Improving Pollen Classification with Less Training Effort

Nhat Rich Nguyen[1], Matina Donalson-Matasci[2], and Min C. Shin[1]

[1]Department of Computer Science, University of North Carolina at Charlotte
[2]Department of Biology, University of Arizona
{nhnguye1,mcshin}@uncc.edu , matina.email@arizona.edu

Abstract

The pollen grains of different plant taxa exhibit various shapes and sizes. This structural diversity has made the identification and classification of pollen grains an important tool in many fields. Despite the myriad of applications, the classification of pollen grains is still a tedious and time-consuming process that must be performed by highly skilled specialists. In this paper, we propose an automatic classification method to discriminate pollen grains coming from a variety of taxonomic types. First, we develop a new feature that captures the spikes of pollen to improve the classification accuracy. Second, we take advantage of the classification rules extracted from the existing pollen types and apply them to the new types. Third, we introduce a new selection criterion to obtain the most valuable training samples from the unlabeled data and therefore reduce the number of needed training samples. Our experiment demonstrates that the proposed method reduces the training effort of a human expert up to 80% compared to other classification methods while achieving 92% accuracy in pollen classification.

1. Introduction

The pollen grains of different plant taxa exhibit many different shapes and sizes, often bearing characteristic ornaments like spines or furrows. This structural diversity has made the identification of pollen grains an important tool in a variety of fields. Aerobiologists identify windborne pollen to warn allergy sufferers in periods of elevated risk. Classifying pollen collected from pollinators like bees, hummingbirds and butterflies provides a record of different flower taxa each individual has visited. In agriculture and conservation biology, pollen classification has practical implications for which plants are actually receiving pollination services, as well as the nutrition and health of the pollinators themselves. Figure 1 shows a dense population of pollen containing various types under an optical microscope.

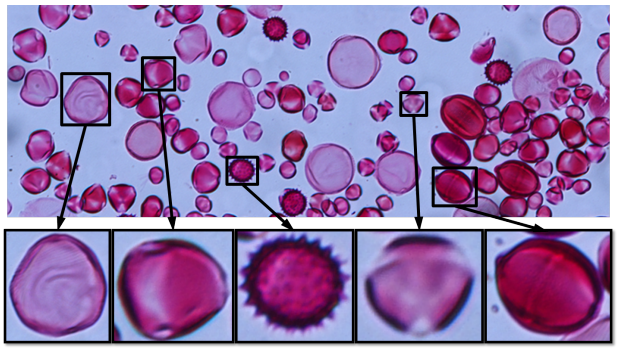
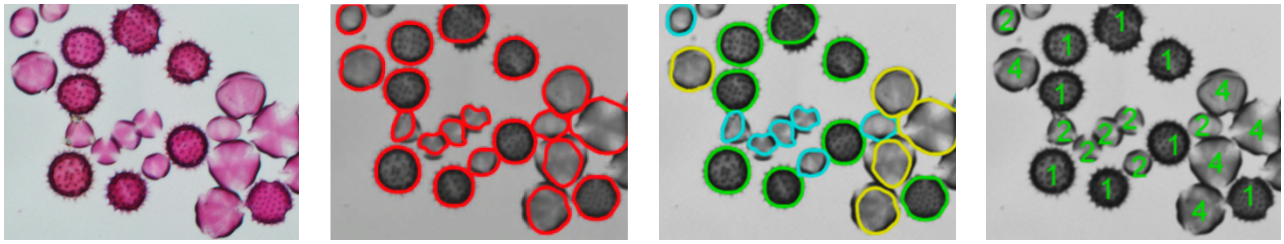


Figure 1. The typical image of pollen contains a dense population of many pollen types viewed under a microscope. Several representative samples from different pollen types are zoomed in to show the diversity in shape and texture of the pollen.

Previous Research Despite the myriad of applications, the classification of pollen is still a tedious and time-consuming process that must be performed by highly skilled palynologists who specialize in the study of pollen. Recently, several groups have developed automated methods for pollen classification. Most such methods involve a small number of known pollen types and require a large number of training samples per type [6, 9, 5]. As the number of pollen types increases, it seems that more and more training samples are required. For example, an automated system proposed by Allen *et al*. requires 40 training samples per type to classify 7 types, but as many as 150 training samples per type for 17 types [1]. To build a more comprehensive classifier that could deal with an increasing number of types over time, the number of samples required for training would thus likely be prohibitive. Another study has shown good performance up to 30 types with as few as 18 training samples per type [2]; however, the classification rules are derived by human experts. Such an approach is likely to scale poorly as the number of types increases, because the complexity of the classification models and the amount of expert knowledge required to build them grow rapidly as



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

Figure 2. An example of pollen classification using the proposed method. (a) A cropped region from a microscopic image containing a few pollen types. (b) The pollen grains are detected with active contours where features are extracted. (c) The classification pollen grains shown in unique colors. (d) The corresponding manual labels from a human expert.

more similar types are added.

