Collection Standards for Semi-Automated Pollen Classification in Forensic Geo-Historical Location Applications

Kimberly C. Riley, Jeffrey P. Woodard, Grace M. Hwang
The MITRE Corporation
7515 Colshire Drive, Mclean, VA, USA
kriley@mitre.org

Surangi W. Punyasena
University of Illinois at Urbana-Champaign,
505 S. Goodwin Avenue, Urbana, IL, 61801, USA

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 labelling, and using fossil data to classify modern data. This is the first known application of training on fossil data to classify modern taxa. Performance 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.*

Index Terms—*computer vision, 3D classification, automation, pollen identification, forensics, feature vector, SIFT, LBP, Hessian-Affine, GIST, pattern recognition, bioinformatics, geolocation*

1. INTRODUCTION

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 that can be applied to geo-historical location locating

applications [1] such as fair trade [2], validating history [3], allergy research [4], agriculture [5], dating rocks for petroleum [6], mining [7] and coal analysis [8]**.** Pollen is a useful tool in the forensics domain because it has the potential to provide substantial amounts of information about an item's age [7,9], provenance [10] and travel path [11]. Additionally, pollen is resilient to damage [8,12].

Current classification methods rely heavily on skilled expert palynologists. Given the limited number of these experts in the world, classification can be a long and costly process. Past studies have shown that palynologists' opinions can be subjective [13] 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 analysed 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 considering 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. Although species-level classification can be a 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 [14]. Additionally, analysts use modern data to classify fossil data. Modern data typically originate from herbarium sheets or directly 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 figure 1 top) at both genus and species levels. Rodriguez-Damian [15,16] 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. Nguyen et al. [17] addressed this issue by counting grain surface spikes on the pollen grain at various angles; in contrast, Allen [18] and Boucher [19] had vast representations of each taxon, which they used to observe the effects of accuracy given decreasing training representations. Ronneberger et al. [20] address this issue by using expensive confocal-microscopy hardware to perform 3-D reconstruction. Similar to Allen and Boucher, we have addressed this issue by representing pollen grains at abundant viewpoint angles in the training data then observing the effects of decreasing training size. Each pollen grain was represented by 23-84 images, which we will refer to as an image stack (see figure 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.

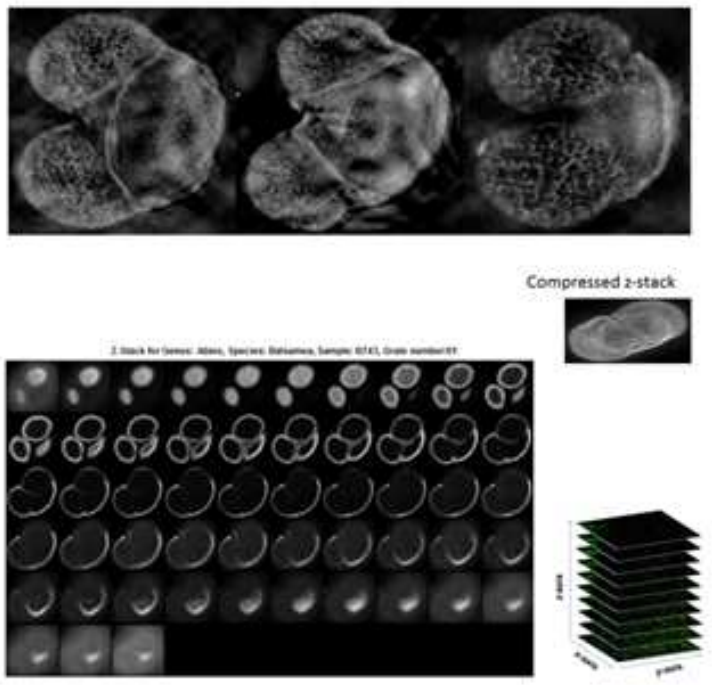

Figure 1.(Top) Morphologically similar genus-level pollen grains of the *Pinaceae* family (from left to right: *Abies*, *Picea*, *Pinus*). Image data from [21]. (Bottom) Image stack (left image) of one grain along with its summed 2D representation (top right). Image stack representation (bottom right) taken from [22].

This paper will begin by discussing a high-level overview of how we believe our proposed semi-automated system would function (section 2). Section 3 describes the data that were used for our study. Section 4 defines our methodology, providing

details on the computer vision methods applied as well a range of classifiers that were explored. Section 5 displays the results of our various studies followed by section 6, which provides further discussion on these results. Lastly, section 7 assesses these studies, providing insight on how these studies answer our hypothesis, and concludes with recommendations for future tests.

2. SYSTEM OVERVIEW

One intended application for pollen classification is for forensic geo-historical location, using pollen to determine where an item originated. Figure 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.

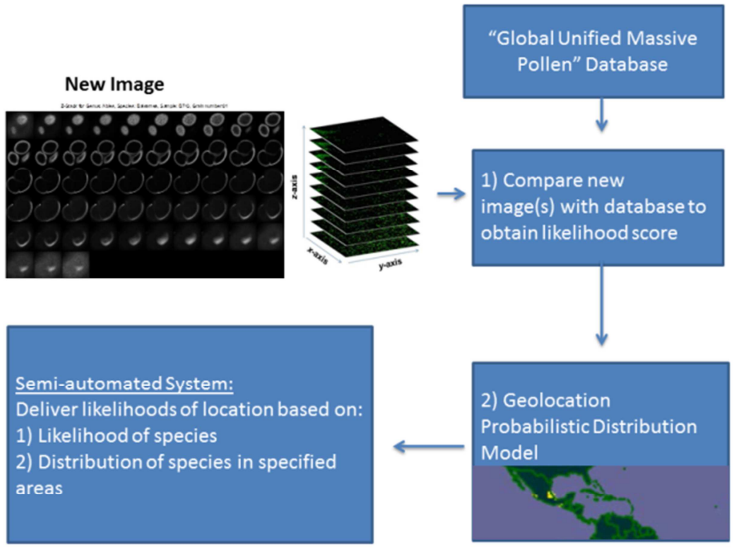

Figure 2. Flow diagram of semi-automated system. A new pollen image stack is introduced to the system. This image is then compared to a set of database images for classification. The estimated class is then used to create a geo-historical location probabilistic distribution model.

3. DATA DESCRIPTION

Two datasets, consisting of modern and fossil grains, were used for our studies. All data were collected with a light microscope. The modern dataset contained 641 grains of which 442 were from spruce (*Picea mariana*, *P. glauca*, and *P. rubens*), 96 were from fir (*Abies balsamea*), and 103 were from pine (*Pinus banksiana*, *P. strobus*, *P. resinosa*, and *P. rigida*). The fossil dataset contained 264 grains of which 103 were from *Picea mariana* and 161 were from *Picea glauca*. Because we used morphologically similar grains, challenging even for expert palynologists to classify, all fossil-image stacks were accompanied by metadata that were labelled with analysts' confidence level (CL) based subjective, self-reported confidence in their judgements. These CL's for modern and fossil samples were 99% and ⩾50%, respectively. Although CL's range from 50% to 99%, our analyses were performed on two subsets of the data: grains with CL ⩾95% and grains with CL between 70 and 95%. It is important to note that lower confidence data tends to be morphologically ambiguous. Classifying these grains can be fairy challenging for both an analyst as well as a computer. Table 1 provides an overview of the data used for this study.

