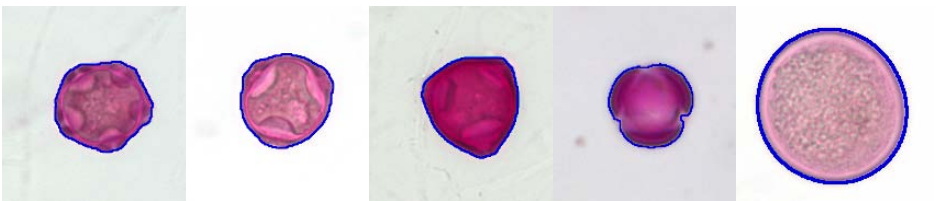


Fig. 1. Example of the five pollen taxa of interest and their segmentation using Otsu's automatic method. The detected contour in blue is enhanced for visualization.

2.3 Description of the Pollen Particles

According to their descriptive characteristics, features can be grouped as general shape features, elliptic Fourier descriptors, texture features, and apertures.

General Shape Features. As mentioned in section 1, shape features are proven to be suitable to characterize pollen for classification purposes [1][3]. The computation of these features is based on a few main variables. Additionally

to *perimeter* and *area*, statistics of the distance between the centroid and the contour are used: *standard deviation, mean (Dmean), minimum (Dmin)* and *maximum (Dmax)*. The first part of Table 1 shows the ten classical features that are employed in the present work, and whose detailed equations are given in [10]. These features focus, besides on estimating the size, on quantizing statistically the complexity of the shape. However, there is no feature that relates to the outline of the pollen, which in general can be described as spherical, oblate or prolate. Because this property is pointed out by palynologists to be discriminative [11], we propose a point-by-point representation of the pollen outline by fitting an ellipse to the contour using the Joon Ahn *et al.* optimal fitting algorithm [12]. From this, useful rotation invariant features are added to the classical shape features: *ratio between major and minor axes length*, and *rms value, mean* and *standard deviation of the error of the fitting*.

Particularly, *perimeter*, *area* and *2eN* are features that are proportional to the shape size. If combined with the rest of the general shape features, these three features could introduce bias related to the pollen size differences in the evaluation of the ability of general shape features to discriminate shape. For this reason, they are not considered at the shape evaluation stage. However, size is an important discriminant of pollen taxa and these three features are included in a global evaluation together with other type of features.

Table 1. List of general shape features grouped in classical and ellipse-fitting-based

Classical shape features

Feature	Short name	Features	Short name
Perimeter	P	Radius dispersion	rdis
Area	A	Ratio Dmax/Dmin	ratio1
Roundness	R	Ratio Dmax/Dmean	ratio2
O'Higgins Undulation	U	Ratio Dmin/Dmean	ratio3
Complexity f	cf	2n Euclidean norm	2eN

Ellipse fitting features

Feature	Short name	Features	Short name
Axes Ratio	EF_ratio	Mean error	EF_mean
RMS error	EF_rms	Std Dev of the error	EF_std

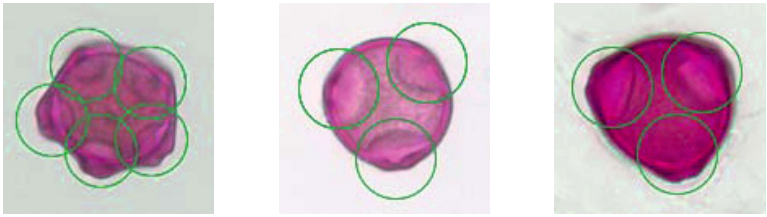
Elliptic Fourier Descriptors (EFD). These descriptors are employed to represent shapes given their contour as point pairs (x, y). The method is based on treating each dimension, x and y, as separate functions, and computing the Fourier coefficients. The coefficients are then transformed into translation, rotation and scale invariant descriptors [13]. The EFD's contain all the information to reconstruct the original shape in an inverse operation. Moreover, the EFD's are ordered in relation to the frequency content of shape contour form low to

high. Iwata *et al.* were able to describe subtle variations of similar root shapes belonging to the same taxa with principal component analysis on EFD's [14]. Due to the strong similarity among the classes, the root representation would have not been effective if general shape features had been employed instead. For our tests, 400 EFD's are computed, from which the first two descriptors are not considered since they are meaningless due to a normalization step, resulting in 398 EFD's.

