Pollen Segmentation and Feature Evaluation for Automatic Classification in Bright-field Microscopy

Rafael Redondo[f,g], Gloria Bueno[f], François Chung[h], Rodrigo Nava[g], J. Víctor Marcos[g], Gabriel Cristóbal[g], Tomás Rodríguez[h], Amelia Gonzalez-Porto[i], Cristina Pardo[j], Oscar Déniz[f], B. Escalante-Ramírez[g]

[a] *VISILAB research group at the University of Castilla La Mancha, av. Camilo José Cela s/n, Ciudad Real 13071 (Spain)*
[b] *Instituto de Óptica, Spanish National Research Council (CSIC), Serrano 121, Madrid 28006 (Spain)*
[c] *Inspiralia Co., Estrada 10, B, Madrid 28034 (Spain)*
[d] *Centro Agrícola de Marchamalo, Guadalajara (Spain)*
[e] *Facultad de Farmacia, Universidad Complutense, Madrid (Spain)*

Email address: email:rafa@optica.csic.es, phone:+34 915616800 (Rafael Redondo)

Pollen Segmentation and Feature Evaluation for Automatic Classification in Bright-field Microscopy

Rafael Redondo[f,g], Gloria Bueno[f], François Chung[h], Rodrigo Nava[g], J. Víctor Marcos[g], Gabriel Cristóbal[g], Tomás Rodríguez[h], Amelia Gonzalez-Porto[i], Cristina Pardo[j], Oscar Déniz[f], B. Escalante-Ramírez[g]

[f] *VISILAB research group at the University of Castilla La Mancha, av. Camilo José Cela s/n, Ciudad Real 13071 (Spain)*
[g] *Instituto de Óptica, Spanish National Research Council (CSIC), Serrano 121, Madrid 28006 (Spain)*
[h] *Inspiralia Co., Estrada 10, B, Madrid 28034 (Spain)*
[i] *Centro Agrícola de Marchamalo, Guadalajara (Spain)*
[j] *Facultad de Farmacia, Universidad Complutense, Madrid (Spain)*

Abstract

Besides the well-established healthy properties of pollen, Palynology and apiculture are of extreme importance to avoid hard and fast unbalances in our ecosystems. To support such disciplines computer vision comes to aleviate tedious recognition tasks. In this paper we present an applied study of the state of the art in pattern recognition techniques to describe, analyze, and classify pollen grains in an extensive dataset specifically collected (15 types, 120 samples/type). We also propose a novel contour-inner segmentation of grains, improving 50% of accuracy. In addition to published morphological, statistical, and textural descriptors, we introduce a new descriptor to measure the grain's contour profile and a logGabor implementation not tested before for this purpose. We found a significant improvement for certain combinations of descriptors, providing an overall accuracy above 99%. Finally,

Email address: email:rafa@optica.csic.es, phone:+34 915616800 (Rafael Redondo)

some palynological features that are still difficult to be integrated in computer systems are discussed.

Keywords: Apiculture, pollen, automatic classification, bright-field microscopy, feature extraction, Fisher discriminant analysis, image processing, morphology descriptors, statistical descriptors, texture descriptors.

---

1. Introduction

A grain of pollen contains the male vegetative and generative cells required for fertilization of plants to ensure the development of seeds and consequently the life of plants. The study of pollen, palynology, is therefore of great interest in so diverse disciplines such as archeology, paleontology, forensics, health (allergies) or agriculture (bee products, and crop forecast). Specifically, bee pollen is collected by worker honey bees which is used as food for the entire colony. For humans it is one of the richest and purest natural foods, with an incredible nutritional and medicinal value [1, 2] and one of the most interesting facts about bee pollen is that it cannot be synthesized in a laboratory. The main nectar source and main pollen source differ widely with latitude, region, season, and type of vegetation, where in scarce nectar periods bees can harvest far away up to 3 km, i.e., in an area of 300-2800 hectares [3]. This reflects their large pollination capacity and the maintenance of plant diversity which directly influences important human activities like agricultural and forestry production. Furthermore, bees are the most common pollinators with strong influence on ecological relationships, ecosystem conservation, and stability, genetic variation in the plant community, biodiversity, specialization, and evolution [4].

The pollen grains manifest a great variety of shapes, sizes, and ornamentation and their description is genetically bound to their botanical family. Externally, pollen grains are protected by a resistant wall called sporoderm, conformed by an internal layer named intine and an outer layer named exine, where the latter exhibits in its surface distinct morphological structures according to the pollen type. Generally, most of them are spheroidal in equatorial view, varying between oblate spheroidal and prolate spheroidal in the range of 8-100 μm.

In the human activities previously mentioned a correct pollen identification is vital in terms of production, bio-preservation, or simply knowledge achievement. The recognition can be accomplished through different techniques which in general are time consuming and require highly trained palynologists who must analyze manually thousands of individual pollen grains: Fourier transformed infra-red from attenuated total reflectance (FTIR-ATR) spectroscopy represents a useful technique for identifying chemical structures [5]; and Polymerase Chain Reaction (PCR) is a recent method for pollen authenticity based on molecular analysis. PCR technique stands out for its specificity for botanical identification. Nevertheless both techniques are expensive in terms of equipment and reagents, and requires several processing days. Finally, the most common and affordable technique is bright-field microscopy. This technique is time consuming too and therefore many efforts have been put on automated classification systems. However it remains a challenge to provide accurate pollen classifications in real scenarios. For a recent study that provides a comparison of the microscopy techniques, see [6].

The first attempt to automate pollen recognition was conducted in 1968 by Flenley by [7], who identified two difficulties attached to bright-field microscopy: images partially focused and multiple grain orientations (views). Both are related with the reduction of 3D objects into 2D captures. The depth of field of optical systems allows visualization of specimens partially in focus. Here, the use of multifocus stacks and recent multifocus fusion techniques [8] could eventually provide more details about pollen's surface, but the way of collecting information is still an open issue. On the other hand, morphology, surface ornamentation, and pori layout are strong indicators of the pollen type, but such information strongly vary with the point of view.

Besides these inherent difficulties in capturing 3D features into 2D, two main obstacles hamper the current progress in this field: a) the extraction of knowledge from expert palynologists and b) the limited access to open pollen databases with a large number of reference pollen per taxa. A previous work in the area of aero-palynology (ASTHMA EU project) used multifocus stacks and reported recognition rates around 97% for 5 pollen types [9]. Other studies demonstrate accuracy ratios between 90-97% [10, 11, 12, 13, 14]. However, such ratios must be considered with care, they are not reliably comparable because their training database usually differ largely in terms of pollen genre and/or number of training samples, which is directly related to obstacle b).

Most of these approaches, if not all, perform morphological and certain statistical description of gray-levels like mean, median, variance, entropy,...etc. Some modern approaches incorporate more sophisticated descriptors through spatial correlations like the Haralick's co-occurrence ma-

trices [15]. For instance, Zhang proposed Gabor transforms and invariant moments [16], Rodriguez-Damian et al. [17] evaluated Fourier descriptors and Run-Length Statistics, Chen et al. [11] incorporated a description of the number of pores and recently Ronneberger et al. proposed 3D invariant moments [18]. An interesting and profuse thesis can be consulted in [19].

In some applications, e.g. images from ambient air, a previous image cleaning from dirt, fungal spores and other non-pollen particles [20] is required. This is also a time consuming process where a robust automatic segmentation is a challenging problem.

In this paper we present a complete applied study of segmentation, description, and classification of bee pollen, reviewing the state of art and proposing some novel techniques. For that, within the EU-funded project APIFRESH, we recollected an important data base of 15 pollen types with 120 samples per type described in Sec. 2. Under the hypothesis that contour and inner of grains typically manifest disparate statistical distributions, we proposed in Sec. 3 a novel segmentation to apply descriptors separately across these two regions. In Sec. 4 an important exercise of knowledge transfer is done from palynology to computer vision together with a complete list of descriptors. Sec. 5 and Sec. 6 describe classification strategies and classifiers. Finally experimental results are presented in Sec. 7 and Sec. 8 concludes the paper addressing unresolved challenging problems.

2. Materials & Preparation: collecting database

Bees collect pollen aggregated in balls and normally of the same pollen type, which guarantees a certain corresponding hue. Therefore, balls were

separated in the laboratory and individualized by color tonality and then labeled with a color code according to the Universal Code Guide PANTONE 747XR. Although pollen can share color, each color corresponds to a pollinic type and a pollinic type can be matched to a larger group of plants (a family), to a middle group of plants (some genera from the same family), to a reduced group of plants (a genus) or more rarely to one species. Balls collected from the same place of origin were classified in colors and for each color we selected 25 pollen balls. Balls were dissolved with glycerogelatin drops and prepared in slices sealed with a coverslip. Through the microscope each botanical group has characteristic features that differentiate it from others like morphology, surface structures or pori layout. For a summarized featured list of the pollen types studied here consult the appendix in Sec. 9.

Although multiple studies have already evaluated a wide range of pollen descriptors, most of them have been done with a reduced dataset and/or a reduced number of pollen types. Without a doubt one of the major efforts in this kind of studies has to do with the compilation, preparation and labeling of datasets. Thus, some of those studies deal with 300-500 total samples and/or 3-5 pollen types [10, 11, 9, 21, 22, 12]. The study from Chica [13] is one the most complete in this respect with 5 pollen types and 1063 total pollen grains. But one impressive case is Ronneberger et al. [14] with 180 000 airborne particles and 22 700 pollen grains. In this case study, we have done an important effort to collect a considerable dataset in order to test computer vision algorithms focused on a real automated pollen classifier.

The 15 pollen types studied were collected mostly from Spain (Guadalajara, Toledo, La Rioja, Madrid and Cantabria). Other types like *Aster* and

Castanea came from Italy (Grosseto, Cosenza and Asti), *Helianthus* from Bulgaria and *Teucrium* from Turkey. They are enumerated in Tab. 1 and some examples are depicted in Fig 1. Neither type is endemic, which means that their presence is common along the whole Mediterranean territory and some of them are present all over the planet. Besides, none of them comes from plants commonly cultivated.