Data Category	Number of classes	Number of grains per class
Modern-Genus	3	96, 103, 442
Modern-Species	5	48, 96, 102, 192, 201
Fossil-Confidence≥70%	2	307, 374
Fossil-Confidence≥95%	2	108, 152

Table 1. Overview of data characteristics

4. METHOD

We describe automation by computer vision algorithms, which have gained popularity in a wide variety of image applications by achieving good performance, sometimes with minimal parameter tuning. Similar methods for automatically classifying pollen grains have been developed by Filipovych [23] and Dahme [24]. One of the challenges of computer vision is finding a way to describe images in a simple yet optimal way. This is done by using feature vectors. A feature vector is an n-dimensional vector of numerical values that represent an image. By reducing the dimensionality of the image, processing becomes easier. There are two types of feature vectors explored in this study: global and local. Global features describe an image from a holistic point of view while local features are selective to characteristics within an image it finds to be relevant. After an image has been converted into a representative vector, it is then ready for the classification. A range of classifiers are explored in this study and further discussed in 4.3.3. A number of researchers have explored nearest neighbor classifiers, SVMs (support vector machine) and decision trees for pollen classification [15, 25, 26]. Our study explored all of these methods while also comparing a range of nearest neighbor alterations. Exploring a variety of nearest neighbor classifiers is believed to be novel.

4.1 Global Feature Vector Description

As stated above, global feature vectors describe an entire image. They provide compact representations of the texture and to a lesser degree spatial shapes in an image. One of the main advantages of these global features is that they are computationally simple yielding a low computational cost. The disadvantages are that they do not typically perform well on images with objects in the presence of clutter or occlusion. The data used in this study was segmented from the background and had neither clutter nor occlusion. Two global feature extractors were used: LBP-HF (Local Binary Pattern Histogram Features) [27] and GIST [28].

Local Binary Pattern: LBP is a simple yet effective texture descriptor. Given that this algorithm accounts for both local spatial patterns as well as gray-scale contrast, it can recognize similar patterns despite variations in luminescence. The LBP-HF transformation is described in figure 4. We see the original image as the first image in the flow diagram. We then transform this image into a higher contrast texture map of the original image, which results in the second image in the flow diagram. We then take this texture map and perform a 2D FFT (Fast Fourier Transform) on this high contrast image, resulting in the third image of the flow diagram. Lastly, the resultant pixel intensity values from the FFT image are binned into a histogram (see the fourth image). The counts of each bin comprise the final texture descriptor.

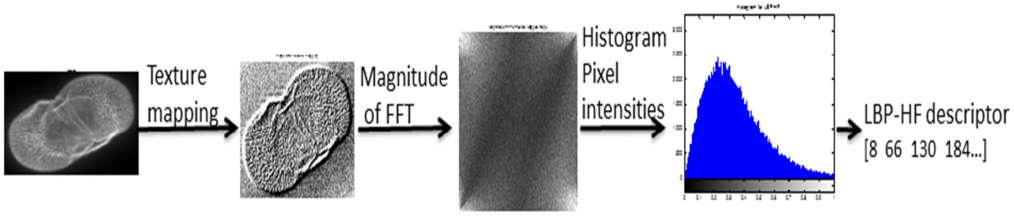

Figure 4: Flow diagram of LBP-HF (Local Binary Pattern- Histogram Fourier Features)

Taking the magnitude of the FFT of the texture mapped image is advantageous because this makes our descriptor invariant to rotation. Because the phase of the FFT is discarded, the features are locally rotation invariant while keeping their highly discriminative attribute.

GIST: GIST captures the "gist" or overall spatial envelope of the scene into a low-dimensional signature vector. This algorithm uses wavelet scale-space image decomposition to compute its texture features. The main advantages of GIST are its computational speed and it low-dimensionality. It does so while preserving perceptually relevant spatial information.

4.2 Local Feature Vector Description

A local feature vector represents only a small part of an image, and is constructed independently of other portions of the image. In contrast to the global case, initially an image has many local features extracted from it. For local features, an unsupervised clustering step is required that is also discussed. The main advantages of local features are that they typically perform well with occlusion and clutter [30]. They may also provide increased discrimination power over global features.

SIFT (Scale Invariant Feature Transform): SIFT [30, 31] generates local features that are robust to challenges such as changes in rotation, frequency-scale and illumination intensities. SIFT operates in two steps: first, the feature is detected, and then it is described. A SIFT feature is detected as a local extremum in a Difference of Gaussian (DOG) space, which is an estimate of the Laplacian of Gaussian (LOG) scale-space. The Laplacian of Gaussian space is beneficial due to its scale invariance properties [30]. Figure 5 depicts an example of SIFT features detected on a single z-slice of a pollen grain image stack. The center of each circle defines the area where the feature was detected while the radius of each circle defines the scale. The arrows inside of each circle define the dominant orientations. SIFT detectors are typically found in regions of high contrast, such as edges and corners.

Figure 4 (top) demonstrates where the features are detected. After areas of interest are established, the algorithm describes these images in a vector representation. The magnitudes of the image gradient are sampled at various normalized orientations in a local region. An array of 16 histograms, each quantized to 8 orientation bins, is then used to create a 128 (16*8) element vector. Typically, many thousands of these 128 vectors are created for each image. Utilizing all of these vectors would be both memory intensive and time consuming. To address this issue, a vector quantization, or VQ, step is performed. In this step, we quantize by comparing each input vector with a precomputed quantizer value, or codeword, determining its closest match in 128-dimensional space. The established set of codewords, or our codebook, is derived from an external dataset that has similar morphological features to the input dataset. If the morphological traits of the external codebook and the testing dataset do not overlap, the vectors will not quantize properly and the resultant feature vectors will not be created properly. This will result in an inaccurate feature vector representation and a decrease in accuracy. In the case of the modern dataset, a mangrove dataset was utilized to establish external codewords. The fossil dataset proved advantageous because it allowed us to allocate subsets of this data, based on analysts' confidences, for testing and training purposes. Since tests were performed on data where analysts' confidence was $\geq$95% and $\geq$70, the remainder of the data could be used to train our VQ codebook. The counts of how often each input vector matches each VQ codeword define final SIFT feature vectors. It is important to note that the locations of where salient features occur are discarded and not used to define the final feature vectors. A visual representation of these codewords can be seen in figure 4 (bottom).