Texture Features. This group of features is focused on recognizing the texture patterns that are formed by the contribution of the outer ornamentation and the inner mass of the pollen. Motivated by results of previous works, we are interested in testing the Haralick measures on the top five german allergenic taxa. Our approach employs eleven Haralick measures from the gray co-occurrence matrix(GLCM): *angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance* and *difference entropy* [15]. 24 different pixel offsets are employed, resulting in a total of 264 features ranging from two to eleven pixels of separation distance (0.5 to 2.9 microns) in different orientations. Because of the compactness of the texture patterns, longer distances are not needed. Only the area inside the contour is considered in order to avoid influence from the particle border and the background.

Aperture Feature. Palynologists state that the type and amount of apertures is very discriminating of the taxa [11]. Apertures are morphological distinctive regions of the pollen wall, usually visible in typical microscopic observations. Due to their inter-class variety, previous approaches have struggled to detect this kind structures efficiently. Moreover, changes in the point of observation increase the variability of their appearance. Figure 2 shows examples of different aperture appearances. Previous recognition strategies have focused on just particular pollen types, which would require the development of new algorithms for other types. For example, Boucher *et al.* detected apertures of *Grass* by means of segmentation techniques, shape and color features [2]. Chen *et al.* detected the aperture and colpus of *Birch* and *Mugwort* based on the intensity profile of the image polar transformation, the Hough transform and a template matching technique [1].

The present strategy employs the proposed method by Lozano Vega *et al.* in which different aperture types can be detected with the same algorithm [16]. This flexibility is an advantage for the recognition of multiple taxa with multiple appearances. Moreover, the method was tested on *Birch* and *Alder* with recall above 80%. The representation of the apertures is through the histogram of primitive visual words contained in the image and densely extracted. These visual words capture most of the variance of the different patterns present in the pollen and are gathered in a codebook. The visual words are created from clusters based on the similarity of local feature vectors computed on small image patches. Local Binary Patterns (LBP) [17] is chosen due to its simple but efficient computation and successful results on the description and classification of complex image



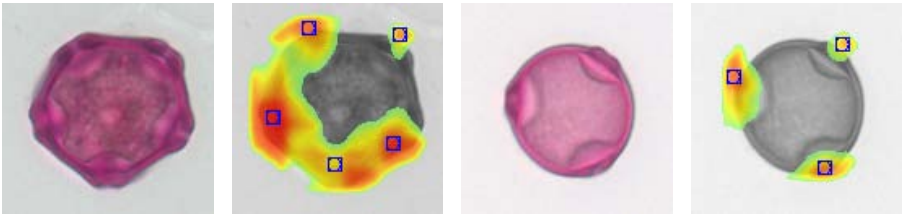
Alder Birch Hazel

Fig. 2. Different aperture appearances due to the change of the viewing angle and taxon variety. Aperture is highlighted by a green circle.

patterns such as faces [18]. The uniform (u2) normalized LBP version with a (8,1) neighborhood is employed as local feature to reduce the length of patterns and take advantage of the rotation invariance [19].

A SVM classifier is trained for the recognition of the aperture region histograms. SVM is a popular method that maps the feature vectors into a higher dimensional space in order to separate the positive and negative classes with a hyperplane with a maximum margin. The radial basis function is used as mapping kernel, and the penalty C and the kernel parameter γ are tuned with a grid search in a three-fold cross validation.

In order to evaluate an unknown pollen, multiple regions covering the particle are classified as *aperture* or *not an aperture* using the aforementioned method. A pixel-wise confidence map is created by the averaged votes of overlapping regions, which can be interpreted as the likelihood of the occurrence of apertures. Finally, apertures are found by estimating the local maxima of the map and the total amount is used as the aperture feature. Our implementation creates 20 visual words and square regions of 30 pixels by side, being able to detect apertures of *Alder, Birch,* and *Hazel* among five pollen types. An example of the confidence map and the detection of apertures is shown in Fig. 3.



(a) (b) (c) (d)

Fig. 3. Original images of *Alder* (a) and *Birch* (c) pollen and their respective confidence maps (b) and (d). Warmer colors (red-orange) on the maps indicate a high likelihood of detecting an aperture. Blue squares indicate the estimated location of the apertures after searching local maxima.

3 Experimental Results

We employed a statistical classification strategy for the evaluation of the feature groups on the discrimination of the five most important pollen taxa in Germany. The source microscope slides were prepared in a laboratory, stained with magenta, and scanned with a Keyence BZ-9000 microscope in brightfield mode with a magnification of 40x and a resolution of 0.26 microns/pixel. The dataset of individually cropped pollen images consisted of 100 labeled images/taxon, except for *Hazel*, for which 48 images were employed due to a defective staining. Automatic segmentation was performed according to the method described in section 2.2 and the complete set of features was computed according to section 2.3: 14 general shape features, 398 EFD's, 264 texture features, and the aperture feature.