Pollen types	*Aster, Brassica, Campanulaceae, Carduus, Castanea, Cistus, Cytisus, Echium, Ericaceae, Helianthus, Olea, Prunus, Quercus, Salix, Teucrium*
Magnification	×40
Original captures	2560×1920 RGB pixels
Cropped grains	from 200 to 600 gray pixels of width and height (variable aspect ratio)
Type grains	120 images/type
Total samples	1 800 images (grains)

Table 1: Pollen database description.

A NIKON E200 microscope (fluoride objective) and a camera NIKON DS-Fi1 were employed to capture the images. Auto-white background balance was previously carried out for every slice capture with the NIS-Elements Nikon software. In Fig. 2 a capture and a example of manual cropping is shown. See Tab. 1 for specifications of the captured and cropped images.

Every sample was manually cropped to ensure an effective surrounding region with no nearby samples or debris, so cropped image samples are not necessarily square. To ensure an optimum focus, every acquisition included a 31-stack image where only the best focused slice was included in the dataset. This best focus was again manually selected. These stages and their automation are beyond the scope of this study.

The colors observed through the microscope, which are not necessarily consistent with the color of the ball which the pollen come from, did not

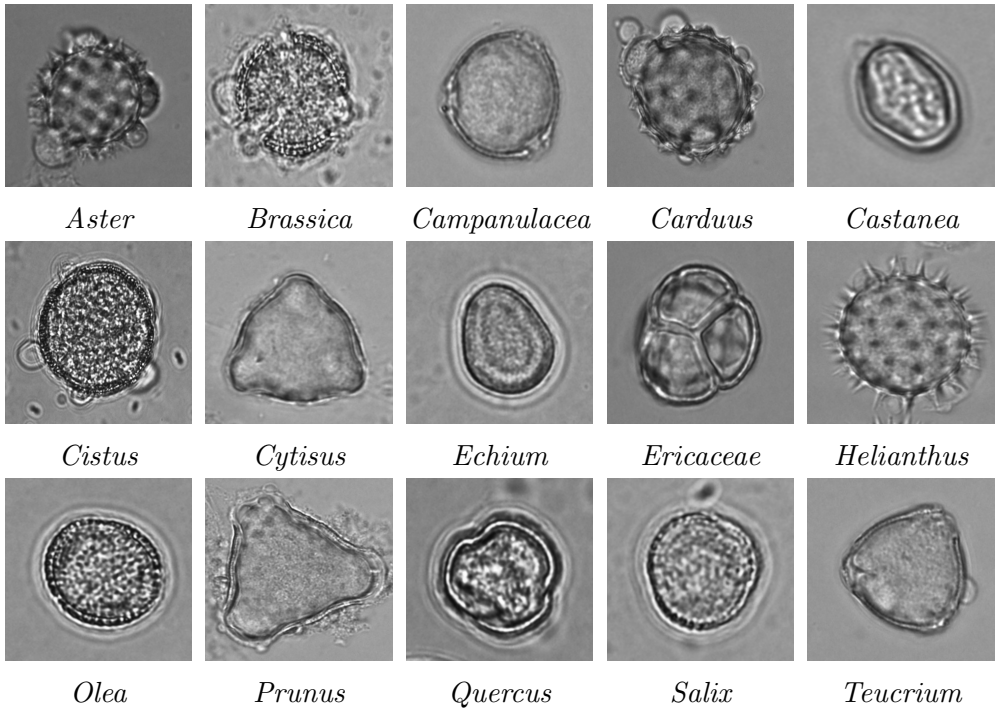


Figure 1: Examples of pollen database. Note these grains have been conveniently scaled here for aesthetic reasons.

Figure 2: Microscope example of *Brassica* genre at magnification ×40 of size 2560×1920 pixels. Framed sample of size 258×243 pixels.

presented evidences of discriminant significance. In addition the white balance performed by CCDs can significantly vary from one to other. Therefore the images were finally converted to grayscale. By doing this we mitigated at the same time the presence of yellowish lipids and other possible colored debris present in the slices which could impair the pollen segmentation and also feature extraction stages.

3. Binary masks: contour-inner segmentation

Binary masks are effective regions where descriptors must be computed, while other regions out of the mask are ignored. The list of published works about automatic pollen segmentation is short and there is still a need for a definitive method without posterior supervision. In this work we present a semi-automatic method which still requires manual outline corrections. Nevertheless the main novelty that we propose is a dual segmentation for inner areas and grain's contours. In these two areas there are visible structural differences in terms of recognition and therefore our hypothesis is that some descriptors should manifest different distributions too. This is not expected to affect morphological descriptors but descriptors relating with statistical properties of the pixels. One way to deal with such a dichotomy could be to implement classifiers able to handle multi-modal distributions. However, in this case we considered this dual segmentation more robust and controlled approach. Doubling the number of effective regions will double dimensionality of the classification domain, but at the same time the classification process gets simplified in terms of class separability according to the *a priori* location knowledge of pollen's textures. The following items enumerate the

sequence of processes we used to give shape to such masks (see also in Fig. 3 a schematic block diagram).

1. *Thresholding binarization*: automatic thresholding segmentation with the maximum histogram value. Similar techniques like Otsu's [22, 10, 13].

2. *Maximum area*: preservation of the biggest area, smaller regions rejected.

3. *Hole filling*: inside holes are filled if present.

4. *Opening*: erosion and dilation with a 15×15 kernel, holes revised again.

5. *Inner-contour segmentation*: erosion with a kernel proportional to the equivalent binary mask diameter[1]. Empirically adjusted to 15%.

These well-known algorithms are present in almost every image processing software and therefore no further details are given in this respect. Due to the presence of debris and some peculiar pollen types more complex to be segmented, the binary masks were afterward manually checked for finishing correction, see some examples in Fig. 4. Note that every pollen type present different exine but the pollen type is not know a priori, therefore such a 15% is necessarily a compromised value that could better fit in some pollen types than others, see Fig. 5. This value corresponds to exine sizes from 1-10μm. At this moment we will leave possible improvements and alternative segmentation strategies for an interesting further research.

[1] diameter of a circle with the same area

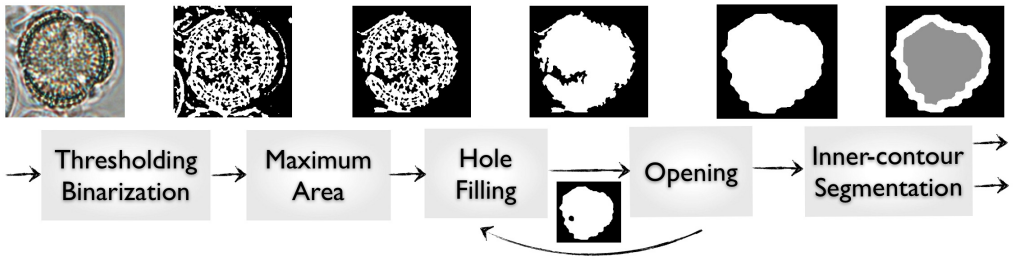
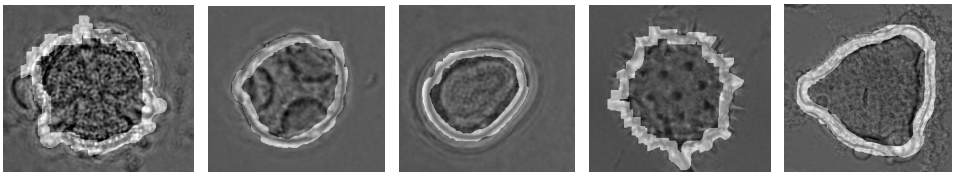


Figure 3: Segmentation example of a pollen grain borrowed from the previous Fig. 2 (258×243 pixels): image thresholding binarization with maximum histogram value; maximum area preservation and removal of smaller areas; inner holes filling-in; opening process with a 15×15 mask (it may need feedback); and inner and contour segmentation (15% grain diameter).

(a) (b) (c)

Figure 4: Examples of contour and interior pollen segmentation adjusted to 15% of their equivalent diameter. **(a)** *Echium* (237×285), **(b)** *Brassica* from Fig. 3 (258×243) and **(c)** *Helianthus* (354×330). Note these grains have been conveniently scaled here for aesthetic reasons.



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

Figure 5: Examples of binary mask segmentation including a contour extraction with a kernel 15% of the grain diameter. **(a)** *Brassica* (368×320), **(b)** *Campanulaceae* (272×256), **(c)** *Echium* (288×272), **(d)** *Helianthus* (384×352) and **(e)** *Prunus* (512×512). Note again these grains have been conveniently scaled here for aesthetic reasons.

4. Pollen Feature Descriptors

As descriptions of pollen species can be found through numerous publications, a special effort must be done in translating such a knowledge in terms of computer vision. To a significant extent the experts' capability to distinguish among similar pollen types comes from a knowledge not necessarily extracted from such bright-field images, but also coming from text descriptions, 3D spatial vision and simply reasoning that humans often do almost effortlessly. Considering that some features are simply imperceptible without such human capability recognizing, which is still not present in the state of the art in computer vision, some other features do describe pollen grains as for accomplishing a helpful automatic classification. In this way, according to our palynologist team we described such pollen features in useful terms for pattern classification in Tab. 2. Although not all the features will be faced here, like for instance apertures which are for the moment under development and some previous works can be found in [11, 23], other features like general morphology or texture will be one of the basis of this study.

Along the subsequent subsections we organized descriptors in groups according to their formulation, which at the same time will help to conduct the later experiments. A brief notion is introduced here and we let the reader to deepen along of plenty of well documented references. See descriptors categorized in Tab. 3. Note that with the exception of the morphological type, the rest of descriptors are computed uniquely in those pixels tagged by the segmented binary masks (contour and inner separately).

Table 2: Feature description of the 15 pollen types in terms of computer vision.