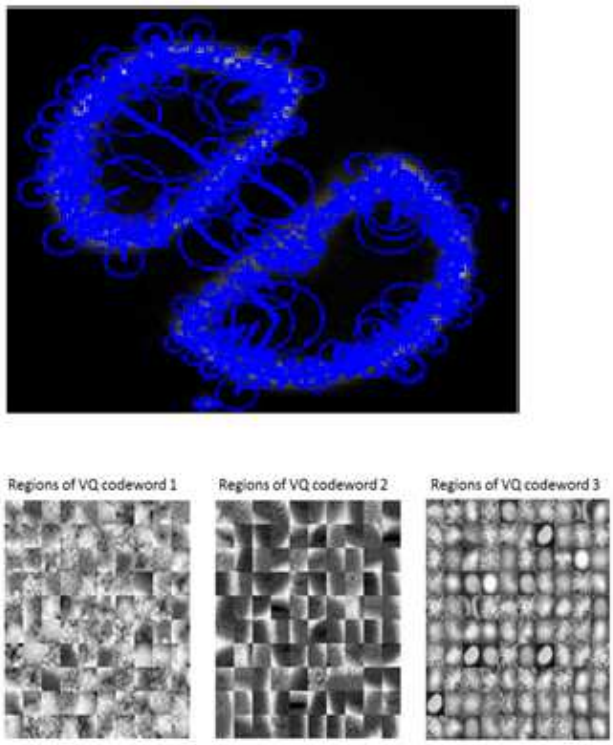

Figure 4: (Top) Overlay of SIFT regions of one z-slice of a pollen grain image stack. (Bottom) Visual representation of how SIFT bins its features.

One interesting observation of these codewords is that numerically close SIFT vectors tend to represent areas of the image that contain visually similar information. Codeword 1 (bin1) appears to describe coral-like shapes while codeword 2 (bin 2) appears to be describing corners. Codeword 3 (bin 3) describes circles and ovals well. It's important to note that these codewords have been trained without human interaction. In most studies, SIFT performs well by distinguishing images based on the relative occurrences of shapes or features. This histogram of clustered vectors is what defines the final local feature vector.

Hessian Affine SIFT: Hessian Affine SIFT [31] has many of the same components of SIFT. The main difference lies with the detection part of the algorithm, where a scale space determines the local maximum and minimum. While SIFT detects local extrema in DOG scale-space, Hessian-Affine SIFT detects local extema in Hessian-Laplacian scale-space. The Hessian space proves beneficial in detecting blob-like structures [32]. Once an extremum is determined, the algorithm assumes the region of interest is elliptical in shape (whereas SIFT assumes the area of interest is circular).

4.3 Classifier Overview

One factor that makes the classification of pollen images so unique and challenging is the three-dimensional aspect. Recall that each test grain is represented with multiple two-dimensional images, where each represents a slice in the third dimension. These image stacks can be conceivably used in a variety of classification strategies. The particular strategy for utilizing the three-

dimensional image data has a significant effect on the classification accuracy. However, the foundation of all such strategies is a conventional machine learning algorithm. Therefore, we briefly review these before moving on to the three-dimensional issues. In computer vision and related data mining applications, a wide variety of supervised machine learning algorithms or classifiers have been explored over the years (e.g., see Andreopoulos and Tsotsos [32]). Especially popular are those based on k-nearest neighbor (K-NN) and support vector machines (SVM). In addition to these two, we also report the performance of several other well-known supervised classifiers: linear discriminant analysis, quadratic discriminant analysis, and decision trees. All of these can be considered as discriminant classifiers (Hastie, et al.33]), as opposed to generative classifiers such as PLSA (Probablistic Latent Semantic Analysis). The discriminant classifiers estimate/learn either a conditional probability distribution of the class label given the observation (image) for the parametric classifiers or a nonlinear classification mapping function for the nonparametric cases. Because all classifiers explored here are well known, we provide brief overviews and refer the reader to standard references. For K-NN classifiers we provide more detail because of the relative ease of incorporating them into more complicated classification strategies involving the three-dimensional image stacks.

SVMs are ubiquitous in machine learning applications and especially in computer vision. These generalized linear methods attempt to separate classes on the basis of hyperplanes, even for the situation where the classes may overlap and are not separable by a linear boundary. We use the public domain implementation (Chang and Lin [34]), where the radial basis function is selected to allow possible non-linear class overlaps. SVMs can be considered generalizations of linear discriminant analysis. The latter relies on a parametric multi-variable Gaussian framework with shared covariance. Quadratic classifiers relax this shared covariance assumption. For implementations of both the linear discriminant and quadratic classifiers we used built-in Matlab functions available in the Statistical toolbox [35]. A decision tree is an example of a graphical classifier, with the trees' nodes and branches partitioning the feature space into separate rectangular regions. A different non-parametric classification method is tailored to each region. We use the recursive binary tree classifier available in the Matlab Statistical toolbox.

K-NN methods are special cases of instance-based methods. Assume for now that each image is represented by one feature vector. Let a test feature vector X exist in a Q-dimensional feature space, so that $X \in R^Q$. This feature vector is compared to N training feature vectors Y_i, i =1:N, $Y \epsilon R^Q$ stored in a library. The comparison is done by some selected distance measure, $D(X,Y_i)$. Each feature vector in the training library has an associated class label, l_j, $j = 1, 2, \ldots, N$. Each class label l is one of a set of V classes, so $l \in \{1, 2, \ldots, V\}$. The nearest neighbor rule is to estimate for a query the closest k training vectors $Y_1^*, Y_2^* \ldots, Y_k^*$, where here Y_j^* represents the j_{th} closest training vector, with associated class labels l_1, l_2, …,l_k. The final class decision is just the majority vote or mode of these K labels. The bias and the variance of the classification error can be traded

off by adjusting the value of K in the K-NN classifier. When K = 1 the bias is lowest but variance is highest. Typical choices for D(·) include those based on the Minkowski norm:

$$D_M(X,Y) \equiv \sqrt[p]{\sum_{j=1}^{Q} \left| x_j - y_j \right|^p} \quad ,$$

where x_j and y_j represent the values of X and Y along the j_{th} dimension, respectively. Familiar special case distances are for p=1 (cityblock), p=2 (Euclidean) , and p = ∞ (Chebyshev). All studies were performed using a grain-fold leave one out cross validation. During training all images from the test grain were left out; the images of the remaining grains were used as training images. In addition, the effect of accuracy on training was explored by using fossil data to train and classify modern data, and conversely.

4.4 Three-Dimensional Nearest Neighbor Classification Strategies

A number of nearest neighbor classification methods were explored in this study to accommodate the three-dimensional (3D) nature of the data. All classification methods were performed on the feature vector representations of the images (see figure 5). As a new pollen grain (represented by a stack of 2D images) is introduced to the system, it is first converted into a feature vector. This feature vector is then compared to feature vectors in the training database resulting in an estimate of the taxon classification. Moreover, when discussing our range of classifiers, our references to both training and testing data used for classification are feature vector representations of these data not the 2D images.

