International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:14, No:2, 2020

Logistic Model Tree and Expectation-Maximization for Pollen Recognition and Grouping

Endrick Barnacin, Jean-Luc Henry, Jack Molinié, Jimmy Nagau, Hélène Delatte, Gérard Lebreton

***Abstract*—**Palynology is a field of interest for many disciplines. It has multiple applications such as chronological dating, climatology, allergy treatment, and even honey characterization. Unfortunately, the analysis of a pollen slide is a complicated and time-consuming task that requires the intervention of experts in the field, which is becoming increasingly rare due to economic and social conditions. So, the automation of this task is a necessity. Pollen slides analysis is mainly a visual process as it is carried out with the naked eye. That is the reason why a primary method to automate palynology is the use of digital image processing. This method presents the lowest cost and has relatively good accuracy in pollen retrieval. In this work, we propose a system combining recognition and grouping of pollen. It consists of using a Logistic Model Tree to classify pollen already known by the proposed system while detecting any unknown species. Then, the unknown pollen species are divided using a cluster-based approach. Success rates for the recognition of known species have been achieved, and automated clustering seems to be a promising approach.

***Keywords*—**Pollen recognition, logistic model tree, expectation-maximization, local binary pattern.

I. Introduction

PALYNOLOGY consists of grouping and then recognizing pollen of the same species in a sample of biological material. This task is arduous and requires hours of work, even for the best experts in the field. Moreover, the result of the analysis is affected by the human factor. Despite that, palynology remains a useful science with many applications. Automation of this practice would reduce time and cost of analysis, but also inter- and intra-palynologist bias between results. This is the reason why many publications have proposed the use of image processing algorithms and tools, which is low cost while conserving reasonable accuracy rate, if implemented correctly. Some of these works have focused on the use of common image descriptors such as those of shape (area, circularity, Hue moments), contours (elliptical Fourier descriptor, Freeman chain code), or textures (Haralick's co-occurrence matrix, Gabor filter). This is particularly the case of the ASTHMA project with the studies of [1]-[4], which have obtained success rates ranging between 77% and 100% for data sets containing 4 to 12 different pollen species. Others such as [5], [6], [7], or [8] have crafted specific pollen descriptors and noticed better recognition rates using their own attributes. Few alternative studies have used less popular supervised learning techniques such as the "paradise neural network" [9], or more sophisticated acquisition methods such as confocal microscopy [10] or scanning electron microscopy [2]. Finally, Daood et al. have studied 30 different pollen species and obtained 94.58% recognition rate using multi-hierarchical classifier [11] and 92.52% [12] using transfer learning and convolutional neural network. Daood et al. have also combined recurrent and convolutional neural networks to recognize sequences of multifocal pollen images acquired by optical microscopy. They obtained a 100% recognition rate [13] for the 10 studied species. In all the previously cited studies, supervised learning has been intensively employed. As a matter of fact, unsupervised learning in the research computer vision field is infrequently used. This is damageable as unsupervised learning could help to group species which are unknown by pollen recognition systems and help to pre-label pollen images which can permit to save time when constructing a new recognition system with non-labeled images. In this paper, we study both supervised and unsupervised learning. In the first, a Logistic Model Tree (LMT) [14] is used to identify whether the pollen studied is known. If it is, the LMT returns the species name of the pollen; if not, the pollen is added to the group of unknown pollen which will later be clustered using an expectation-maximization method [15].

Endrick Barnacin is with the LAMIA laboratory at University of Antilles, Guadeloupe (corresponding author, e-mail: endrick.barnacin@etu.univ-antilles.fr).

Jean-Luc Henry, Jimmy Nagau, and Jack Molinié are with the LAMIA laboratory at University of Antilles (e-mail: jean-luc.henry@univ-antilles.fr, jimmy.nagau@univ-antilles.fr, Jack.Molinie@univ-antilles.fr).

Hélène Delatte and Gérard Lebreton are with CIRAD of Réunion (e-mail: helene.delatte@cirad.fr, gerard. lebreton@ cirad.fr).