				POLAR VIEW		MERIDIAN VIEW		
TYPE	INSIDE	CONTOUR	SIZE	PERIMETER	APERTURES	PERIMETER	APERTURES	USUAL VIEW
Quercus	slightly granulated	thick	medium	circular or oval	3 depressions	circular or oval	enlarged towards poles	none
Echium	smooth	thin	small	oval (pear)	negligible	oval (pear)	negligible	meridian
Cistus	reticular	medium	medium	circular	negligible	circular	negligible	none
Olea	partially reticular	thick+reticular	medium	circular	3 subtle openings	circula r	negligible	none
Salix	wide reticular	thick+reticular	medium/small	circular	3 profound openings	oval	enlarged towards poles	none
Brassica	reticular	medium	medium	tri-lobular	3 depressions	circular	3 lineal fissures	none
Retama	reticular	medium	medium/small	triangular	3 depressions	circular	3 lineal fissures	none
Carduus	dense prickles	thick+dense prickles	medium/large	circular	3 protuberances	circular	protuberances	none
Aster	prickles	thick+prickles	medium/small	circular	3 tiny depressions	circular	3 lineal fissures	none
Helianthus	prickles	thick+large prickles	medium	circular	3 tiny depressions	circular	3 lineal fissures	none
Teucrium	smooth	thick poles	medium	oval	3 tampons	triangular	two tampons	none
Cytisus	transparent	thin+non reticular	medium/small	triangular	3 thin openings	rhomboidal	rhombus opposite sides	none
Ericaceae	transparent	thick+non reticular	medium/large	tetrahedral	Tri-lobed	tetrahedral	qua/bi-lobulated	none
Castanea	transparent	non reticular	small	circular	3 depressions	Oval	several depressions	none
Campanulacea	trickles	thin+tiny prickles	small	circular	3 tiny depressions	Oval	3 small depressions	polar
Prunus	mainly striate	double+striate	medium/large	triangular	3 openings	Oval	central + large	polar

$small \sim 10 - 25\mu m \quad \cdot \quad medium \sim 25 - 50\mu m \quad \cdot \quad large \sim 50 - 100\mu m$

CATEGORY	DESCRIPTOR	ANNOTATION	TOTAL DESCRIPTORS
Morphological	Area, Perimeter, Shape, Eccentricity, Fullness, Contour Profile	6 features (3 eccentricities)	6
Statistical	1^{st} Order	13 features	13
	2^{nd} Order Haralick	$distance = 1, 3, 5$ $direction = 0°, 45°, 90°, 135°$	19x3x4 = 241
Transformed space	LBP	mean, variance, asymmetry and kurtosis	4
Moments	Hu	7 moments	7
Space-frequency	Fourier	4 scales	241x4 = 964
	Wavelets	4 scales (3 orientations)	241x4 = 964
	Gabor	4 scales (6 orientations)	241x4 = 964

Table 3: List of descriptors grouped in testing categories.

4.1. Morphological Descriptors

The binary masks described in the previous section establish an appropriate framework to compute morphological features related to pollen's contour and area.

4.1.1. Area

This descriptor can be calculated as the sum of pixels in the binary mask of size MxN given $B \in (0, 1)$:

$$Area = \sum_{n=1}^{N} \sum_{m=1}^{M} B(m, n) \tag{1}$$

4.1.2. Perimeter

This descriptor is the number of pixels that belong to the object and which have at least one neighbor belonging to the background.

$$Perimeter = \sum_{n=1}^{N} \sum_{m=1}^{M} P(n,m)$$

$$P(m,n) = \begin{cases} 1 \;\; if \;\; \exists \;\; B(m \pm 1, n \pm 1) = 1 \\ \\ 0 \;\; otherwise \end{cases} \tag{2}$$

4.1.3. Shape

This descriptor measures the elongation of an object. For a circle its value is equal to 1. It is calculated in the following way:

$$Shape = \frac{4 \cdot \pi \cdot Area}{Perimeter^2} \tag{3}$$

4.1.4. Eccentricity

these descriptors also reflect elongation but in relation with the object's center of mass, also called centroid and defined as:

$$(m_c, n_c) = \left(\frac{1}{Area} \sum_{(m,n) \in Area} m \cdot B(m,n), \frac{1}{Area} \sum_{(m,n) \in Area} n \cdot B(m,n) \right) \tag{4}$$

The first $Eccentricity_1$ is defined as a quotient of the maximum and minimum distance between the centroid and object's border, also called outer and inner circumference radius.

$$Eccentricity_1 = \frac{Outerradius}{Innerradius} \tag{5}$$

Similarly $Eccentricity_2$, is calculated as quotient of the semi-axes of the

Mean	$\mu = \sum_{i=0}^{H-1} i \cdot h(i)$
Mode	$i = argmax(h(i))$
Variance	$\sigma = \sum_{n=0}^{H-1} (i-\mu)^2 \cdot h(i)$
1^{st} *Quartile*	$\mu_{q1} = \sum_{i=3\lceil H/4 \rceil}^{H} i \cdot h(i)$
2^{nd} *Quartile*	$\mu_{q2} = \sum_{i=2\lceil H/4 \rceil}^{3\lceil H/4 \rceil} i \cdot h(i)$
3^{rd} *Quartile*	$\mu_{q3} = \sum_{i=\lceil H/4 \rceil}^{2\lceil H/4 \rceil} i \cdot h(i)$
Interquartile Range	$\mu_{q3} - \mu_{q1}$
Minimum	$min(h(i))$
Maximum	$max(h(i))$
Range	$max(h(i)) - min(h(i))$
Entropy	$\sum_{i=0}^{H-1} h(i) \cdot log(h(i))$
Asymmetry	$\frac{1}{\sigma^3} \sum_{n=0}^{H-1} (i-\mu)^3 \cdot h(i)$
Kurtosis	$\frac{1}{\sigma^4} \sum_{n=0}^{H-1} (i-\mu)^4 \cdot h(i)$

Histogram $h(i)$, bins number H, floor operator $\lceil\ \rceil$.

Table 4: First order statistical descriptors.

best fitting ellipse for the object and $Eccentricity_3$ is a ratio of the inertia moments of the two semi-axes of the best fitting ellipse (see forwards for a description of moments).

4.1.5. Fullness

is the ratio of the object area to bounding rectangle area.

4.2. Statistical Descriptors

4.2.1. 1st-order Statistical: histogram

These descriptors, listed in Tab. 4, measure typical statistics in image histogram $h(i)$. These group of descriptors are sensible to global variation of gray pixel levels, but they ignore their local correlation.

4.2.2. 2st-order Statistical (Haralick): co-ocurrence matrix

These descriptors, listed in Tab. 5, measure statistics in co-ocurrence matrix $c(m, n)$ defined as the distribution of co-occurring neighbor gray values. For a complete guide to statistical description consult [15].

4.3. Contour Profile Descriptor

In this section we introduce a novel descriptor to describe micro structures present along the perimeter of grains. As described in Tab. 2, some pollen types have reticular exines, which is translated into corrugated contours at the zenithal microscope view. In computer vision terms this means that the variance of gray levels along the pollen contour is higher than in pollen with no reticular surface. Although this measure could not classify by itself one specific pollen type, it can work as an efficient support tool for discriminating among pollen groups of highly, medium and low reticular exines.

The first step uses the center of mass of binary masks described in Sec. 3 to accomplish square cropping. One simple way of making this calculation is by means of moment equations described in Sec. 4.5. After that Cartesian coordinates (x, y) are transformed into polar logarithmic coordinates $\rho = log(\sqrt{(m - m_c)^2 + (n - n_c)^2})$, $0 \leq \rho \leq \rho_{max}$ and $\omega = arctan((m - m_c)/(n - n_c))$, $0 \leq \omega \leq 2\pi$ [23]. See in Fig. 6-(b) and (e) a couple of examples of polar logarithmic transformations.

The second step computes a snake algorithm from the bottom of the polar transformed images (outskirts in Cartesian coordinates), i.e. from the maximum radius. Starting from a horizontal line, each location (pixel) of that line is moved upwards if the gray level at the current location is higher than a given threshold (we used 30% of the maximum graylevel). When a dark gray

Energy	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j)^2$
Variance	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}(i-\mu)^2 \cdot c(i,j)$
Contrast	$\sum_{n=0}^{H-1} n^2 \Big(\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j)\Big), \quad \lvert i-j\rvert = n$
Dissimilarity	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} \lvert i-j\rvert \cdot c(i,j)$
Correlation	$\frac{1}{\sigma_x\sigma_y}\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} i\cdot j\cdot c(i,j) - \mu_x\mu_y$
Autocorrelation	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} i\cdot j\cdot c(i,j)$
Measure of Correlation 1	$\frac{T-HXY1}{max(HX,HY)}$
Measure of Correlation 2	$(1-exp[2\cdot(HXY2-T)])^{0.5}$
Cluster Shade	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}(i+j-\mu_x-\mu_y)^3 \cdot c(i,j)$
Cluster Prominence	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}(i+j-\mu_x-\mu_y)^4 \cdot c(i,j)$
Maximum Probability	$max(c(i,j)), \quad i=[0...H-1],\, j=[0...H-1]$
Entropy	$T=-\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j)\cdot log(c(i,j))$
Sum Average	$\sum_{i=0}^{2(H-1)} i\cdot c_{x+y}(i)$
Sum Entropy	$\sum_{i=0}^{2(H-1)} c_{x+y}(i)\cdot log(c_{x+y}(i,j))$
Sum Variance	$-\sum_{i=0}^{2(H-1)}(i-SumEntropy)^2\cdot c_{x+y}(i)$
Difference Entropy	$-\sum_{i=0}^{H-1} c_{x-y}(i)\cdot log(c_{x-y}(i,j))$
Difference Variance	$\sum_{i=0}^{H-1} i^2\cdot c_{x-y}(i)$
Homogeneity 1	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}\frac{c(i,j)}{1+\lvert i-j\rvert}$
Homogeneity 2	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}\frac{c(i,j)}{1+(i-j)^2}$

H bins number, HX and HY entropy of p_x and p_y.

$\mu_x = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} i\cdot c(i,j); \qquad \mu_y = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} j\cdot c(i,j)$

$c_x(i) = \sum_{j=0}^{H-1} c(i,j); \qquad c_y(j) = \sum_{i=0}^{H-1} c(i,j)$

$\sigma_x = \sqrt{\sum_{i=0}^{H-1} c_x(i)(i-\mu_x)^2}; \qquad \sigma_y = \sqrt{\sum_{j=0}^{H-1} c_y(i)(i-\mu_y)^2}$

$c_{x+y}(k) = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j); \quad i+j=k,\, k=[0...2(H-1)]$

$c_{x-y}(k) = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} p(i,j); \quad \lvert i-j\rvert = k,\, k=[0...H-1]$

$HXY1 = -\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j) log(c_x(i)\cdot c_y(j))$

$HXY2 = -\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c_x(i)\cdot c_y(j)\cdot log(c_x(i)\cdot c_y(j))$

Table 5: Second order statistical descriptors (Haralick).