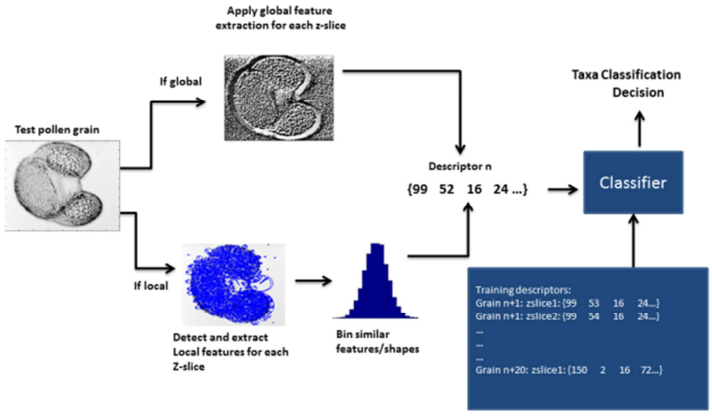

Figure 5. Overview of algorithm: A new pollen grain (represented by an image stack) is introduced to the system. It's then converted to a feature vector representation (also called a descriptor). Lastly, it is compared to training feature vectors in order to estimate its classification value. Figure is modified from [36].

Mode of mode

For each z-slice of each test image, we determined the closest matching z-slices among all z-slices of all training images. The label for a testing z-slice was estimated to be the mode of the true labels of its closest K-NN training z-slices, e.g. by conventional K-NN classification. Once each testing z-slice was assigned an estimated label, we then calculated the mode of all the testing z-slice labels of the test grain: this label was the final label estimate of the testing grain, as shown in figure 6.

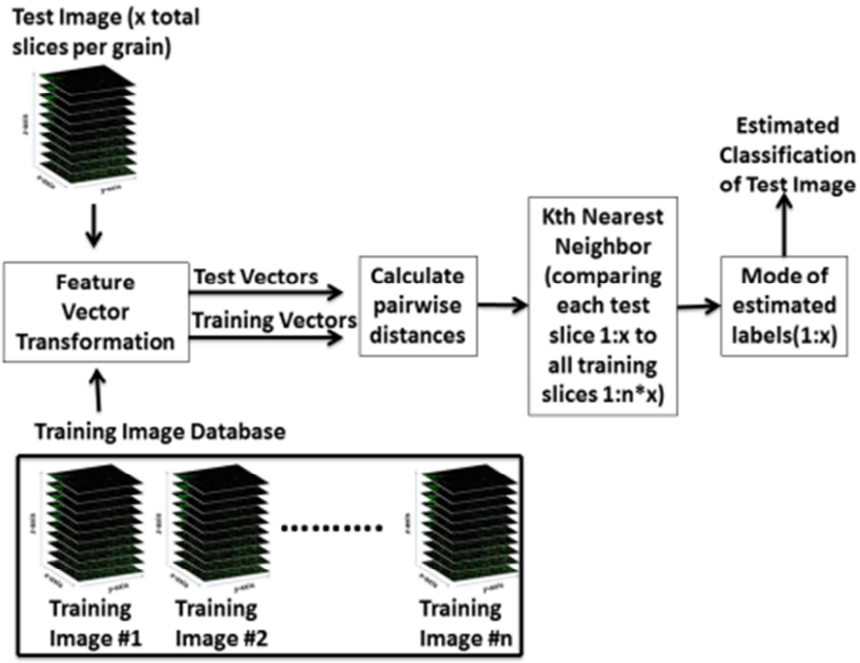

Figure 6. Flow diagram of 3D nearest neighbor classification. Image stack representation taken from [22].

Mode of Mode of Smallest Distances

This method modifies the previous technique by taking into account pair-wise distances. Specifically, it uses the mode of the labels related to the lowest distance values. This method was performed for the lowest 10% and 50%. For example, if a test grain has 50 z-slices and therefore 50 label estimates, only the labels related to the smallest 5 (or 25 for the 50% method) distances were utilized.

Mean of Class

When a test grain is compared with a training set of grains, we average the distances between that test grain z-slice versus all training z-slices related to a particular class. Each test z-slice is then assigned a label based on the smallest mean. This is essentially a kernel method of classification. After the test grain has been assigned estimated labels (which should equal the

length of the number of z-slices in the testing grain), the mode of these label estimates is then computed, yielding a label estimate for the test grain itself.

Tapered Mean of Class[1]

This method is an alteration of the Mean of Class method. Instead of comparing a test z-slice to all training images of each class, we compared the test z-slice to the top 10% and top 50% of each class based on lowest distance values.

Nearest neighbor (grain-level, each grain gets a value)[1]

This method computes one distance between the test grain and each training grain. Conventional K-NN then proceeds from here.

Summing Z-slices

This method sums or averages all z-slices so that we are left with only one image per grain (see figure 2, top right), followed by conventional K-NN classification.

Slide-level bias study

This method insures that data from the testing set are not taken from the same slide as data from the training set. A mode of mode nearest neighbor approach was used for this assessment.

4.5 Other Experiments

Aside from classification experiments we also performed the following studies.

Location vs. Accuracy

Tests were performed in order to determine whether the 3-D pollen imagery had specific "sweet spots" that consistently yielded high levels of accuracy. Note that the number of z-slices per grain was not consistent across the dataset, which prevented a one-to-one comparison based on the location of the slice within the image stack. Therefore, we averaged our compiled image stack error rates into 20 bins to gain a relative regional understanding of which areas of the image stack typically result in higher error rates.

Additionally, to better understand why these errors may be heightened at specific areas of the image stack, we created an image montage which displayed visual representations of each designated bin. To compute these images, the images within each bin were summed then divided by the number of images per bin. For simplicity, we observed image stacks that contained 60 total z-slice images, allowing us to allocate 3 images to each bin. An example of this montage can be seen at the bottom of figure 7.

[1] These methods yielded universally poor results and therefore are not further discussed in this paper.

Determining Training Size

A reverse k-fold method was used to understand the effects of decreasing training size on accuracy. As is common in machine learning, stratified k-fold cross-validations ration the data into k subsets [37].One k-subset is used for testing while the remaining subsets are used for training. This process is performed k-times so that each subset is used once for testing and k-1 times for training. For reverse k-fold experiments, as the value of k increased, the size of the training set decreased while the size of the testing set increased. Recall that the number of images per class within each dataset was not consistent. Considering this issue, the k-values ranged from 1 to the smallest class representation size per test, ensuring that each class had at least one potential match in the training set. One test was performed for each k-fold analysis.

5. RESULTS

5.1 Overview of Accuracy Assessment

Table 2 is an overview comparing our current and past error rates.