Initially, each group of features was evaluated individually in order to evaluate the performance of the subsets and to be able to discover the most discriminant features. The aperture feature was not tested individually because results from a single-value feature are not relevant. Moreover, palynologists support firmly the discriminant importance of this characteristc [11]. Although SVM classification was selected especially for the high-dimensional EFD's and texture features, it was employed also for the rest of the experiments in order to keep comparability. The performance measure was the accuracy in a three-fold cross validation with stratified sampling and a training/testing ratio of 0.67/0.33. This method allows to verify the robustness to unseen data. We also tried five and ten folds in order to validate changes of the training/testing ratio (0.8/0.2 and 0.9/0.1 respectively). Since results were comparable, they are not shown here.

The optimal set of features can be found by examining all combinations, strategy called brute force feature selection. However, this solution is not computational efficient, particularly for high dimensional vectors, as in the case of EFD's and texture groups. For those cases, Sequential Forward Selection (SFS) is the method employed for discovering the most relevant features and for reducing the feature space dimensionality. SFS is a wrapper approach that searches the best set of features based on the classification performance. It begins with the empty subset of selected features $X = \{\emptyset\}$ and starts an iterative search of features $\{x_1, x_2, \ldots, x_n\}$ to grow X. In each iteration, all the combinations of X with a new feature is evaluated. The feature x_i that delivers the best performance is integrated permanently to X. The iteration is stopped until no better performance is obtained, and then the current X is regarded as the best subset. This method avoids evaluating all the combinations of features and it is especially fast when the optimal set is small. To keep consistency, the performance metric was a three-fold cross validation using a SVM classifier with RBF kernel.

For the individual experiments with the **general shape features**; *perimeter, area* and *2eN*; related directly to the pollen size; were excluded intentionally to avoid influence of the size differences in the evaluation of the shape recognition. Because of the reduced size of this group, brute force feature selection was feasible and employed in this feature group, with the advantage of finding the best subset of features.

The achieved accuracy was **83.49% ±2.19%** with *cf, rdis, ratio2, ratio3, EF_rms* and *EF_mean*. As expected, ellipse-fitting features came up as an important extension to classical features. Interestingly, neither *R* nor *EF_ratio*, which measure shape circularity, were selected. We consider that the intra-class variation of the shape of the pollen when projected into a 2D plane is strong enough to cause also a strong variation of the shape circularity, decreasing the discrimination power of this type of measures.

The accuracy with the **EFD's group** was **76.18% ±2.10%** with 34 descriptors using SFS, discarding 90% of the features. It is relevant to notice that 50% of the selected descriptors correspond to the first 90 positions, containing the lower frequencies of the contour.

The **texture group** reached an accuracy of **83.26% ±1.39%** with 42 features after applying SFS, discarding 84% of the features. The most discriminant Haralick measures were *angular second moment, contrast* and *sum average* with 21 features. The most frequent offsets belong to equivalent distances of two and ten pixels, which correspond together to 16 selected features.

Finally, we evaluated the performance of the classification using together **all the feature groups**. Only selected features from the individual tests were input in addition to *area, perimeter, 2eN* and the aperture feature for a total of 86 features. The classification accuracy increased up to **94.42% ±1.27%**, more than 10% above better than the individual groups. The confusion matrix is shown in Table 2 and the comparison with individual results in Fig. 4. Most errors are due to the mix up of *Alder and Birch*, which belong to the same family *betulaceae*, due to their mutual similarity in size, ornamentation and aperture number.

In summary, results show the importance of the contribution of different features to the classification accuracy. The combination of features of different nature allows to capture the diversity of visual information present in the pollen, which is linked to palynological properties. While the general shape features measure the shape complexity, EFD's can describe finely the pollen outline. Texture features capture appropriately pollen ornamentation patterns. Size and

Table 2. Confusion matrix of the classification of the five pollen taxa using all the feature groups together

		True				
		Alder	Birch	Hazel	Mugwort	Grass
Predicted	Alder	**91**	8	0	1	1
	Birch	8	**88**	1	1	0
	Hazel	1	2	**47**	0	0
	Mugwort	0	2	0	**98**	0
	Grass	0	0	0	0	**99**

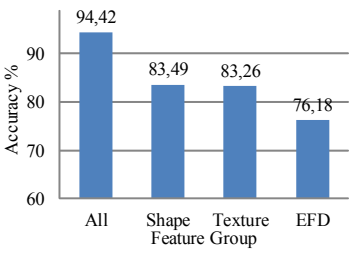


Fig. 4. Accuracy of the different tested groups. Using all groups together performed much better that the individual groups.

apertures are concepts which are statistically and semantically related to each taxon. The combination of the groups contributes to enhance the classification performance and the robustness against unknown data and errors on the contour.