Algorithm 1: Contour Profile Descriptor

Transform image into polar coordinates $PIMAGE(radius, angle)$;
$SNAKE(angle) = radius_{max}$;
$radius = radius_{max}$;
while *angles exist and are not anchored* **do**
  if $[(SNAKE(angle \pm 1) < aCurvature)$ **and**
  $(PIMAGE(SNAKE(angle) + 1, angle) > aThreshold)]$ **then**
    $SNAKE(angle) = radius - 1$;
  else
    anchor $SNAKE(angle) = radius$;

pixel is found it may belong highly probably to the pollen contour and the snake at that point is fixed. Some elastic properties are given to the snake, so that it can fit to the curved contour. We used 1 pixel maximum curvature. Previously the polar image is smoothed by a 5×5 uniform filter to remove spurious values and outskirts debris. Such contours found by snakes do not necessarily match with those binary contours found in Sec. 3 for the binary masks. Other descriptors that operate globally in a given region could not require a segmented region extremely precise. However this contour profile descriptor in concrete requires a path as much precise as possible. In any case such snaked-contours could be also applied for all descriptors as some studies revealed some improvements [10]. Considering the snake as a uni-dimensional function, the whole algorithm is described in Algorithm 1.

The third and final step draws a profile of gray levels along the snake and measures its variance in relation to the mean local value obtained by smoothing in our case the gray profile with a 21-bin uniform filter. See Fig. 6-(c) and (f). A high contour profile variance will indicate that the grain contour, the exine, is probably reticulated and a low variance means

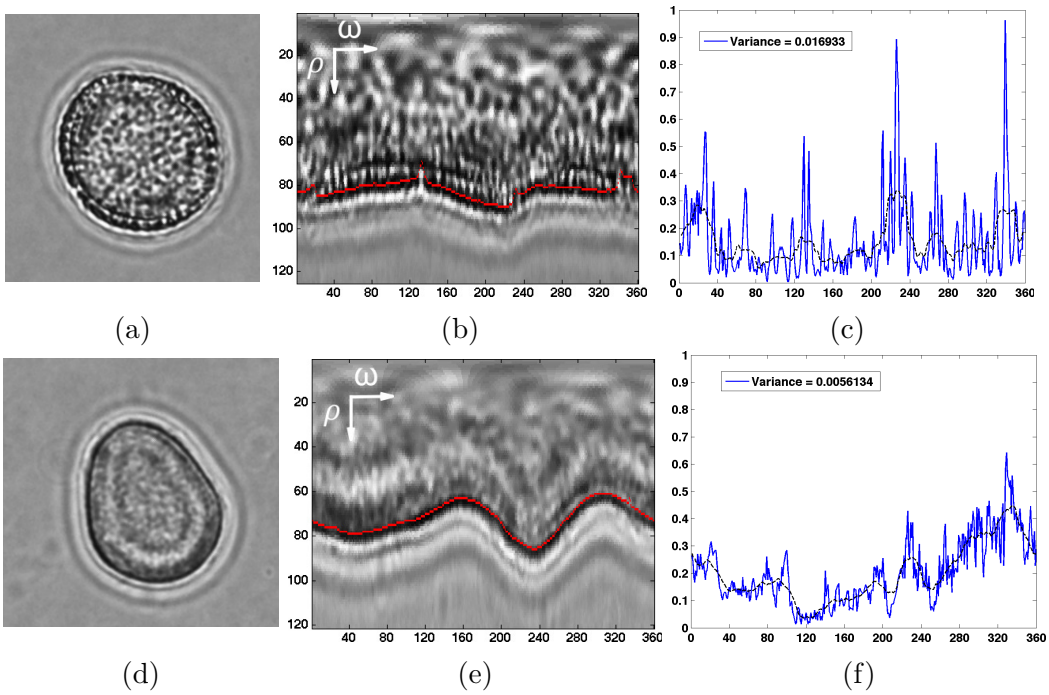


Figure 6: Reticular descriptor of contour profile (exine). (a) *olea* and (d) *echium* samples, (b) and (e) polar transformation, (c) and (f) gray level profile and variance measurement (blue line) and local mean obtained with a 21-bin uniform filter (black dotted line).

no reticulation.

4.4. Local Binary Patterns

The Local Binary Pattern (LBP) operator [24] is based on the idea that textural properties within homogeneous regions can be mapped into patterns, which represent micro-features. It uses a 3×3 square mask called "texture spectrum," to compare masked values with their central pixel, those ones lesser are labeled with "0" otherwise with "1". The labeled pixels are multiplied by a fixed weighting function and summed to obtain a label: $LBP(g_c) = \sum_{p=0}^{7} s(g_p - g_c)2^p$, where $\{g_p | p = 0, \ldots, 7\}$ are the neighbors of

g_c and the comparison function is defined as: $s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Ojala et al. [25] improved their proposal by including a circular mask denoted by the subscript (P, R) where P is the number of sampling points and R is the radius of the neighborhood. If sampling coordinates, $(x_p, y_p) = (x_c + R\cos(\frac{2\pi p}{P}), y_c - R\sin(\frac{2\pi p}{P}))$, do not fall at integer positions, then the values are bilinearly interpolated. Furthermore, they observed that over 90% of patterns can be described with few LBP patterns, so, they introduced a uniformity measure $U(LBP_{P,R}(g_c)) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)|$, which corresponds to the number of transitions (0/1) in the labeled LBP.

In this way, the uniform-LBP ($LBP^{uni}_{P,R}$) can be obtained as:

$$LBP^{uni}_{P,R}(g_c) = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R}(g_c)) \leq 2 \\ P + 1 & \text{otherwise} \end{cases} \quad (6)$$

After this process is completed a labeled image is generated and the pixel-wise information is encoded as a histogram, so that it can be interpreted as a fingerprint or a signature of the analyzed object. $LBP^{uni}_{P,R}$ produces $(P + 2)$-bin histograms [26]. Then from all statistical descriptors only mean, variance, asymmetry and kurtosis are computed assuming a studied trade-off between overloaded dimensionality vs. accuracy impairment [27].

4.5. Hu Moments

Image moments, originally proposed by [28], describe not only invariant morphological features of shapes but also high order statistical features. They are formulated as follows:

$$\mu_{pq} = \sum_{m} \sum_{n} (m - m_c)^p \cdot (n - n_c)^q \cdot g(m, n) \tag{7}$$

Where $m_c = \frac{r_{10}}{r_{00}}$ and $n_c = \frac{r_{01}}{r_{00}}$.

4.6. Space-frequency Descriptors

These do not really constitute descriptors themselves but transformations where features, somehow hidden, arise with higher visibility. It is in this transformed domains where features are measured by applying the mathematical previously introduced descriptors, in our case the statistical descriptors. For every sub-band there are 241 statistical descriptors (13 1^{st}-order statistical plus 19 2^{nd}-order statistical with 3 distances and 4 orientations) and for the whole transformed domain with 4 scales the total number of descriptors is 964, see Sec. 4.2.

4.6.1. Fourier Transform

It is the first formal proposal to analyze spectral contain of a signal where 2D frequencies arise in this context from graylevel variations along features like contour, edges, stripes and other periodic structures like textures. The Fourier spectrum is split in octave bands apart and averaged as follows:

$$Fourier_l = \sum_{r=\frac{1}{2^l}}^{\frac{1}{2^{l-1}}} \sum_{\theta=0}^{360} |I(r, \theta)| \; dr d\theta, , \quad l \in \{1, .., L\} \tag{8}$$

Where I is the Fourier transform of the image in polar coordinates and L the number of levels. For the experiments we selected 4 decomposition levels, which is a common value in most computer vision algorithms. Note the continuous DC-component (luminance) is disregarded and binary mask cannot be used here since the spatial dimension is lost.

4.6.2. Wavelet Transform

Although some pollen types present stationary structures (textures) through their inner regions, they mostly present a different spectral content not only along the interior but also along their contour. The Fourier transform loses the space localization where this happens and frequencies from different areas are mixed. In the late 80's wavelets were firstly proposed with Daubechies and Mallat as main precursors [29]. We used the overcomplete version and 5 stem long of Daubechies basis to build our descriptors as the energy on every scaled level as follows:

$$Wavelet_l = \sum_{o=1}^{3} \sum_{u=1,v=1}^{U,V} |W_{l,o}(u,v) \cdot B(u,v)|, , \quad l \in \{1, .., L\} \tag{9}$$

Where W is the wavelet transform of the image for the scale l and orientation o whose vertical and horizontal coordinates extend to $U = M$ and $V = N$. Based on preliminary observations, the number of scales was set to $L = 4$. Note that to achieve orientation invariance all bands in the same scale must be summed, i.e the vertical, horizontal and diagonal bands. The residual DC-component was discarded. As previously stated, prior to this wavelet average the coefficients are masked by the corresponding contour and inner binary masks $B(u, v)$.

4.6.3. LogGabor Transform

Firstly proposed by Dennis Gabor in 1946 [30], the Gabor filters are different versions of a Gaussian-shaped window modulated by a sinusoid. The result is the partition of the Fourier plane into bands modulated in orientation and octave bands apart in frequency. Gaussian shape ensures an optimum spreading in both dimensions, i.e. space location vs. frequency discrimination, while one weakness of wavelets is the pronounced frequency overlapping. In addition, the Gaussian envelop is modulated by a complex exponential with odd and even phases, which is effective for analyzing features with odd phase like ridges and even phase like edges. In this study we used the overcomplete implementation of logGabor filters proposed by Fischer et al. [31] never tested before for this task. Similarly to wavelets, the logGabor descriptor is formed by calculating the energy at every scaled level:

$$\begin{aligned} Gabor_l &= \textstyle\sum_{o=1}^{O} \sum_{u=1,v=1}^{U,V} |F^{-1}\left(G_{lo} \cdot I\right) \cdot B(u,v)|, , \quad l \in \{1,..,L\} \\ G_{lo} &= \exp\left(-\tfrac{1}{2}\left(\tfrac{\rho-\rho_l}{\sigma_\rho}\right)^2\right) \exp\left(-\tfrac{1}{2}\left(\tfrac{\theta-\theta_{pl}}{\sigma_\theta}\right)^2\right) \end{aligned} \tag{10}$$

Where F^{-1} is the inverse Fourier transform, G_{lo} is the logGabor filter with L scales and O orientations in log-polar coordinates (ρ, θ) and $(\sigma_\rho, \sigma_\theta)$ are the angular and radial bandwidths, see [31] for more implementation details. Again $L = 4$, $O = 6$ and the residual DC-component is discarded.

