Review of Palaeobotany and Palynology 221 (2015) 117–127

Contents lists available at ScienceDirect

Review of Palaeobotany and Palynology

journal homepage: www.elsevier.com/locate/revpalbo

Progress towards establishing collection standards for semi-automated pollen classification in forensic geo-historical location applications

Kimberly C. Riley [a], Jeffrey P. Woodard [a], Grace M. Hwang [a,b], Surangi W. Punyasena [c]

[a] The MITRE Corporation, 7515 Colshire Drive, Mclean, VA, USA
[b] System Planning Corporation, 3601 Wilson Blvd, Arlington, VA 22201, USA
[c] University of Illinois at Urbana-Champaign, 505S. Goodwin Avenue, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 21 December 2014
Received in revised form 28 May 2015
Accepted 10 June 2015
Available online 18 June 2015

Keywords:
Pollen classification
Automated palynology
Computer vision
3D classification
Forensics
Pattern recognition

ABSTRACT

The digitization of pollen grain images would permit the creation of a semi-automated system that could aid the expert palynologists in pollen classification. It would reduce cost and time-to-answer as well as improve analyst productivity. These issues are particularly critical in forensic applications. There are numerous factors that should be considered when establishing a digital database intended for semi-automated pollen classification. This paper explores a number of these issues through computer vision and machine learning assessments. The main topics evaluated are morphologically similar species-level classification, optimal training data size, how best to utilize three-dimensional data, accuracy changes due to the availability of metadata, i.e., fluctuations in analysts' confidence in taxa labeling, and using fossil data to classify modern data. This is the first known application of training on fossil data to classify modern taxa. Performances of 95.4% and 93.8% correct classification were achieved on two distinct sets of morphologically similar species-level data, surpassing previous records. We determined that a minimum of 5–10 training images per class was required to yield reasonable performance. Additionally, we established that all depth dimension slices associated with each grain were required to yield the best performance possible. Lastly, the error rate doubles due to decreasing analyst confidence and almost triples when using data from grains of varying ages, further solidifying the importance of comprehensive metadata.

1. Introduction and review

The association of goods or people to place-of-origin is a term broadly referred to in forensics as geographic attribution or simply geolocation. However, here we will use the term geo-historical location to differentiate it from real-time tracking. For years, palynologists have studied pollen grains to infer information for geo-historical location applications (Hwang and Masters, 2013), including fair trade (Bryant and Jones, 2006), validating history (Bertino, 2012), allergy research (Scharring et al., 2006), agriculture (Jones and Bryant, 1992), dating rocks for petroleum (Cross, 1964), mining (Gonzalez et al., 2002) and coal analysis (Nichols and Jacobson, 2005). Pollen is a useful tool in the forensic domain because it has the potential to provide substantial amounts of information about an item's age (Bryant and Mildenhall, 1998; Gonzalez et al., 2002), provenance (Stoney et al., 2011) and travel path (Korejwo et al., 2007). Additionally, pollen is resilient to damage, so has a high preservation potential (Nichols and Jacobson, 2005; Traverse, 2008).

Current pollen classification methods rely heavily on skilled expert palynologists. Given the limited number of these experts, classification can be a long and costly process. Past studies have shown that palynologists' opinions can be subjective (Allen et al., 2008) and that their knowledge can be localized to specific world regions. We hypothesize that the creation of a semi-automated system via computer vision could provide a capability that will allow more data to be analyzed in a fraction of the time.

In order to build a successful system, careful consideration needs to be given to what makes a strong database. Currently, there is no unified global database of pollen types that accounts for both morphological attributes, as well as geo-historical location information. Limited data sources may cause an expert palynologist to use multiple data sources as a basis of comparison. When combining multiple data sources, an understanding is required as to which metadata fields contain parameters that could degrade performance in pollen grain classification (and associated geo-historical location predictions). In the future, metadata could allow the expert to assign a level of confidence to the classification result. The metadata parameters explored in this study were pollen age and analysts' confidence in their own identification of training samples. Additionally, understanding the potential and limits of automation in the pollen domain is critical.

Routine analysis typically identifies pollen grains at the genus level and rarely classifies at the much finer species level (Mander and Punyasena, 2014). Although species-level classification can be a

E-mail address: kriley@mitre.org (K.C. Riley).

http://dx.doi.org/10.1016/j.revpalbo.2015.06.005
0034-6667/© 2015 Elsevier B.V. All rights reserved.

challenge for even a seasoned palynologist, the geographic information that can be recognized with species-level classification provides much greater spatial accuracy and precision compared to genus-level classification (Hwang et al., 2013a). Additionally, analysts use modern pollen specimens to classify fossil data. Modern specimens typically originate from herbarium sheets or from vouchers collected from plants in the field. Therefore, the metadata (i.e., labels) of these data provide ground truth. Fossil data are normally extracted from a core sample, while in forensics, data may be extracted from samples from an article of clothing or a package. The labels associated with these data are opinions based on the knowledge founded from modern data samples. Given the potential for limitations in the representation of taxa within training, tests were also performed using fossil data to classify modern data. These tests examine whether it is possible to use high confidence fossil data to classify unknown pollen grains.