Definition of Analysis	% Error (this study)	% Error from prior study [21]
Modern data/Genus-level Analysis	1.7	4.8
Modern data/Species-level Analysis	6.2	6.7
Fossil data/Species-level/Confidence≥95%	4.6	5.8
Fossil data/Species-level/Confidence≥70%	10.9	22.5
Fossil data/Confidence≥95% *slide-level leave one out	4.6	6.2
Classifying modern data using fossil data: query study	13.7	Not performed
Classifying fossil data using modern data: query study	12.1	Not performed

Table 2: Summary of best accuracy

5.2 Comparing nearest neighbor methods

Table 3 compares the error rates for a variety of nearest neighbor methods for the modern dataset at the genus level:

Percent Error for Modern Data: Genus-level Study				
	Mode of K-NN (K=1)	Mean of Class	Mode of Smallest distance (10%)	Mode of smallest distance (50%)
LBP	1.7	54.3	10.1	3.1
GIST	4.5	38.8	11.5	7.0
SIFT	12.3	83.8	24.0	11.0
Hessian-Affine SIFT	11.9	46.7	37.8	18.7

Table 3: Modern data: Genus-level study

Table 4 compares the error rates for various nearest neighbor methods for the modern dataset at the species-level:

Percent Error for Modern Data: Species-level Study				
	Mode of K-NN (k=1)	Mean of Class	Mode of Smallest distance (10%)	Mode of smallest distance (50%)
LBP	6.2	63	36	12.5
GIST	15.8	67.4	24.3	19.3
SIFT	13.9	70.7	41.8	22
Hessian-Affine SIFT	25.3	62.1	54.8	40.4

Table 4: Modern data: Species-level study

Table 5 compares the error rates for a variety of nearest neighbor methods for the fossil dataset when analysts' confidence is greater than or equal to 95%. For all fossil level studies, only species-level analysis was performed.

Percent Error for Fossil Data with Analysts' Confidence ≥95%				
	Mode of K-NN (K=1)	Mean of Class	Mode of Smallest distance (10%)	Mode of smallest distance (50%)
LBP	7.6	36	9.85	9.1
GIST	15.2	34.1	20.8	18.6
SIFT	4.6	55.7	12.9	8.7
Hessian-Affine SIFT	8.3	17.8	23.5	9.1

Table 5: Fossil data with analysts' confidence ≥95%

Table 6 compares the error rates for a variety of nearest neighbor methods for the fossil dataset when analysts' confidence is greater or equal to 70%.

Percent Error for Fossil Data with Analysts' Confidence ≥70%				
	Mode of K-NN (K=1)	Mean of Class	Mode of Smallest distance (10%)	Mode of smallest distance (50%)
LBP	14.6	44.7	17.8	14.0
GIST	24.3	45.1	30.0	26.8
SIFT	10.9	54.0	21.0	15.0
Hessian-Affine SIFT	17.5	20.2	38.0	20.1

Table 6: Fossil data with analysts' confidence ≥70%

5.3 Comparing Our Best Nearest-Neighbor Classifier to Non-NN Classifiers

Table 7 provides an overview of performance of our best K-NN classifier to other classifiers for genus-level evaluations. For tables 7 through 10, the computer vision methods that yielded the best K-NN results were used for these classifiers. For table 7, LBP vectors were utilized.

Classification Method	% Error
Mode of mode K-NN	1.7
SVM-radial	4.3
Pseudo Linear Discriminant Analysis	4.3
Pseudo Quadratic Discriminant Analysis	5.8
Decision Tree	5.8

Table7: Modern-genus comparison

Table 8 provides error rates for our modern species-level evaluations. As with the modern genus-level study, LBP was utilized for each classifier.

Classification Method	% Error
Mode of mode K-NN	6.2
SVM-radial	19.7
Pseudo Linear Discriminant Analysis	16.0
Pseudo Quadratic Discriminant Analysis	16.0
Decision Tree	13.4

Table 8: Modern-species comparison

Table 9 provides error rates for the fossil species-level evaluation using SIFT vectors. These studies were performed on data where analysts' confidence was greater than or equal to 95%

Classification Method	% Error
Mode of mode K-NN	4.6
SVM-radial	10.0
Pseudo Linear Discriminant Analysis	7.7
Pseudo Quadratic Discriminant Analysis	8.9
Decision Tree	8.5

Table 9: Fossil confidence ≥95%

Table 10 provides error rates for the fossil species-level evaluation using SIFT vectors. These studies were performed on data where analysts' confidence was greater than or equal to 70%

Classification Method	% Error
Mode of mode K-NN	10.9
SVM-radial	16.5
Pseudo Linear Discriminant Analysis	15.4
Pseudo Quadratic Discriminant Analysis	18.6
Decision Tree	14.0

Table 10: Fossil confidence ≥70%

5.4 Comparing error rates when analyzing image stacks to 2D rendered images

Table 11 compares error rates when utilizing the image stack, a set of 2D image slices, versus compressing the image stack by summing all images to yield a single 2D representation of the stack.

Feature Vector Method	Image stack % Error	2D % Error
LBP	1.7	13.6
GIST	4.5	7.5
SIFT	12.3	19.8
Hessian-Affine SIFT	11.9	24.5

Table 11: Comparison of utilizing all z-slices of image stack vs. summing image stack into a single 2D image.

5.5 Z-slice Location versus Accuracy

Figure 7 (left) shows the relationship between image stack region and percent error for species-level studies on modern data. Figure 7 (right) depicts the relationship between image stack region and percent error for the species-level studies on fossil data.