5. Discriminant Analysis

In case of task-specific methods like this study, descriptors are either chosen after a comprehensive literature review study but also after empirical experiment feedback. With a generalized descriptor extraction, the large set of image descriptors provides an extensive numeric description of the image content [32]. However, descriptors that are discriminant for one specific dataset may not be discriminant for others, probably because they describe features that are widely spread along all classification groups or because they are redundant (correlated) with respect to other descriptors. In that case such descriptors provide useless information that moreover will likely degrade the classification performance not only in terms of accuracy but also in terms of speed due to the higher dimensionality [33, 34]. To reduce the number of irrelevant descriptors, discriminant analysis minimizes the classification error for the smallest possible subset of descriptors. Thousands of descriptors are extracted from the methods here described and it is a fact widely studied that beyond a certain limit an increasing number of descriptors not only provokes an increasing computational time but also impairs classification [35]. Therefore a feature selection process is then required to remove redundant information.

5.1. Floating Selection

A preliminary study was carried out to elucidate which are individually the most discriminant descriptors. For that we employed the Sequential Forward Feature Selection (SFFS) [36]. It constructs an incremental priority subset of descriptors by adding the descriptor in the excluded subset that

1	L	asimmetry_LBP.Org	26	S	Dissi_1_135.Org	51	S	Dissi_3_0.KEx	76	S	Contr_5_0.Org
2	L	curtosis_LBP.Org	27	S	Dissi_1_90.Org	52	S	Dvarh_5_135.KEx	77	S	Entro_3_0.Org
3	M	Perimeter.Org	28	S	Dissi_1_45.Org	53	S	Denth_5_45.Org	78	S	Denth_3_0.KEx
4	M	Area.KEx	29	S	Denth_1_135.Org	54	S	Dissi_3_45.Org	79	S	Denth_1_135.KEx
5	M	Area.Org	30	S	Dvarh_3_0.KEx	55	S	Denth_5_90.Org	80	S	Dissi_3_135.KEx
6	M	EquivDiameter.Org	31	S	Denth_3_0.Org	56	S	Denth_5_135.Org	81	S	Dissi_3_45.KEx
7	M	EquivDiameter.KEx	32	S	Dissi_1_0.KEx	57	S	Entro_1_0.Org	82	G	OrgS_Gbf6
8	M	Perimeter.KEx	33	S	Denth_1_90.Org	58	S	Dissi_1_90.KEx	83	S	Denth_1_0.KEx
9	S	Entropy.Org	34	S	Dissi_3_90.Org	59	S	Dissi_1_45.KEx	84	S	Denth_1_90.KEx
10	S	Entropy.KEx	35	S	Dvarh_5_0.KEx	60	S	Dissi_1_135.KEx	85	S	Entro_3_90.Org
11	S	Dvarh_3_0.Org	36	S	Dvarh_3_45.KEx	61	S	Dissi_5_90.Org	86	S	Denth_3_45.KEx
12	S	Dvarh_1_135.Org	37	S	Denth_1_0.Org	62	G	OrgS_Gbf26	87	S	Denth_1_45.KEx
13	S	Dvarh_3_135.Org	38	S	Dvarh_3_135.KEx	63	S	Entro_1_135.Org	88	S	Denth_5_0.KEx
14	S	Dvarh_3_90.Org	39	S	Dvarh_3_90.KEx	64	S	Contr_3_135.Org	89	S	Denth_3_135.KEx
15	S	Dvarh_3_45.Org	40	S	Denth_1_45.Org	65	S	Dvarh_1_90.KEx	90	S	Denth_3_90.KEx
16	S	Dvarh_5_0.Org	41	S	Denth_3_45.Org	66	S	Contr_3_90.Org	91	S	Contr_3_45.Org
17	G	OrgS_Gbf2	42	S	Dvarh_1_135.KEx	67	S	Dissi_5_0.KEx	92	S	Contr_1_0.Org
18	G	OrgS_Gbf14	43	S	Denth_3_90.Org	68	S	Entro_1_90.Org	93	S	Contr_5_90.Org
19	S	Dissi_1_0.Org	44	S	Dissi_3_135.Org	69	S	Entro_1_45.Org	94	G	OrgS_Gbf8
20	S	Dvarh_5_45.Org	45	S	Denth_3_135.Org	70	G	OrgS_Gbf12	95	S	Denth_5_45.KEx
21	S	Dvarh_5_90.Org	46	S	Dvarh_5_45.KEx	71	G	OrgS_Gbf4	96	S	Contr_1_45.Org
22	S	Dvarh_1_90.Org	47	S	Denth_5_0.Org	72	S	Dvarh_1_45.Org	97	S	Dissi_5_90.KEx
23	G	OrgS_Gbf1	48	S	Dvarh_5_90.KEx	73	S	Contr_1_135.Org	98	S	Denth_5_90.KEx
24	S	Dvarh_5_135.Org	49	S	Dissi_5_0.Org	74	S	Dissi_3_90.KEx	99	S	Denth_5_135.KEx
25	S	Dissi_3_0.Org	50	S	Contr_3_0.Org	75	G	OrgS_Gbf10	100	S	Entro_3_135.Org

Table 6: The 100 most discriminant descriptors listed in order of importance. *L* stands for LBP, *M* for morphological, *S* for statistical and *G* for logGabor. Extension *Org* means that the descriptor was calculated in the pollen kernel and *KEx* in the outer exine.

increments the highest the classification rate. For every step the priority subset is re-arranged and re-examined in case of one of the descriptors impairs classification with the new formed group. As a result this algorithm often converges to a ordered list by discrimination capacity. In Tab. 6 there is such a list for the first 100 features. The percentage of LBPs descriptors is only 2% although they are on the top. The percentage in that list of morphological descriptors is 6%, logGabor 9% and the statistical descriptors 83%. Although this is an interesting list for elucidating some discriminant descriptors, one should bear in mind that later on their contribution will be altered by the LDA transformation (see Sec. 5.3).

5.2. Correlation

Some overloaded features like Fourier, wavelets and Gabor are treated like new domains themselves where the whole bank of statistical descriptors can be calculated and extracted from each decomposition band. This means that the total number of descriptors becomes 4 times larger given a 4-level decomposition transform. This approach increases the workload to a cuttered extend to be easily computable. Therefore, for these feature groups of space-frequency transform we decided to remove those statistical variables that are highly correlated.

To remove redundant information we firstly used the correlation coefficient as the similarity measure between two or more features. Hence, a threshold value must be defined for determining the correlation value from which features are considered redundant. This measure has been commonly adopted for unsupervised feature selection [37]. In our study, an empirical threshold of 98% was adopted. Such a decision reduced the number of descriptors in such a way that for instance a 4-level Fourier bands with 964 initial descriptors gets shortened to 224, 4-level wavelet decomposition from 964 to 143 or the actual statistical descriptors from 964 to 147. In summary with this technique it is achieved an overall 75-80% dimensionality reduction and the corresponding computing time. In return classification accuracy, according to calculations shown afterwards, is insignificantly affected.

5.3. Linear Discriminant Analysis

Since the previous methods generate high dimensional feature vectors and a limited dataset is available in our context, Linear Discriminant Analysis (LDA) [38] constitutes an efficient tool for dimensionality transformation.

Since oversized spaces crowd together classes which impairs classification, LDA transforms the original space into an orthogonal and linear space where feature vectors are prioritized in order of importance while others are rejected. This implies that classes must be linearly separated which is not always fulfilled. LDA also requires unimodal Gaussian likelihoods which was so validated.

6. Training & Classification Techniques

Dimensionality reduction is a fundamental step in any classification problem. In most cases we cannot assume parameter independence, which prevents from separately assessing each parameter from the rest. This issue is the so called Model Selection Problem (MSP). In this case we validated normal distributions by means of normality test of K.S. normality test, Levene's homocedasticity and the analysis of variance (ANOVA) [39, 40].

10-*fold* cross-validation is a simple and yet widely employed method for model evaluation that randomly splits up data into 10 disjoint subsets of approximately equal size. Each fold is then classified separately by using the remaining 9 subsets to train the model. In the end the average of all folds provides an estimation of the classification accuracy of the model. A similar procedure was exercised with groups of 1 element, also called *leave-one-out*. Both training methods threw similar results and for the sake of simplicity only leave-one-out will be presented in the experiments.

Although many classifiers can be found and some of them could significantly improve accuracy rates, it is not the main purpose here to carry out a thorough analysis of classifiers' performance, but again to discover which

descriptor or combination of descriptors better discriminate between pollen types. Hence, although we do compare an extensive bank of classifiers like nearest-neighbor, k-means, Parzen classifier, decision tree, neural networks, quadratic Bayes normal classifier, Fisher classifier, linear discriminant or support vector machine, we selected here three representative ones, which in turn were three of the best classifiers tested.

6.1. Fisher classifier

Fisher's linear classifier finds a linear discriminant function by minimizing the errors in the least square sense [35]. This linear discriminant is based on finding a direction in the feature space such that the projection of the data minimizes Fisher's criterion, i.e., the ratio of the squared distance between the class means and averaged class variances. The linear classifier is then perpendicular to this projection.

6.2. Support Vector Machine

Support Vector Machines (SVM) finds a discriminant function by maximizing the geometrical margin between positive and negative samples [41]. Thus, the space is mapped so that examples from different classes are separated by a gap as wide as possible. Besides linear classification, SVMs act as a non-linear classifier by using the so-called kernel trick. This trick can be considered a mapping of the inputs onto a high-dimensional feature space in which classes become linearly separable. SVMs minimize both training error and geometrical margin. The latter accounts for the generalization abilities of the resulting classifier. SVMs are one of the best classifiers available and have been applied to many real-world problems.