With this in mind, we performed assessments on morphologically similar data from the *Pinaceae* family (see Fig. 1 top) at both genus and species levels. Rodriguez-Damian (Rodriguez-Damian et al., 2004, 2006) also performed similar studies on morphologically similar species, focusing on the *Urticaceae* family. Furthermore, gaining an understanding of training data size on performance is critical when establishing a database. It is well accepted that classifying morphologically similar species is one of the most difficult problems in pollen analysis and the 3-D nature of the pollen grain further increases this complexity (Mander and Punyasena, 2014). Nguyen et al. (2013) addressed this issue by counting grain surface spikes on the pollen grain at various angles. Boucher et al. (2012) and Allen et al. (2006) had vast representations of each taxon, which they used to observe the effects of accuracy given decreasing training representations. Ronneberger et al. (2002) used confocal microscopy to perform 3-D reconstructions. Similar to Boucher et al. (2012) and Allen et al. (2006), we have addressed this issue by representing pollen grains at abundant viewpoint angles in the training data followed by observing the effects of decreasing training size. Each pollen grain was represented by a series of 23–84 axial images, which we will refer to as an image stack (see Fig. 1 bottom). More salient features may be prevalent in specific regions of the image stack. If utilizing specific regions yields comparable or superior results to using the entire image stack, this discovery would be extremely beneficial when considering memory storage limitations.

This paper will begin by discussing a high-level overview of how we believe our proposed semi-automate system would function (Section 2). Section 3 describes the data that were used for our study as well as our methodology. We provide details on the computer vision methods applied as well as a range of classifiers that were explored. Section 4 displays the results of our various studies followed by Section 5, which provides further discussion of these results. Lastly, Section 6 assesses these studies, providing insight on how these studies answer our hypothesis, and concludes with recommendations for future tests.

2. Description of system

One intended application for pollen classification is for forensic geohistorical location, using pollen to determine where an item originated. Fig. 2 gives a high-level overview of how we visualize this system functioning. When new images are introduced into the system, the classifier (step one) determines matches for this new image based upon its previous knowledge provided by the database. Once the pollen grains have been classified, a probabilistic distribution model is created (step 2) by utilizing occurrence data from plant or pollen databases. This model estimates the possible regions of item origination and produces associated maps (see yellow in step 2). As the effects of collection parameters and morphological parameters are better understood, they should be incorporated into a much more detailed diagram.

3. Materials and methods

Two datasets, consisting of modern and fossil grains, were used for our studies. All image data were collected with a Zeiss Apotome fluorescence microscope. The modern dataset contained 641 grains of which 442 were from spruce (*Picea mariana*, *Picea glauca*, and *Picea rubens*), 96 were from fir (*Abies balsamea*), and 103 were from pine (*Pinus banksiana*, *Pinus strobus*, *Pinus resinosa*, and *Pinus rigida*). The fossil dataset contained 264 grains of which 103 were from *Picea mariana* and

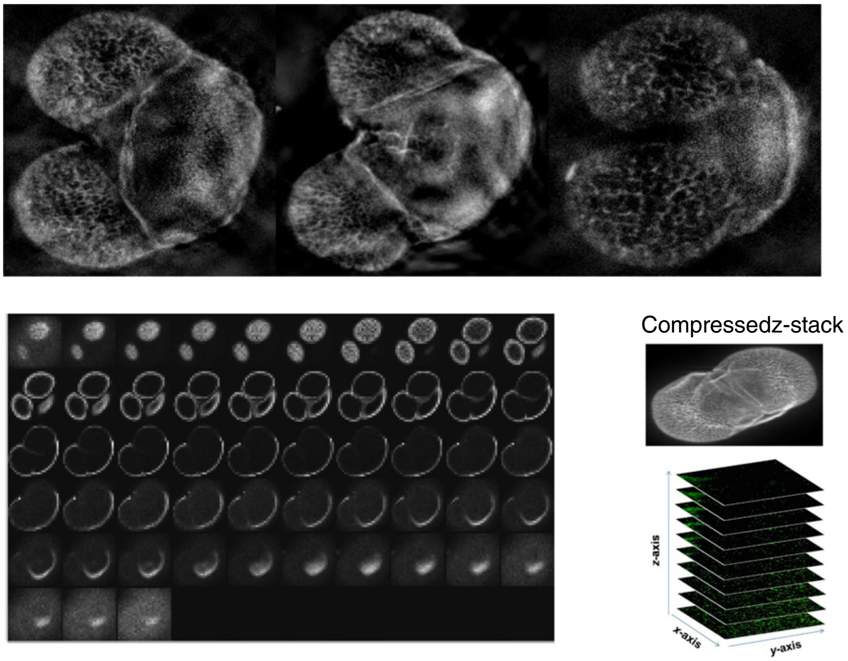

Fig. 1. (Top) Morphologically similar genus-level pollen grains of the *Pinaceae* family (from left to right: *Abies*, *Picea*, *Pinus*). Image data from Punyasena et al. (2012). (Bottom) Image stack (left image) of one grain along with its summed 2D representation (top right). Image stack representation (bottom right) taken from (http://bioimagel.com/yahoo_site_admin/assets/images/stack.333130145.jpg).