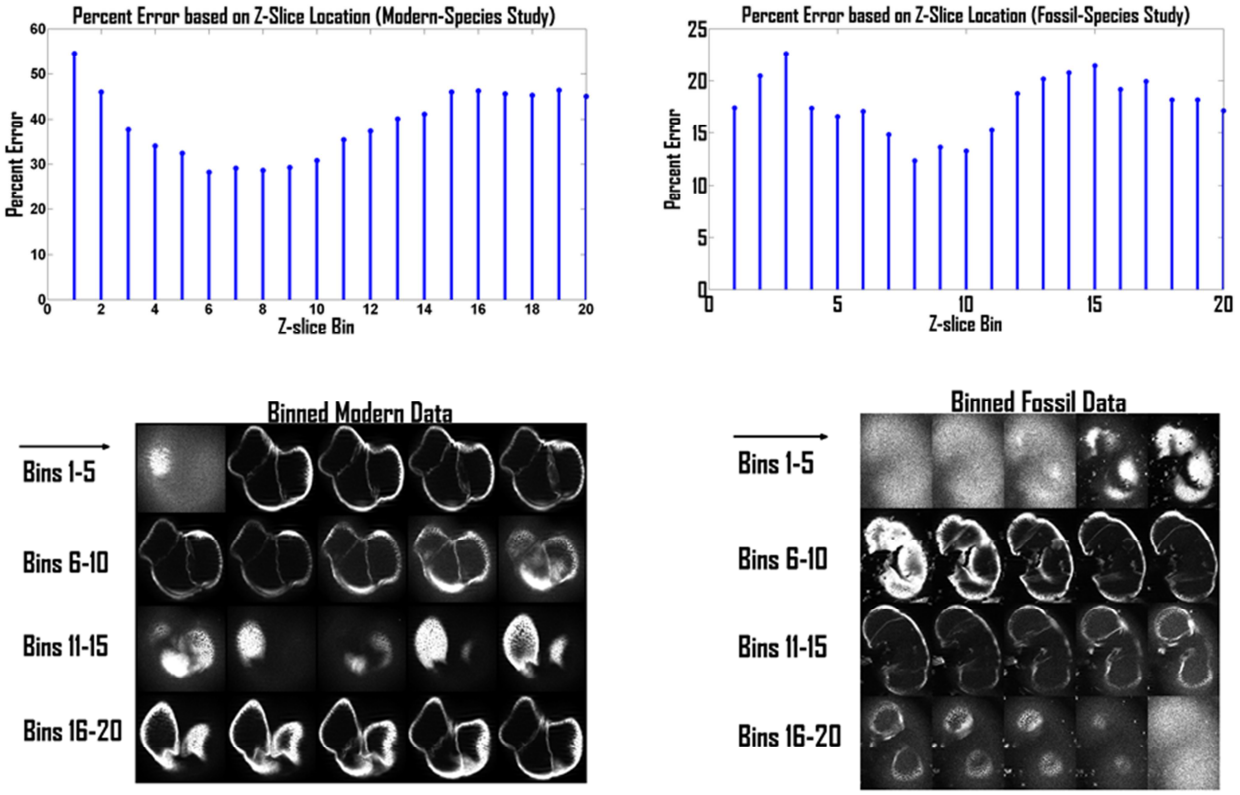

Figure 7: (Top, left) Percent error versus z-slice bin location for modern data: Species-level study (bottom, left) montage of binned z-stack images (first bin located at top left, last (20th) bin located at bottom right) for the modern data study. (Top, right) Percent error based on z-slice location for the fossil-species study (bottom, right) montage of binned z-stack images (first bin located at top left, last (20th) bin located at bottom right) for the fossil-species study.

5.6 Training Data Size versus Accuracy

Figure 8 shows the relationship between training size and error for the best methods for both modern tests as well as the fossil tests when analysts' confidence is greater or equal to 95%. The method for this test is discussed in section 4.5.

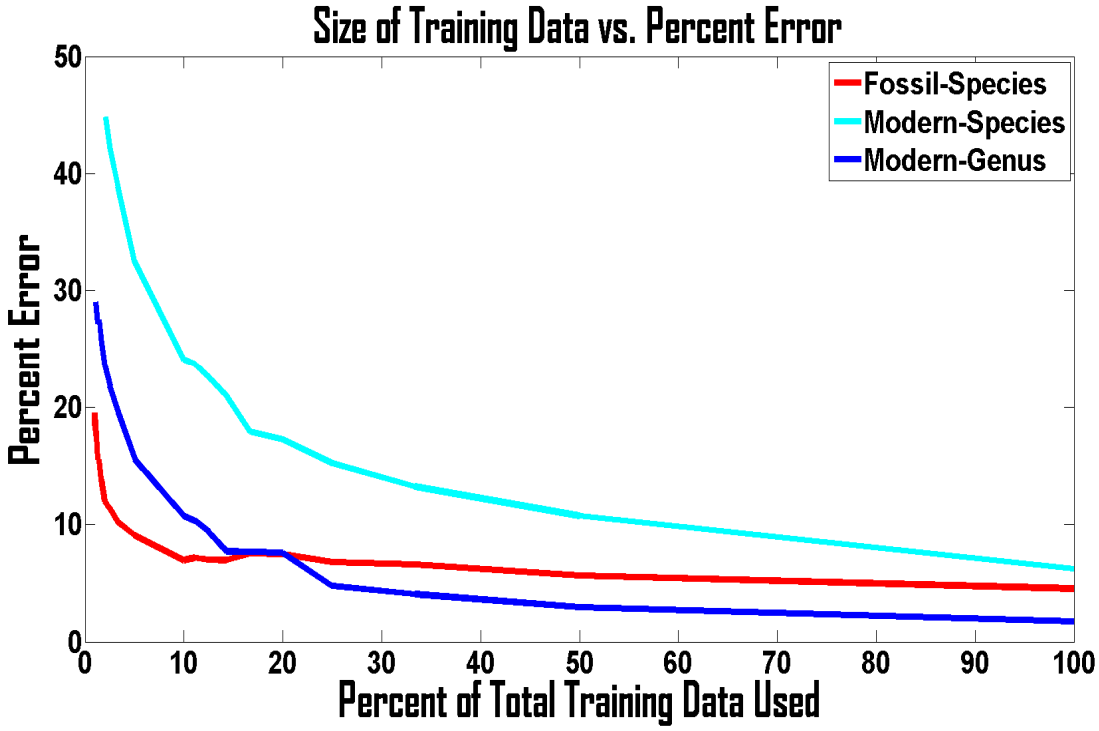

Figure 8: Percent error versus training data size.

6. DISCUSSION

6.1 Accuracy Results

When comparing our level of accuracy with our previous results [21] (see table 2), it is important to note that in the current study the only data analyzed were the images themselves. Our previous study utilized both pollen grain size (which was indirectly estimated by accounting for the area by manually created bounding boxes) as well as the images in order to classify the data. With the exception of the query test for classifying modern data using fossil data, a nearest neighbor classifier using the mode of all label estimates gave the best performance (see tables 3-10). For that particular query, utilizing only 50% of the z-slice estimates (labels related to the smallest pairwise distances) yielded a percent error of 13.7% while using the entire image stack produced an error of 16.8%. It is not surprising that a basic nearest neighbor classifier of one yielded the best performance. Given the 3D nature of the training images, some training images may not closely resemble their matching grains due to their difference in aspect angle. The nearest neighbor classifier allows the algorithm to choose the grain with the closest morphological features and the closest angle. Additionally, we found that using the image stack in its entirety yielded better results than rendering the image stack into a 2D representative image.

For the modern data studies, LBP yielded the best results using a minimum pairwise city block distance for genus and Euclidean distance for species. Also, the modern data showed slightly increased accuracy when data were median filtered

(kernel size 6) prior to applying the LBP transformation. For the fossil data study, SIFT yielded the best results with a minimum pairwise Euclidean distance. In contrast to LBP, applying a median filter slightly degraded performance. Given that median filters help smooth speckle noise, it is not surprising that in some cases applying a median filter can be advantageous.

When comparing the accuracy of these experiments, it is important to note that the success of the local methods (SIFT and Hessian-Affine SIFT) is reliant on a strong VQ codebook. In other words, if a proper outside dataset is not provided, the SIFT vectors will not be vector quantized properly and will provide poor performance. With the modern dataset, a dataset of images of mangrove pollen was provided. Although the mangrove data did contain some shape features that were similar to the modern dataset, we did not feel the physical characteristics of these images were close enough to those of the modern dataset. Therefore, our performance was not as good as expected. One advantage of the fossil data was the analyst confidence levels. Given that just two tests were performed on the data (utilizing only the data related to analysts' confidence greater than or equal to 95% and only the data related to analysts' confidence greater than or equal to 70%), the remaining data (data related to analysts' confidence less than 70%) was used to train our VQ codebook. It is not surprising, then, that the local method performed well on the fossil dataset.

As mentioned in the introduction, understanding whether grains collected from the same slides introduce bias was a concern in our previous study [21]. The potential for slide-level bias can occur due to similarities in background. While our previous results saw a decrease in accuracy of 0.4% (see table 2) when not controlling for which slide grains were imaged on, our more recent methods did not result in a change in accuracy.