6.3. Random Forest

A Decision Tree (DT) is a conceptually simple, yet robust, and widely used tool for decision support in which classification is performed through a tree graph [42]. The classification starts from an initialization node (*root node*) from which a given test sample is tested at each stage (*internal node*) of the classification, all the way down to the end of a tree branch (*leaf* or *terminal node*) [43]. The path followed by the sample depends on threshold-based conditions associated to each internal node.

To select the optimal threshold-based conditions, DT algorithms make use of a brute force method, which consists of testing all potential variables and selecting the variable that maximizes a given criterion. When building the DT, this criterion characterizes the quality of the split created by the transition from an internal node to its associated leaves [43].

To improve classification accuracy and robustness, the Random Forest (RF) classifier, built upon an ensemble of DTs, learn from different subsets of the training dataset and no pruning is performed after their construction [42]. Each DT is built using the values of random feature vectors in a way that all DTs from the RF possesses the same distribution. The random feature vectors may be generated using several techniques, such as bagging [42], random split selection [44] and the so-called random subspace technique [45]. When classifying an unknown sample, its feature vector is tested using all DTs of the RF. Their outputs constitute votes for the most popular class, which in turn is the RF prediction. Nowadays, the RF classifier is considered as one of the most accurate learning algorithms and its performance has been proven on many datasets [46].

7. Results

Given the total number of descriptors is 6 320, considering that space-frequency, moments, LBPs and statistical descriptors are duplicated due to the contour-interior pollen segmentation. Given also the feature reducing and transformation algorithms (correlation and LDA) and the three selected classifiers (Fischer, SVM and Random Forest), the number of possible experiments is considerable. We organized the results in several experiments to show concrete aspects of descriptors and classifiers.

Although the experiments are driven according to the best accuracy/error rates, one should bear in mind that the obtained absolute values could be hardly compared quantitatively to other studies due to discrepancies in feature vector, dimensionality reduction and/or classifiers. One example case could be [14] with an astonishing number of particles/grains and classification rate around 98.5%, but not comparable because 1) 3D information is additionally incorporated 2) pollen grains were only 12% of the training date set and 3) allergenic pollen is not necessarily the same pollen than for beekeeping.

7.1. Experiment 1: whole and contour+inner segmentation.

Descriptors were tested here according to their mathematical definition type. At the same time descriptors were compared when applied for both segmentation cases: whole-segmented grain and contour+inner segmentation. In order to simplify the case study, the same classifier was used for all of them, in this case Fisher but with no particular reason, except for its accurate performance behavior. Albeit similar results can be observed with

the remaining classifiers.

From the plot in Fig. 7 most descriptors lead to similar classification error around $\varepsilon \sim 0.3$ when they operate individually. Two of them are above 0.7 though, i.e. LBP and moments. This corroborates that macro-features derived form morphology descriptors already provide competitive accuracy on a par with local micro-feature analysis performed by (spatio)-frequential descriptors. We have no plausible explanation for the lower rates delivered by LBP and moments. Note that although some LBPs had an important discriminant capacity in Fig. 6, all together combined do not perform as high as other texture descriptors. This could be to the reason that the number of LBP descriptors 4 is not actually enough for the current database. The McNemar's significance test [47] provides a confidence value to accept that methods are statistically significant provided a minimum threshold typically chosen 95% of confidence. That value is found by projecting the classification discrepancies of both methods through a chi-quadratic function as a expectation model for binomial distributions. McNemar's threshold delivered here a more than amply averaged value $T = 230.98 > \chi^2_{1,0.95} = 3,84$ except for 5 pairs which can not be considered significative: Wavelets vs. Fourier, morphological and statistical; Fourier vs. statistical and morphological vs. statistical on averaged $T = 1.95$.

Another favorable observation reveals a significant improvement when computing doubly but separately for contour and inner pollen regions, as we hypothesized. Overall, one can observe a drastic decline of errors achieving an overall accuracy improvement of 50% ($T = 344.62$). Note that morphological descriptors have no counterpart for contour+inner segmentation. All

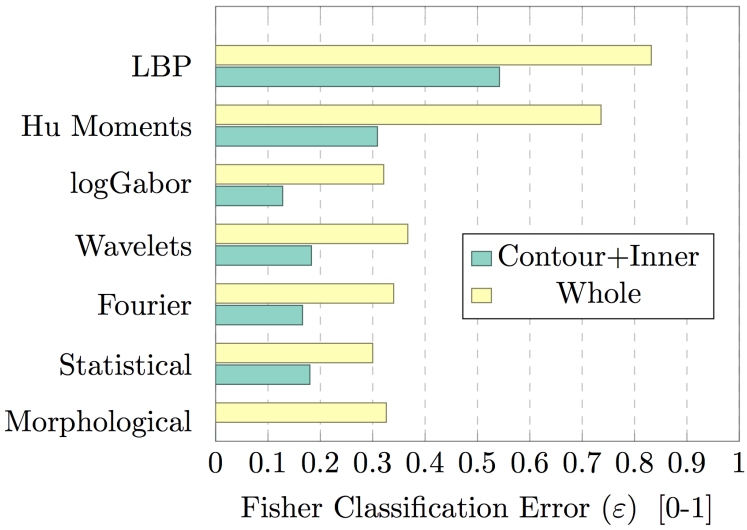


Figure 7: Comparing descriptors types under the Fisher classification error.

subsequent experiments will consider exclusively contour-inner segmentation.

To justify the rejection of the space-frequency descriptors above a 98% correlation threshold (see Sec. 5.2), the classification error obtained for instance with logGabor is $\varepsilon = 0.124$, while with the 100% of descriptors a similar error is obtained $\varepsilon = 0.128$ and the significance test is low $T = 0.78$.

7.2. Experiment 2: descriptor types combinations.

The combination of different types of descriptors can strengthen the discrimination capacity. In Fig. 8 morphological and statistical descriptors constitute a baseline for comparison since they provided high accuracy rates and bring together the two main discriminant features: shape and texture. From this chart, when statistical descriptors are combined with morphological descriptors provided together an improvement of 51.1%. Adding to these two groups the space-frequency descriptors separately provided an over-

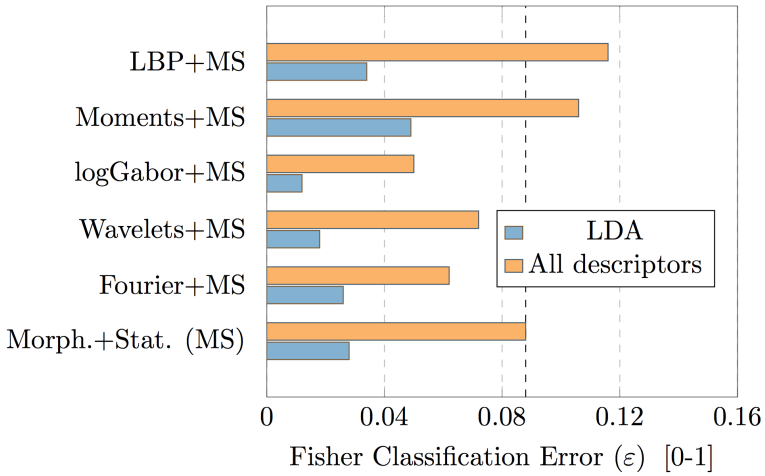


Figure 8: Accuracy performance of Morphological+Statistical (MS) and remaining descriptors with contour+inner segmentation.

all accuracy improvement of 34% (averaged $T = 81.31$), while the logGabor+morphological+statistics are the most accurate with $\varepsilon = 0.05$. Note that moments and LPB produces slight impairments. Significance test between pairs delivered an averaged value ($T = 23.63$), except for Fourier vs. Wavelets ($T = 1.35$) and Moments vs. LBP ($T = 1.04$). This tendency is confirmed in Fig. 9 where morphological+statistical combined with space-frequency descriptors achieve the lowest error rates. Particularly the combination of morphological+statistical+all space-frequency descriptors provided the lowest error rate $\varepsilon = 0.032$ ($T = 31.05$ averaged with comparative cases). Note also that neither LBP nor moments barely affect performance.

7.3. Experiment 3: LDA dimensionality reduction.

In previous Figs. 8 and 9 already compared the improvement achieved by LDA. All combinations of descriptors augmented their accuracy significantly and reduced the overall classification error around 70% (averaged $T = 89.54$).

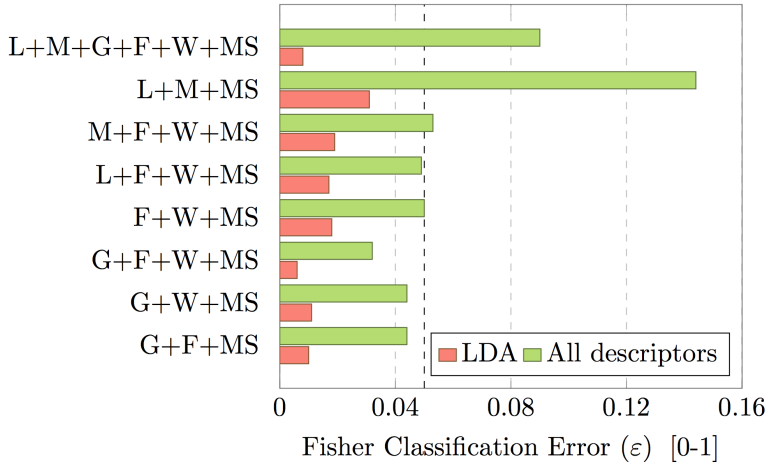


Figure 9: Accuracy performance (contour+inner segmentation) of Morphological+Statistical (MS) and combinations of remaining descriptors: Wavelets (W), Fourier (F), logGabor (G), Moments (M), LBP (L). Note that the vertical dashed line corresponds to previous logGabor+MS error rate 0.05.

This suggests that, besides unimodal Gaussian likelihoods tested in advance, most descriptors can be linearly separated. Now most combinations classify with less than 2% error and an outstanding case is the combination of morphological, statistical descriptors with the three space-frequency descriptors achieving 99.4% accuracy rate ($\varepsilon = 0.006$), see its confusion matrix in Tab. 7. Note also that LDA successfully deals with the addition of LBP and moments achieving a similar error $\varepsilon = 0.008$ and low significant difference $T = 0.35$ for these two cases.