II. Pollen Image Acquisition

Several physicochemical treatments were carried out in order to separate the pollen from the various components of the honey. First, 15 g of honey were collected, placed in a 500-mL beaker and diluted in hot distilled water. This is to dilute the sugars, which is the main compound of honey. After that, the method used in the samples is acetolysis [16]. The pellet obtained after acetolysis is stored in a phenol glycerin solution. In order to observe the pollen content of the samples, 50 µl of the preparation is collected and mounted between the slide and the lamella (24*50 mm), using the free mounting method. A qualitative and quantitative analysis was then conducted for each slide. They consist of recognizing pollen types present in the preparation and then counting the pollen encountered for each type. To do this, an optical microscope was used with a magnification x400 (immersion optics x40) coupled to a camera. The count is performed on three arbitrary lines, the first on the upper quarter of the slide, the second in the middle, and the third on the lower quarter. All the pollens on these three lines are photographed to be counted and

identified.

III. Segmentation

In order to well-use Otsu threshold algorithm [17], the RGB color images were converted into HSV images [18]. The algorithm has been applied on the saturation channel to extract the pollen before a hole-filling method. As shown in Fig. 1, the background color is predominant and uniform; moreover, the pollens are clearly distinguishable.

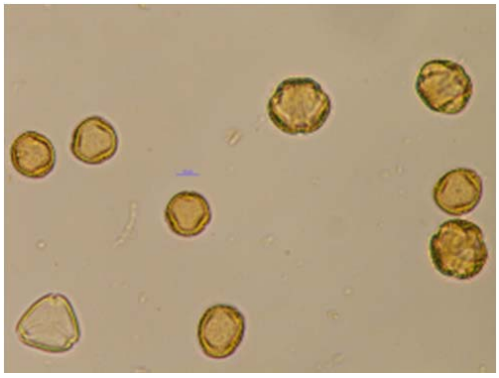


Fig. 1 Honey pollen slide image

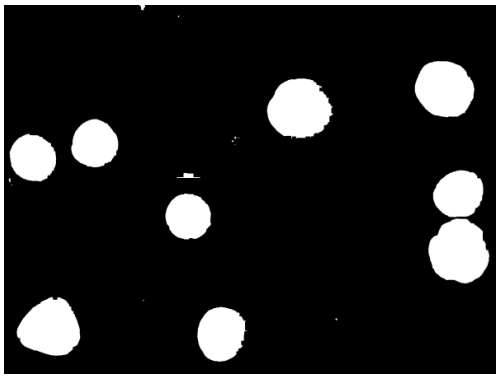
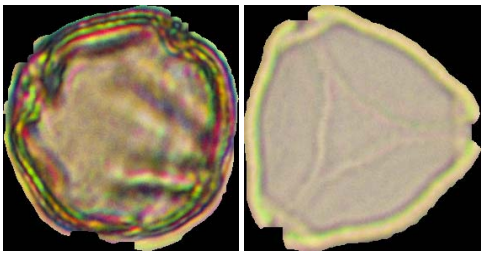


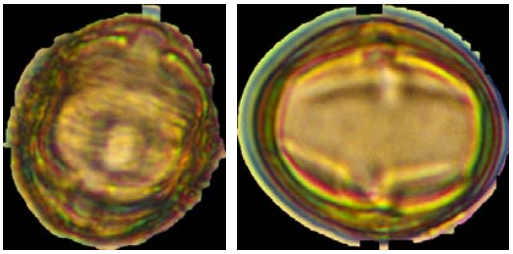
Fig. 2 Honey pollen slide image after Otsu thresolding

IV. Dataset

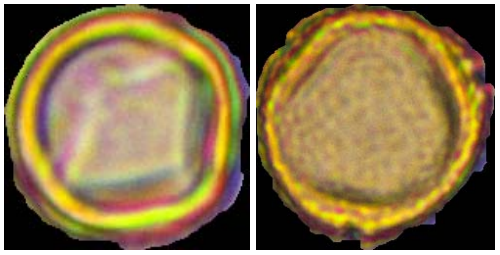
The dataset used in this study is composed of ten different pollen species, shown in Fig. 3. The number of pollen per species contained in the dataset is shown in Table I.