6.2 Location vs. Accuracy

One reoccurring question that arose while performing this study was whether specific z-slices were more important than others. Observing figure 7 we see a jump in error inside the first 5% of the z-stack (bin 1) for species-level analyses. It is not surprising that the first few slices tend to be out-of-focus given that the structured illumination mechanism used to take the florescence images does not work well when a small part of the grain is at that focal length. Given the increase in error, we performed additional tests where we exclude the first 5% of the z-slices for each grain then reran our best classifier (nearest neighbor by z-slice). Since the error does increase slightly at the second half of the z-stack, we also performed a study utilizing only bins 2 through 10 out of 20 (or the top half minus the top 5%). Studies were performed at both the genus and species levels. The quality of the second half of the images tended to be inferior due to interference and shadowing caused by material at higher focal planes. For the genus-level study, we find that not including the top 5% decreased the error from 1.72% to 1.56%. However, the error rate increased up to 2.18% when we disregarded the bottom half of the z-stack along with the top 5%.

For the species-level study, we find that the accuracy holds constant at 6.24% error regardless of whether we discount the top 5% or the top half minus the top 5%. Given that the above two methods either show no effect on accuracy or slightly increase the accuracy, these methods were again performed for the fossil species-level study. We find that the accuracy remains constant at 4.55% for both methods.

In evaluating figure 7 (bottom images), we visually observe the first 5% of the data is the sparsest. When comparing the binned stem plot to the montage, it is not surprising that we find the highest error exists in these less detailed bins.

6.3 Training data vs. accuracy

One crucial question for any automated system is what makes a strong training set. The data used in this study were ideal in that there were a large number of images per class and only 2-5 classes per test. Understanding how drastically performance decreases as our training set decreases is crucial. Figure 8 shows this relationship for our three main studies. As mentioned before, the data were divided in a stratified fashion. In other words, while the subset of training data decreases, the original training data class ratio remains constant. For the modern data/genus-level study, the original training data size had three classes divided as follows: class 1: 96; class 2:103; class 3: 442 (as stated in table 1). Given that the smallest class representation was 96 grains while the largest was 442, the largest number we could divide our training data by was 96 in order to ensure that every class had at least one possible training representation. Since 103 and 442 both have remainders when divided by 96, the worst case error (29.55%), seen as the first datapoint in figure 8 (dark blue), occurs when there is only one representation of class 1, one to two representations of class 2 and four to five representations of class 3.

The same study was performed on the modern dataset at the species level. Again, the five classes used for the species-level classification were not equally represented (see table 1). Figure 8 (light blue) shows that our percent error increases from 6.24% up to 44.83% when we decrease our minimal representation from 48 grains per class to 1. Lastly, the analysis was performed on the fossil dataset. Observing the red plot in figure 8, the error rate increases from 4.6% to 19.5% when decreasing our minimum class representation from 108 grains per class to 1 grain per class. For all cases, we see that as we approached using 20% of the training data, or a minimum of 5-10 images per class, the error rose nonlinearly and drastically, which suggests that 20% may be a reasonable minimal threshold size for determining training set size requirements.

7. CONCLUSIONS

In this paper, we proposed exploring a number of collection and physical parameters of pollen to determine dependencies on classification accuracy. First, we determined that morphological information alone may be all that is necessary to correctly

classify grains. Second, we found that utilizing an entire image stack instead of a 2-D summed representation of the image stack yields significantly improved performance. While we did find that the first 5% of an image stack contained visually sparse data, we found the decrease in error related to excluding these data to be marginal and unlikely to be statistically significant. Third, we learned that our current algorithms did not show bias when handling grains from the same slide. This discovery is important because collecting a single pollen grain per slide would be fairly time consuming and costly.

It is important to note that our dataset was optimistic in the sense that we had few taxa and many examples per taxa. There were only three to five classes per test and each test was represented by 48 to 442 images. Tests performed on reduced training data size solidified the importance of a strong training dataset.

Our findings show great potential towards automating the classification of pollen grains while enforcing the need for a strong understanding of classification dependencies. While all of our studies yielded fairly low error rates, they also confirmed the importance of comprehensive metadata. We determined that the error rates doubled when analysts' confidence was decreased from $\geqslant$95% down to $\geqslant$70% and almost tripled when using data of varying ages. Future studies should concentrate on evaluating other collection parameters over a more diverse set of data and, more importantly, explore the relationship between the accuracy of geo-historical location to that of accurate pollen classification. Further evaluations founded on either computer vision or machine learning could establish collection parameters allowing for the creation of a robust semi-automated system. Such a system would reduce latency and cost while improving productivity for the analysts.

8. REFERENCES

[1] G. Hwang and D. Masters, Special Issue. "Palynology and Geolocation. Forensic geolocation challenge: is pollen analysis the answer? (http://www.palynology.org/newsletter)," *AASP - The Palynological Society*, pp. 1-78, 2013.

[2] V. Bryant and G. Jones, "Forensic palynology: Current status of a rarely used technique in the United States of America," *Forensic Science International*, vol. 163, pp. 183-197, 2006.

[3] A. Bertino, *Forensic Science: Fundamentals and Investigations 2012 Update*. Mason, OH: South-Western Cengage Learning, 2012.

[4] S. Scharring, A. Brandenburg, G. Breitfuss, H. Burkhardt, W. Dunkhorst, M. von Ehr, et al., "Online monitoring of airborne allergenic particles (OMNIBUSS)," in *Biophotonics: Visions for Better Health Care*, J. Popp and M. Strehle, Eds., ed Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. , 2006, pp. 31-87.

[5] G. Jones and V. Bryant, "Melissopalynology in the United States; a review and critique," *Palynology*, vol. 16, pp. 63-71, October 1, 1992 1992.

[6] A. Cross, "Palynology and its relation to the exploration for oil," *The Mountain Geologist*, vol. 1, p. 91, 1964.

[7] F. Gonzalez, C. Moreno, R. Saez, and G. Clayton, "Ore genesis age of the Tharsis Mining District (Iberian Pyrite Belt): a palynological approach," *Journal of the Geological Society*, vol. 159, pp. 229-232, 2002.

[8] D. Nichols and S. Jacobson, Palynology in coal systems analysis-the key to floras, climate and stratigraphy of coal-forming environments, Geological Society of America Special Paper 387: 51-58

[9] V. Bryant and D. Mildenhall, "Forensic palynology: a new way to catch crooks," in *New developments in palynomorph sampling, extraction, and analysis*, V. M. Bryant and J. W. Wrenn, Eds., ed Dallas, Texas: American Association of Stratigraphic Palynologists Foundation, 1998, pp. 145-155.

[10] D. Stoney, A. Bowen, V. Bryant, E. Caven, M. Cimino, and P. Stoney, "Particle combination analysis for predictive source attribution: Tracing a shipment of contraband ivory," *Journal of American Society of Trace Evidence Examiners*, vol. 2, pp. 13-72, 2011.

[11] D. Korejwo, J. Webb, D. Willard, and T. Sheehan, "Pollen Analysis: An Underutilized Discipline in the U.S. Forensic Science Community," presented at the Trace Evidence Symposium. , Quantico, Virginia, 2007.