7.4. Experiment 4: Classifiers

In Fig. 10 Fisher classifier was compared with SVM and Random Forest by using four reference groups of descriptors. Although there is a slight improvement tendency in favor of Random Forest, outcomes did not show preference in all groups, nor even for other groups not shown here. Thus no

ACTUAL	PREDICTED *Aster*	*Brassica*	*Campanula.*	*Carduus*	*Castanea*	*Cistus*	*Cytisus*	*Echium*	*Ericaceae*	*Helianthus*	*Olea*	*Prunus*	*Quercus*	*Salix*	*Teucrium*
Aster	120	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brassica	0	119	0	0	0	0	0	0	0	0	1	0	0	0	0
Campanula.	0	0	118	0	0	0	2	0	0	0	0	0	0	0	0
Carduus	0	0	0	119	0	0	0	0	0	0	0	1	0	0	0
Castanea	0	0	0	0	120	0	0	0	0	0	0	0	0	0	0
Cistus	0	0	0	0	0	120	0	0	0	0	0	0	0	0	0
Cytisus	0	0	0	0	0	0	120	0	0	0	0	0	0	0	0
Echium	0	0	0	0	0	0	0	120	0	0	0	0	0	0	0
Ericaceae	0	0	0	0	0	0	0	0	120	0	0	0	0	0	0
Helianthus	6	0	0	0	0	0	0	0	0	114	0	0	0	0	0
Olea	0	1	0	0	0	0	0	0	0	0	119	0	0	0	0
Prunus	0	0	0	0	0	0	0	0	0	0	0	120	0	0	0
Quercus	0	0	0	0	0	0	0	0	0	0	0	0	120	0	0
Salix	0	0	0	0	0	0	0	0	0	0	0	0	0	120	0
Teucrium	0	0	0	0	0	0	0	0	0	0	0	0	0	0	120

Table 7: Confusion matrix for the best case combining Morphological+Statistical+Fourier+Wavelets+logGabor descriptors and LDA.

pair of group and classifier revealed evident superiority here. This leads to the point that descriptors and classifiers must be selected as two parts that work together accordingly.

8. DISCUSSION: Grain features, limits and possibilities

Several important contributions have been made in this work. A considerable data base of pollen grains has been elaborated. It was compound of 15 pollen types and 120 samples per type manually cropped. Grains have been automatically segmented (manually revised) to make binary masks. A bench of the state of the art in morphological, statistical and texture descriptors together with a new contour profile descriptor has been exhaustively tested for classifying the 15 pollen types. Some of them like texture descriptors have not been evaluated before in this field. Furthermore we proposed a novel contour+inner segmentation which provided an overall 50% improvement up

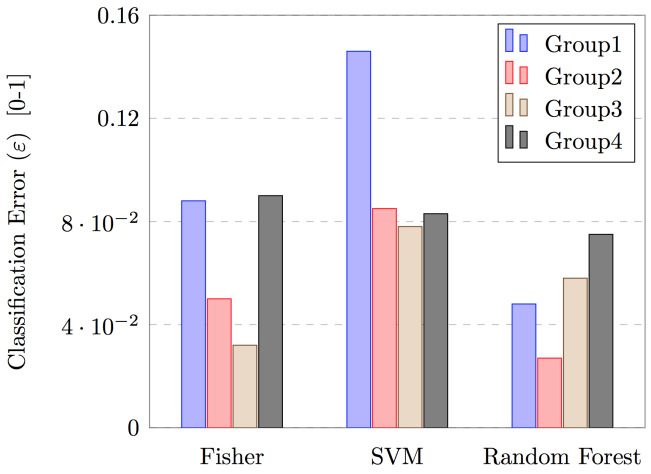


Figure 10: Performance comparison of classifiers Fisher, SVM and Decision Tree with Group1 (MS), Group2 (MS+F+W) and Group3 (MS+G+W+F) and Group4 (MS+W+F+G+M+L).

to 99.4% accuracy. We concluded that the traditional morphological and statistical descriptors together with space-frequency representations, specifically logGabor, provided the best classification accuracy rates. Moreover the dimensionality reduction with LDA improved classification by 70%. Along this research we also come across with other several challenging issues addressed in the following.

After long conversations with palynologists, they argue that the geometric shape and number of apertures are the first aspects they look for in a preliminary screening. Although size is an effective feature to discriminate between broad groups of pollen types, this should be taken carefully since size could vary more than 10 microns in some circumstances like the substrate conditions or how much water received the plant. Thus it could be recommended to simplify the size in two classes: small and medium-large. Other morphological refinement would consider triangularity, rhomboicity or

even more complex shapes adapted to the wide range of shapes exhibited by the pollen. After that, depending on the possible pollen types according to their morphology, the search would resume for some other details like reticular texture, spikes, apertures, exine width,... etc. This leads again to the idea of a hierarchical-tree classification.

Unfortunately some pollen types are almost equal even for experts which shows up the difficulty of this task. For instance *Cruciferae* and *Olea* are similar except for the polar area and apertures. *Retama* and *Cytisus* belong to the same family *Fabaceae* and are consequently similar except for the almost negligible aperture and triangular view of the former.

In addition to the obvious place of origin, the first and effective way of classifying pollen is the color ball. Bees rarely harvest pollen of different types in the same ball, therefore the knowledge of a given color drastically reduces the number of possible candidates. Such a description can be easily incorporated in an automatic recognition software by displaying a color chart where the user could select the most similar ball tonality. This pre-processing stage could help computer vision tasks not only by reducing computation time but also reducing error classification rates (decision tree). As previously indicated, the color ball is not consistent with the color later observed through the microscope and furthermore we found no evidence that color could have any discriminant capacity beyond this point, therefore images were converted to grayscale.

A pollen type may present different appearances according to their view with respect to the z-view and consequently their morphological and also statistical descriptors can drastically vary. Although it is still unclear how

to handle such a difficulty, three alternatives are discussed. Multimodal classifiers deal with complex probability distributions functions made of several monomodal pdf's. This could be the case of some morphological descriptors like 'circularity' (and related) that clearly present two or more probability means depending on the view. This approach is more database consistent, although increases the complexity. Another approach could split those multi-appearance types in two or more different subclasses (polar, meridian,...) which refines database description and simplifies classification, however this doubles the effort for labeling each pollen type and collecting more training samples. Hybrid tree-monomodal classification or decision rule could also be employed for classifying in two steps. Firstly the monomodal classifiers could be used for those descriptors not affected by the z-laying and secondly for those sensible to view, without making separated classes for each z-laying. This alternative demands a smaller training dataset than the previous option, however it needs for a decision rule adjustment and still needs for tedious polar/meridian/others labeling.

Precise segmentation is a critical point for the whole classification process. Some studies using snakes and other computer vision techniques showed remarkable results [48]. Effectively segmentation should be done accurately, although our purpose in the current study focuses on comparing descriptor's discriminant capacity while optimum classification rates remain a secondary goal. Therefore we do not pursue perfect binary masks, but suitable enough to be equally shared by all descriptors. However since one of the main contributions here is the strategy of splitting grains in contour and inner parts, therefore segmentation techniques will require a further study.

In a real scenario there is a need for an 'unknown' class, also known as outlier detection. This class contains samples that do not belong to any of the trained classes. Furthermore rejection class is also required to embrace those ambiguous samples due to malformations or a bad cropping. Considering a commercial software the goal is to discover the origin of the harvested balls and a pollen ball contains hundreds of grains whose majority belongs to the same pollen type due to a smart habit of bees. In this scenario it is not therefore so critical the ratio of false-negative (attributed to a outlier-class), since there are hundreds of attempts to find out the principal pollen type. Instead false-positive ratio should be minimized as much as possible. In other words, once a pollen grain is considered to belong to a certain class, one must be highly confident on that assertion. Such a confident threshold has to be modeled according to not only the classification error but also according to population ratio present in every pollen ball or slice preparation.

9. Appendix:

Description of the main features of the 15 pollen types studied in this paper.

Aster - isopolar, radially symmetric, medium size ($P = 22 - 31\mu m$, $E = 20 - 29\mu m$), spheroidal to prolate (elliptic), in equatorial view ($P/E = 0.96 - 1.20$), circular or trilobulate in polar view, 3-zonocolporate, ornamentation echinate-perforate.

Brassica - isopolar, radially symmetric, medium size ($P = 21 - 30\mu m$, $E = 27 - 27\mu m$), oblate spheroidal to prolate (elliptic) in equatorial view ($P/E = 0.90 - 1.28$), circular or trilobulate in polar view, 3-zonocolpate, ornamenta-

tion reticulate.

Campanulaceae - isopolar, radially symmetric, small or medium size ($P = 18 - 29\mu m$, $E = 20 - 34\mu m$), oblate spheroidal to spheroidal in equatorial view ($P/E = 0.80 - 0.96$), subtriangular in polar view, 3-zonoporate, ornamentation echinate.

Carduus - isopolar, radially symmetric, medium to big size ($P = 31 - 50\mu m$, $E = 31 - 51\mu m$), oblate spheroidal to prolate (elliptic) in equatorial view ($P/E = 0.85 - 1.20$), trilobulate in polar view, 3-zonocolporate, ornamentation echinate-perforate-finely reticulate.

Castanea - isopolar, radially symmetric, small size ($P = 14 - 15\mu m$, $E = 9 - 11\mu m$), prolate (elliptic) in equatorial view ($P/E = 1.27 - 1.55$), triangular in polar view, 3-zonocolporate, ornamentation rugulate.

Cistus ladanifer - isopolar, radially symmetric, small to medium size ($P = 41 - 55\mu m$, $E = 50 - 53\mu m$), spheroidal to prolate spheroidal in equatorial view (P/E=0.91-1.19), circular in polar view, 3-zonocolporate, ornamentation reticulate.

Cytisus - isopolar, radially symmetric, medium to small size ($P = 20 - 33\mu m$, $E = 15 - 29\mu m$), spheroidal to prolate spheroidal (elliptic or subrhomboid) in equatorial view ($P/E = 0.86 - 1.55$), circular to trilobulate in polar view, 3-zonocolporate, ornamentation finely reticulate.

Echium - heteropolar, radially symmetric, small size ($P = 13 - 25\mu m$, $E = 8 - 15\mu m$), prolate (pyriform) in equatorial view ($P/E = 1.30 - 1.87$), trilobulate in polar view 3-zonocolporate, perforate-finely reticulate.