4 Conclusions and Future Work

The performance of our proposal is among the best pollen recognition processes, comparable to that of Chen *et al.* with the addition of two more taxa and using a single method for aperture detection. More than 400 pollen samples from five pollen taxa and validation with different training and testing sets support the reliability of the results. Feature selection enables to identify relevant features and to reduce the feature space dimensionality. Assembling different feature groups facilitates to sum up strengths and to minify weaknesses, resulting in a more robust classification.

Focusing on the rich information (usually disregarded) in the palynology literature, we suspect that building a knowledge-structured classification system based on palynological properties could improve robustness while maintaining high accuracy. Those properties could be detected by taking advantage of the relevant features that are proposed in the present work.

Acknowledgements. The authors are grateful for their financial support to the German Bundesministerium für Wirtschaft und Technologie under the program Zentrales Innovationsprogramm Mittelstand ID KF2848901FR1, to the Conseil Regional de Bourgogne in France and to the Fond Europeen de Developpement Regional (FEDER). The authors are also grateful to Celeste Chudyk, Yann Ryann and Morad Larhriq for scanning and preparing the pollen datasets and to the Max Plank Institute for Chemistry for permitting the use of their facilities.

References

1. Chen, C., Hendriks, E.A., Duin, R.P., Reiber, J., Hiemstra, P., De Weger, L., Stoel, B.: Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort. Aerobiologia 22, 275–284 (2006)
2. Boucher, A., Hidalgo, P.J., Thonnat, M., Belmonte, J., Galan, C., Bonton, P., Tomczak, R.: Development of a semi-automatic system for pollen recognition. Aerobiologia 18(3), 195–201 (2002)
3. Rodríguez-Damián, M., Cernadas, E., Formella, A., González, A.: Automatic identification and classification of pollen of the urticaceae family. In: Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIVS 2003), pp. 38–45 (2003)
4. Li, P., Treloar, W.J., Flenley, J.R., Empson, L.: Towards automation of palynology 2: the use of texture measures and neural network analysis for automated identification of optical images of pollen grains. Journal of Quaternary Science 19(8), 755–762 (2004)
5. Ranzato, M., Taylor, P.E., House, J.M., Flagan, R.C., LeCun, Y., Perona, P.: Automatic recognition of biological particles in microscopic images. Pattern Recognition Letters 28(1), 31–39 (2007)

6. Ronneberger, O., Wang, Q., Burkhardt, H.: 3D invariants with high robustness to local deformations for automated pollen recognition. In: Proceedings of the 29th DAGM Conference on Pattern Recognition, pp. 425–435 (2007)
7. Byun, H.-R., Lee, S.-W.: Applications of support vector machines for pattern recognition: A survey. In: Lee, S.-W., Verri, A. (eds.) SVM 2002. LNCS, vol. 2388, pp. 213–236. Springer, Heidelberg (2002)
8. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man and Cybernetics 9, 62–66 (1979)
9. Suzuki, S., Abe, K.: Topological Structural Analysis of Digitized Binary Images by Border Following. Computer Vision, Graphics, and Image processing 30(1), 32–46 (1985)
10. Da Costa, L., Cesar Jr., R.: Shape analysis and classification: theory and practice. CRC Press Inc. (2001)
11. Erdtman, G.: An Introduction To Pollen Analysis. Chronica Botanica Company, U.S.A. (1943)
12. Joon Ahn, S., Rauh, W., Warnecke, H.: Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola. Pattern Recognition 34(12), 2283–2303 (2001)
13. Nixon, M., Aguado, A.S.: Feature Extraction and Image Processing, 2nd edn. Academic Press (2008)
14. Iwata, H., Niikura, S., Matsuura, S., Takano, Y., Ukai, Y.: Evaluation of variation of root shape of Japanese radish (Raphanus sativus L.) based on image analysis using elliptic Fourier descriptors. J. Euphytica, 143–149 (1998)
15. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics 3(6), 610–621 (1973)
16. Lozano-Vega, G., Benezeth, Y., Marzani, F., Boochs, F.: Classification of Pollen Apertures Using Bag of Words. In: Petrosino, A. (ed.) ICIAP 2013, Part I. LNCS, vol. 8156, pp. 712–721. Springer, Heidelberg (2013)
17. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29 (1996)
18. Huang, D., Shan, C., Ardabilian, M., Wang, Y., Chen, L.: Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 41(6), 765–781 (2011)
19. Ojala, T., Pietikäinen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 971–987 (2002)