[12] A. Traverse, "Paleopalynology: Second Edition," ed: Springer, 2008, pp. 58-63.

[13] G. Allen, R. Hodgson, S. Marsland, and J. Flenley, "Machine vision for automated optical recognition and classification of pollen grains or other singulated microscopic objects," in *Mechatronics and Machine Vision in Practice, 2008. M2VIP 2008. 15th International Conference on*, 2008, pp. 221-226.

[14] G. Hwang, K. Riley, C. Christou, G. Jacyna, and J. Woodard, "Semi-automated pollen identification system for forensic geolocation applications," *AASP - The Palynological Society*, vol. 46, p. 28, 2013.

[15] M. Rodriguez-Damian, E. Cernadas, A. Formella, M. Fernandez-Delgado, and S. Pilar De, "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*, vol. 36, pp. 531-542, 2006.

[16] M. Rodriguez-Damian, E. Cernadas, A. Formella, and R. Sa-Otero, "Pollen classification using brightness-based and shape-based descriptors," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 212-215 Vol.2.

[17] N. Nguyen, M. Donalson-Matasci, and M. Shin, "Improving pollen classification with less training effort," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, 2013, pp. 421-426.

[18] G. Allen, B. Hodgson, S. Marsland, G. Arnold, R. Flemmer, J. Flenley, et al., "Automatic recognition of light microscope pollen images," *In Proc of Image Vision and Computing New Zealand*, 2006.

[19] A. Boucher, P. Hidalgo, M. Thonnat, J. Belmonte, C. Galan, P. Bonton, et al., "Development of a semi-automatic system for pollen recognition," *Aerobiologia,* vol. 18, pp. 195-201, 2002.

[20] O. Ronneberger, E. Schultz, and H. Burkhardt, "Automated pollen recognition using 3D volume images from fluorescence microscopy," *Aerobiologia,* vol. 18, pp. 107-115, 2002.

[21] S. Punyasena, D. Tcheng, C. Wesseln, and P. Mueller, "Classifying black and white spruce pollen using layered machine learning," *New Phytologist,* vol. 196, pp. 937-944, 2012.

[22] *http://bioimagel.com/yahoo_site_admin/assets/ images/stack.333130145.jpg. Accessed 08/2012.*

[23] R. Filipovych, N. Sublette, E. Ribeiro, and M. Bush, "Pollen Recognition in Optical Microscoy by Matching Multifocal Image Sequences (in review)," ed.

[24] G. Dahme, E. Ribeiro, and M. Bush, "Spatial Statistics of Textons," in *International Conference of Computer Vision Theory and Applications - VISAPP*, ed. Setubal, Portugal, 2006.

[25] O. Ronneberger, H. Burkhardt, and E. Schultz, "General-purpose object recognition in 3D volume data sets using gray-scale invariants - classification of airborne pollen-grains recorded with a confocal laser scanning microscope," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, pp. 290-295 vol.2.

[26] C. Travieso, J. Briceno, J. Ticay-Rivas, and J. Alonso, "Pollen classification based on contour features," in *Intelligent Engineering Systems (INES), 2011 15th IEEE International Conference on*, 2011, pp. 17-21.

[27] T. Ahonen, J. Matas, C. He, and M. Pietikainen, "Rotation invariant image description with local binary pattern histogram fourier features," *Lecture Notes in Computer Science,* vol. 5575, pp. 61-70, 2009.

[28] A. Oliva and A. Torralba, "Chapter 2 Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research.* vol. Volume 155, Part B, S. Martinez-Conde, et al., Eds., ed: Elsevier, 2006, pp. 23-36.

[29] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, pp. 1150-1157 vol.2.

[30] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int J Comput Vision,* vol. 60, pp. 91-110, 2004.

[31] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, et al., "A Comparison of Affine Region Detectors," *Int J Comput Vision,* vol. 65, pp. 43-72, 2005/11/01 2005.

[32] A. Andreopoulos and J. Tsotsos, "50 Years of object recognition: Directions forward," *Computer Vision and Image Understanding,* vol. 117, pp. 827-891, 2013.

[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Prediction, Inference and Data Mining, Second Edition*. New York Springer, 2009.

[34] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans Intell Syst Technol*, vol. 2, pp. 1-27, 2011.

[35] MATLAB version 8.1.0.604 Natick, Massachusetts: The MathWorks Inc., 2013.

[36] G. Hwang, K. Riley, C. Christou, G. Jacyna, J. Woodard, R. Ryan, et al. (in press), Automated pollen identification system for forensic geo-historical location applications. *The 13th annual IEEE Conference on Technologies for Homeland Security*.

[37] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." *IJCAI*. Vol. 14. No. 2. 1995

Manuscript received (editor provides date)

Footnotes:

Page 12: These methods yielded universally poor results and therefore are not further discussed in this paper.

Affiliations of authors:

1) Kimberly C. Riley, Jeffrey P. Woodard and Grace M. Hwang
The MITRE Corporation
7515 Colshire Drive, Mclean, VA, USA

2) Surangi W. Punyasena

University of Illinois at Urbana-Champaign,
505 S. Goodwin Avenue, Urbana, IL, 61801, USA

Acknowledgement of financial support:

Data collection was funded by the University of Illinois. The MITRE Corporation and the University of Illinois both funded the studies and evaluations discussed in this paper.

Figure Captions

1. Page 3, Figure 1: (Top) Morphologically similar genus-level pollen grains of the *Pinaceae* family (from left to right: *Abies*, *Picea*, *Pinus*). Image data from [21]. (Bottom) Image stack (left image) of one grain along with its summed 2D representation (top right). Image stack representation (bottom right) taken from [22].
2. Page 4, Figure 2: Flow diagram of semi-automated system. A new pollen image stack is introduced to the system. This image is then compared to a set of database images for classification. The estimated class is then used to create a geo-historical location probabilistic distribution model.
3. Page 6, Figure 3: Flow diagram of LBP-HF (Local Binary Pattern- Histogram Fourier Features)
4. Page 8, Figure 4: (Top) Overlay of SIFT regions of one z-slice of a pollen grain image stack. (Bottom) Visual representation of how SIFT bins its features.
5. Page 10, Figure 5: Overview of algorithm: A new pollen grain (represented by an image stack) is introduced to the system. It's then converted to a feature vector representation (also called a descriptor). Lastly, it is compared to training feature vectors in order to estimate its classification value. Figure is modified from [36].
6. Page 11, Figure 6: Flow diagram of 3D nearest neighbour classification. Image stack representation taken from [22].
7. Page 17, Figure 7: (Top, left) Percent error versus z-slice bin location for modern data: Species-level study (bottom, left) montage of binned z-stack images (first bin located at top left, last (20th) bin located at bottom right) for the modern data study. (Top, right) Percent error based on z-slice location for the fossil-species study (bottom, right) montage of binned z-stack images (first bin located at top left, last (20th) bin located at bottom right) for the fossil-species study.
8. Page 18, Figure 8: Figure 8: Percent error versus training data size.