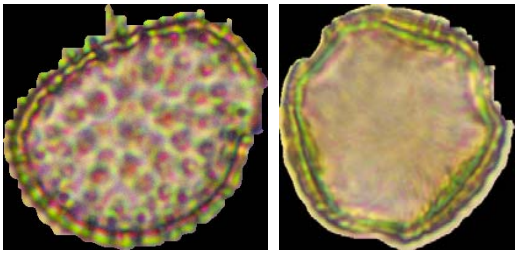
(a) *Anacardiaceae Schinus terebinthifolius* (b) *Myrtaceae Syzigium jambos*



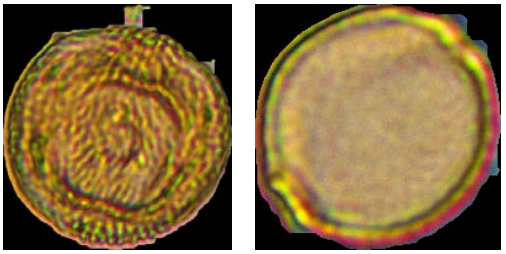
(c) *Aphloiaceae Aphloia theiformis* (d) *Sapotaceae* type Mimusops



(e) *Cunoniaceae Wein mannia tinctoria* (f) *Euphorbiaceae Corde-moya integrifolia*



(g) *Pandanaceae Pandanus* spp. (h) *Sapindaceae Dorato- xylon apetalum*



(i) *Euphorbiaceae Corde moya integrifolia* (j) *Cannabaceae Trema orientalis*

Fig. 3 A sample of each pollen type of our dataset

TABLE I
Quantity of Pollen Per Species

Name	Quantity
Anacardiaceae Schinus terebinthifolius	288
Myrtaceae Syzigium jambos	1162
Aphloiaceae Aphloia theiformis	468
Sapotaceae type Mimusops	63
Cunoniaceae Weinmannia tinctoria	128
Euphorbiaceae Cordemoya integrifolia	752
Pandanaceae Pandanus spp	414
Sapindaceae Doratoxylon apetalum	351
Euphorbiaceae Cordemoya integrifolia	48
Cannabaceae Trema orientalis	87

V. Global Scheme of the System

Fig. 4 shows a general schematic of the system. Pollen images are first segmented, and the LBP feature vector of each pollen is extracted. Then, the LMT method is used for the detection of unknown species. It classifies and considers as unknown pollen those for which the probability of recognition is lower than 0.85. Because experimentally, it is enough to exclude all unknown samples while conserving a recognition rate higher than 95% on the known samples. Unknown pollens are grouped together, the BoVW feature vectors are extracted, and they are clustered using expectation-maximization. Finally, the classes of known pollen are returned by the LMT.

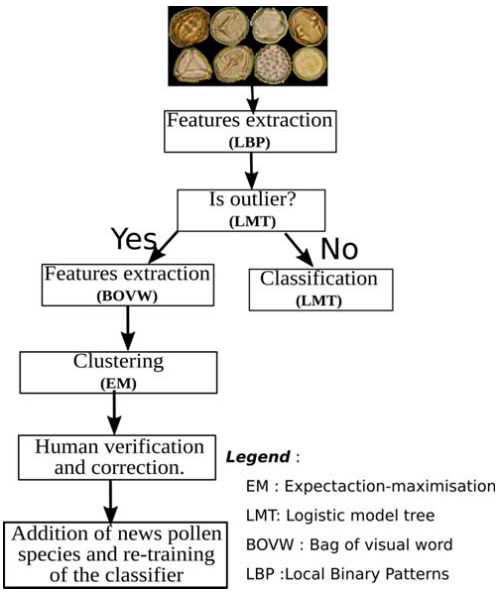


Fig. 4 Global scheme of the system

VI. Feature Extraction

A. Local Binary Pattern

The LBP [19] are constructed as follows:

For each pixel of an image, neighboring pixels P within a circle of radius R are selected. The values of the neighboring pixels P are subtracted from the value of the current pixel. The Heaviside function allows us to keep only the positive values for the calculation of the LBP.

$$LBP_{P,R}(x_c,y_c) = \sum_{p=0}^{p-1} 2^p \delta(g_p - g_c) \quad (1)$$

P and R represent the number of neighboring pixels used in the calculation and the radius of the neighborhood circle, respectively. The "g_c" is the central pixel, "g_p" is the neighboring pixel, and δ is the Heaviside function. From the resulting image, a histogram is formed and used as a feature vector.