Ericaceae - tetragonal tetrads, medium to big size ($P = 27 - 67\mu m$), pollen 3-zonocolporate, ornamentation psilate to verrucate.

Helianthus - isopolar, radially symmetric, medium size ($P = 27-31\mu m$, $E = 27 - 33\mu m$), oblate spheroidal to spheroidal in equatorial view ($P/E = 0.90-1.00$), circular in polar view, 3-zonocolporate, ornamentation echinate-perforate.

Olea - isopolar, radially symmetric, small to medium size ($P = 20 - 27\mu m$, $E = 19 - 31\mu m$), spheroidal to prolate (elliptic) in equatorial view ($P/E = 1.05 - 1.31$), circular or trilobulate in polar view, 3-zonocolporate, ornamentation reticulate-verrucate.

Quercus - isopolar, radially symmetric, medium size ($P = 19-33\mu m$, $E = 15-34\mu m$), oblate to prolate (elliptic) in equatorial view ($P/E = 0.86-1.35$), circular or triangular in polar view, 3-zonocolpate, ornamentation granulate-verrucate.

Rubus - isopolar, radially symmetric, small to medium size ($P = 16 - 28\mu m$, $E = 14 - 24\mu m$), spheroidal to prolate (elliptic) in equatorial view ($P/E = 1.00 - 1.57$), circular or trilobulate in polar view, 3-zonocolporate or 3-zonocolporoidate, ornamentation reticulate.

Salix - isopolar, radially symmetric, small to medium size ($P = 16 - 27\mu m$, $E = 16 - 23\mu m$), prolate (elliptic) in equatorial view ($P/E = 1.10 - 1.20$), subtriangular in polar view, 3-zonocolporate, ornamentation reticulate.

Teucrium - isopolar, radially symmetric, medium to big size ($P = 35 - 66\mu m$, $E = 26 - 45\mu m$), spheroidal to prolate (elliptic) in equatorial view ($P/E = 1.05 - 1.65$), circular or triangular in polar view, 3-zonocolpate, ornamentation echinate-perforate.

Acknowledgements

This work has been supported by the European project APIFRESH FP7-SME-2008-2 "Developing European standards for bee pollen and royal jelly: quality, safety and authenticity" and we would like to thank to Mr. Walter Haefeker, President of the European Professional Beekeepers Association (EPBA). J. Victor Marcos is a "Juan de la Cierva" research fellow funded by the Spanish Ministry of Economy and Competitiveness. Rodrigo Nava thanks Consejo Nacional de Ciencia y Tecnología (CONACYT) and PAPIIT grant IG100814.

References

[1] S. Bogdanov, K. Bieri, G. Gremaud, D. Iff, A. Känzig, K. Seiler, et al., Swiss food manual chapter 23 b: Bienenprodukte-pollen, BAG (Swiss Federal Office for Public Health), Berne.

[2] S. Bogdanov, Pollen: nutrition, functional properties, health: A review, Bee Product Science.

[3] J. E. Eckert, The flight range of the honeybee, Journal of Agricultural Research 47 (1933) 257–285.

[4] N. Bradbear, et al., Bees and their role in forest livelihoods: a guide to the services provided by bees and the sustainable harvesting, processing and marketing of their products., Non-wood Forest Products (19).

[5] E. Gomez-Ordonez, P. Ruperez, FTIR-ATR spectroscopy as a tool for

polysaccharide identification in edible brown and red seaweeds, Food Hydrocolloids 25 (2011) 1514–1520.

[6] M. Sivaguru, L. Mander, G. Fried, S. Punyasena, Capturing the surface texture and shape of pollen. a comparison of microscopy techniques, PLOS One 7 (2012) e39129.

[7] J. Flenley, The problem of pollen recognition, in: Problems in picture interpretation, CSIRO, Camberra, 1968, pp. 141–145.

[8] R. Redondo, F. Šroubek, S. Fischer, G. Cristóbal, Multifocus image fusion using the log-Gabor transform and a multisize windows technique, Information Fusion 10 (2) (2009) 163–171.

[9] A. Boucher, P. Hidalgo, M. Thonnat, J. Belmonte, C. Galan, P. Bonton, R. Tomczak, Towards automation of palynology 3: pollen pattern recognition using Gabor transform and digital moments, J. of Quaternary Science 19 (2004) 763–768.

[10] M. Rodriguez-Damian, E. Cernadas, A. Formella, M. Fernández-Delgado, P. De Sa-Otero, Automatic detection and classification of grains of pollen based on shape and texture, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 36 (4) (2006) 531–542.

[11] C. Chen, E. Hendriks, R. Duin, J. Reiber, P. Hiemstra, L. A. Weger, B. Stoel, Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort, Aerobiologia 22 (2006) 275–284.

[12] J. R. Ticay-Rivas, M. del Pozo-Baños, C. M. Travieso, J. Arroyo-Hernández, S. T. Pérez, J. B. Alonso, F. Mora-Mora, Pollen classification based on geometrical, descriptors and colour features using decorrelation stretching method, in: Artificial Intelligence Applications and Innovations, Springer, 2011, pp. 342–349.

[13] M. Chica, Authentication of bee pollen grains in bright-field microscopy by combining one-class classification techniques and image processing, Microscopy Research and Technique 75 (11) (2012) 1475–1485.

[14] O. Ronneberger, Q. Wang, H. Burkhardt, 3d invariants with high robustness to local deformations for automated pollen recognition, in: Pattern Recognition, Springer, 2007, pp. 425–435.

[15] R. M. Haralick, K. Shanmugam, I. H. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics (6) (1973) 610–621.

[16] Y. Zhang, D. Fountain, R. Hodgson, J. Flenley, S. Gunetileke, Towards automation of palynology 3: pollen pattern recognition using Gabor transform and digital moments, J. of Quaternary Science 19 (2004) 763–768.

[17] M. Rodriguez-Damian, E. Cernadas, P. Sa-Otero, Pollen classification using brigthness-based and shape-based descriptors, in: Proceedings of the 17th Int. Conf. on Pattern Recognition, 2004, pp. 212–215.

[18] O. Ronneberger, Q. Wang, H. Burkhardt, 3D invariants with high robustness to local deformations for automated pollen recognition, in:

F. Hamprecht, C. Schnorr, B. Jahne (Eds.), Lecture Notes in Conputer Science, Vol. 4713 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 425–435.

[19] N. Q. Haas, Automated pollen image classification, Master's thesis, University of Tennessee (2011).

[20] S. Landsmeer, E. Hendriks, L. De Weger, J. Reiber, B. Stoel, Detection of pollen grains in multifocal optical microscopy images of air samples, Microscopy Res. and Technique 72 (2009) 424–430.

[21] P. Carrión, E. Cernadas, P. Sá-Otero, E. Dıaz-Losada, Could the pollen origin be determined using computer vision? an experimental study, in: IASTED International Conference on Visualization, Imaging, and Image Processing, 2002, pp. 74–79.

[22] C. M. Travieso, J. C. Briceño, J. R. Ticay-Rivas, J. B. Alonso, Pollen classification based on contour features, in: Intelligent Engineering Systems (INES), 2011 15th IEEE International Conference on, IEEE, 2011, pp. 17–21.

[23] J. Angulo, Polar modelling and segmentation of genomic microarray spots using mathematical morphology, Image Analysis and Stereology 27 (2) (2008) 107–124.

[24] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, in: Proceedings of the 12th International Conference on

Pattern Recognition - Conference A: Computer Vision Image Processing (IAPR), Vol. 1, 1994, pp. 582–585. doi:10.1109/ICPR.1994.576366.

[25] T. Ojala, M. Pietikäinen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[26] R. Nava, B. Escalante-Ramírez, G. Cristóbal, Texture image retrieval based on log-Gabor features, in: L. Alvarez, M. Mejail, L. Gomez, J. Jacobo (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Vol. 7441 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 414–421. doi:10.1007/978-3-642-33275-3_51.
URL http://dx.doi.org/10.1007/978-3-642-33275-3_51

[27] W. R. Schwartz, F. R. de Siqueira, H. Pedrini, Evaluation of feature descriptors for texture classification, Journal of Electronic Imaging 21 (2) (2012) 023016–1.

[28] M. K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Info. Theory IT-8 (1962) 179–187.

[29] I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Applied Math. XLI (41) (1988) 909–996.

[30] D. Gabor, Theory of communication, J. Inst. Elec. Eng. (London) 93III (1946) 429–457.

[31] S. Fischer, F. Sroubek, L. Perrinet, R. Redondo, G. Cristóbal, Self invertible gabor wavelets, International Journal of Computer Vision 75 (2007) 231–246.

[32] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, I. G. Goldberg, Wnd-charm: Multi-purpose image classification using compound image transforms, Pattern Recognition Letters 29 (11) (2008) 1684–1693.

[33] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.

[34] H. Liu, H. Motoda, Feature selection for knowledge discovery and data mining, Springer, 1998.

[35] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, Wiley, New York, 2001.

[36] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.

[37] P. Mitra, C. Murthy, S. K. Pal, Unsupervised feature selection using feature similarity, IEEE transactions on pattern analysis and machine intelligence 24 (3) (2002) 301–312.

[38] G. McLachlan, Discriminant analysis and statistical pattern recognition, Vol. 544, Wiley. com, 2004.

[39] J. Stevens, Applied multivariate statistics for the social sciences, Mahwah, 2002.

[40] R. E. Walpole, R. H. Myers, S. L. Myers, Probability & statistics for engineers & scientists, Pearson Education, 2006.

[41] C. Cortes, V. Vapnik, Support-vector networks, in: Machine Learning, 1995, pp. 273–297.

[42] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[43] P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson international Edition, Pearson Addison Wesley, 2006.

[44] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine learning 40 (2) (2000) 139–157.

[45] T. K. Ho, The random subspace method for constructing decision forests, Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (8) (1998) 832–844.

[46] R. Caruana, N. Karampatziakis, A. Yessenalina, An empirical evaluation of supervised learning in high dimensions, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 96–103.

[47] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (2) (1947) 153–157.

[48] O. Ronneberger, Q. Wang, H. Burkhardt, Fast and robust segmentation of spherical particles in volumetric data sets from brightfield microscopy, in: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, IEEE, 2008, pp. 372–375.