In this study, we have chosen to use a 5- and 10-pixel radius with 32 neighbors and a 20-pixel radius with 64 neighbors. Although the radius was tested individually, the best result was obtained with a combination of all of them.

B. Visual Bag-of-Words Using Texture Features

Bag-of-words (BoW) was originally a method destined to text classification which had been extended to computer vision [20]. Finally, Lozano-Vega et al. [5] have used the BoW strategy with LBP in order to detect the apertures of pollen. In this paper, we use it as a feature vector for the expectation-maximization clustering algorithm. The process is as follow:

Firstly, the image is subdivided into patches of 4-pixel height and 4-pixel width. Then, for each patch, the mean and the standard deviation of the hue, and saturation channels is computed. K-means with 15 clusters is applied to extract 15 visual words from the patches.

Finally, the histogram of the occurrence of all the visual words for each image is computed and will serve as a feature vector.

VII. Methodology and Experimentation

We used Weka data-mining software to do the experimentation.

A. Supervised Learning

For each pollen image, the LBP features presented in the previous section was extracted, and a cross-validation classification with k=10, using LMT classifier, was performed.

We have chosen the LMT classifier because this is the classifier which has obtained the best accuracy score experimentally.

B. Unsupervised Learning

Due to the huge computation time required to compute the codebook of visual words and to cluster the pollen, the dataset has been reduced to only forty pollen per species.

Expectation-maximization has been selected because it has obtained the best accuracy score experimentally.

The number of clusters has been determined using a cross-validation method which works as follows:

1. The number of clusters is set to 1
2. The dataset is split randomly into 10 folds
3. EM is performed 10 times using the 10 folds
4. The loglikelihood is averaged over all 10 results
5. If loglikelihood has increased, the number of clusters is increased by 1, and the program continues at step 2.

C. Results

The results obtained for the two learning methods using the dataset presented in Section IV are exposed in Table II:

TABLE II
Classification Rates

Learning method	Supervised (LMT)	Unsupervised (EM)
Recognition Rate	97.21%	77.38%

The proposed method achieved 97.21% classification rate and 77.38% of correctly clustered instances.

VIII. Conclusion and Perspectives

This study focused on the construction of a system capable

of grouping pollen by species with and without supervision. The proposed process consists first of segmenting pollen using the Otsu algorithm, then extracting LBP features to detect unknown species using the LMT classifier. At the end of this first stage, two groups of pollen are obtained: the known species group and the unknown species group. The first is classified using the results obtained previously with the LMT method. The second is divided into sub-groups using the expectation-maximization method with bag of visual words as features. The LMT achieved a 97.21% recognition rate and the expectation-maximization correctly clustered 77% of the samples.

In future work, we will focus on unsupervised learning, and improve the BoVW method by choosing more descriptive feature such as LBP, GLCM, or even Gabor filter to build more descriptive visual words.

REFERENCES

[1] Zhang, Y., D. W. Fountain, R. M. Hodgson, J. R. Flenley, and S. Gunetileke. “Towards Automation of Palynology 3: Pollen Pattern Recognition Using Gabor Transforms and Digital Moments.” Journal of Quaternary Science 19, no. 8 (2004): 763–768. doi:10.1002/jqs.875.

[2] Treloar, W. J., G. E. Taylor, and J. R. Flenley. “Towards Automation of Palynology 1: Analysis of Pollen Shape and Ornamentation Using Simple Geometric Measures, Derived from Scanning Electron Microscope Images.” Journal of Quaternary Science 19, no. 8 (2004): 745–754. doi:10.1002/jqs.871.

[3] Ticay-Rivas, Jaime R., Marcos del Pozo-Baños, Carlos M. Travieso, Jorge Arroyo-Hernández, Santiago T. Pérez, Jesús B. Alonso, and Federico Mora-Mora. “Pollen Classification Based on Geometrical, Descriptors and Colour Features Using Decorrelation Stretching Method.” Artificial Intelligence Applications and Innovations (2011): 342–349. doi:10.1007/978-3-642-23960-1_41.

[4] C. Chudyk, H. Castaneda, R. Léger, I. Yahiaoui and F. Boochs, "Development of an Automatic Pollen Classification System Using Shape, Texture and Aperture Features," *LWA 2015 Workshops: KDML, FGWM, IR, and FGDB*, 2015 .

[5] G. Lozano-Vega, "Image-based detection and classification of allergenic pollen,"2015.

[6] Chen, Chun, Emile A. Hendriks, Robert P. W. Duin, Johan H. C. Reiber, Pieter S. Hiemstra, Letty A. de Weger, and Berend C. Stoel. “Feasibility Study on Automated Recognition of Allergenic Pollen: Grass, Birch and Mugwort.” Aerobiologia 22, no. 4 (December 2006): 275–284. doi:10.1007/s10453-006-9040-0.

[7] Nguyen, Nhat Rich, Matina Donalson-Matasci, and Min C. Shin. “Improving Pollen Classification with Less Training Effort.” 2013 IEEE Workshop on Applications of Computer Vision (WACV) (January 2013). doi:10.1109/wacv.2013.6475049.

[8] Kaya, Yılmaz, S. Mesut Pınar, M. Emre Erez, Mehmet Fidan, and James B. Riding. “Identification of Onopordumpollen Using the Extreme Learning Machine, a Type of Artificial Neural Network.” Palynology 38, no. 1 (January 2, 2014): 129–137. doi:10.1080/09500340.2013.868173.

[9] France, I, A.W.G Duller, G.A.T Duller, and H.F Lamb. “A New Approach to Automated Pollen Analysis.” Quaternary Science Reviews 19, no. 6 (February 2000): 537–546. doi:10.1016/s0277-3791(99)00021-9.

[10] Ronneberger, Olaf, Qing Wang, and Hans Burkhardt. “3D Invariants with High Robustness to Local Deformations for Automated Pollen Recognition.” Pattern Recognition (n.d.): 425–435. doi:10.1007/978-3-540-74936-3_43.

[11] Daood, Amar, Eraldo Ribeiro, and Mark Bush. “Pollen Recognition Using a Multi-Layer Hierarchical Classifier.” 2016 23rd International Conference on Pattern Recognition (ICPR) (December 2016). doi:10.1109/icpr.2016.7900109.

[12] Daood, Amar, Eraldo Ribeiro, and Mark Bush. “Pollen Grain Recognition Using Deep Learning.” Lecture Notes in Computer Science (2016): 321–330. doi:10.1007/978-3-319-50835-1_30.

[13] Daood, Amar, Ribeiro, Eraldo, AND Bush, Mark. "Sequential Recognition of Pollen Grain Z-Stacks by Combining CNN and RNN" Florida Artificial Intelligence Research Society Conference (2018): n. pag. Web. 30 Jul. 2019

[14] Landwehr, Niels, Mark Hall, and Eibe Frank. “Logistic Model Trees.” Lecture Notes in Computer Science (2003): 241–252. doi:10.1007/978-3-540-39857-8_23.

[15] Dempster, A. P., N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data Via the EM Algorithm.” Journal of the Royal Statistical Society: Series B (Methodological) 39, no. 1 (September 1977): 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x.

[16] G. Erdtman, "The acetolysis method, a revised description." *Svensk Botanisk Tidskrift,* vol. 54, (1960).

[17] Otsu, Nobuyuki. “A Threshold Selection Method from Gray-Level Histograms.” IEEE Transactions on Systems, Man, and Cybernetics 9, no. 1 (January 1979): 62–66. doi:10.1109/tsmc.1979.4310076.

[18] Gonzalez, Rafael C., Richard E. Woods, and Barry R. Masters. “Digital Image Processing, Third Edition.” Journal of Biomedical Optics 14, no. 2 (2009): 029901. doi:10.1117/1.3115362, pp. 407–413.

[19] Ojala, T., M. Pietikainen, and D. Harwood. “Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions.” Proceedings of 12th International Conference on Pattern Recognition (n.d.). doi:10.1109/icpr.1994.576366.

[20] Joachims, Thorsten. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” Lecture Notes in Computer Science (1998): 137–142. doi:10.1007/bfb0026683.