In this paper, we introduce a machine learning approach designed to reduce the amount of training effort required to classify an increasing number of pollen types. Without loss of generality, we assume there are some existing pollen types with available labels. As the number of types increases, the new types usually have a small number of training samples initially. We propose to reduce the number of training samples for the classification on the new types by leveraging the knowledge learned from the existing types. To this end, a classifier is trained on a few existing types (source types) and is able to extrapolate some of that knowledge into new types (target types). Furthermore, we propose a selection criterion to search for the most confusing samples that are most likely to belong to the target types from a large pool of unlabeled samples. Using a set of pollen images containing various mixtures of nine pollen types, we demonstrate that the proposed method reduces the training effort of a human expert as much as 80% compared to other classification methods while achieving 92% classification accuracy. Additionally, we show an application of the proposed method in pollen counting. Figure 2 illustrates an example of the pollen classification using the proposed method.

2. Methods

Our aim is to automatically identify and classify pollen grains of various types in images captured by an optical microscope. We first localize the boundary of pollen grains, then extract representative features from the boundary shape and the texture region (Section 2.1). Next, we propose to minimize the amount of effort to train the classifier by transferring the classification rules (Section 2.2) and selecting the most valuable training samples (Section 2.3).

2.1. Pollen Detection and Feature Extraction

To localize the precise boundary of each pollen grain, we utilize active contours [11] which have been used extensively in many applications. Since the shapes of pollens are varied from circular to slightly elliptical, we estimate the initial boundary of active contours by detecting circles using the circular Hough transform.

Previous studies have compiled a set of shape and texture features derived from the detected boundary [6, 1, 9]. The shape features computed from the boundary of the active contours include area, diameter, ratio of area and perimeter, compactness, roundness, rates of changes, thickness, elongation, centroid, Euclidean norm, mean size, eccentricity, and circularity. While shape features utilize the extracted boundary, the texture features are derived from the rectangular region which encloses the extracted boundary. Our texture features include the first-order statistics, Haralick's Coefficients from the gray level co-ocurrence matrix (GLCM), and the gray-level run length. The computation details of these features are described in [9].

Spike Count Spikes are important features that can be discriminative among some pollen types. Figure 3 provides details on the extraction of the spikes. Due to the inherent image quality and resolution, the active contours are unable to capture the spikes which often are too noisy and faint. However, knowing the final position of the active contour will help identify these spikes. Thus, we extract a "ring-like" binary mask along the active contour to estimate the region of the spikes. Within this mask, a spike usually appears as a fluctuation in intensity. For each pollen grain, the average intensity of pixels at each angle is computed. Then, a 1-D signal formed by the intensity values at every angle is generated. A local minimum in the signal is detected as a spike if its difference to the values at both adjacent local maxima are within a range of $[0.05, 0.40]$. Note that we are able to distinguish a spike from the border of a neighbor grain which yields a large intensity difference (as seen in Figure 3).

2.2. Transfer of Classification Rules

Transfer learning has been shown to achieve high classification accuracy from a small number of training samples

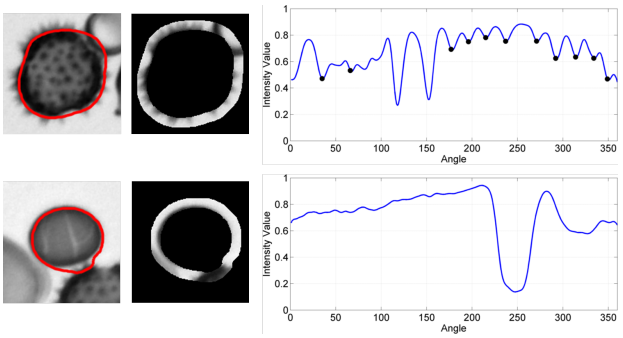


Figure 3. The spike count discriminates pollen type which has many spikes (upper row) from no spike (lower row). Left column: the active contour is unable to capture the spikes. Middle column: the radial mask indicates the estimated region of the spikes. Right column: a radial intensity curve consists of the average intensity of the pixels at each angle. The local minima which satisfies a specific condition are counted as spikes (as black dots along the curve). Note that although the pollen in the lower row has a neighbor with spikes, no spike is counted as the intensity fluctuation at the boundary is too large.

in several applications [3, 12, 8]. In our approach, we take advantage of some classification rules, in the form of base classifiers, extracted from the existing pollen types (source types). These classification rules then are applied to the new pollen types (target types) given only a small number of labeled samples.

Extracting Classification Rules In order to extract the classification rules from the source types, we employ a well-known learning algorithm, adaptive boosting (AdaBoost) [4], to distinguish the labeled samples of one source type from another. We collect the base classifiers from all possible pairs of the source types into a candidate set according to the *transferability* to the target type. The highest transferability occurs when one of the source type is equivalent to the target type. In such a case, the transfer learning would become a traditional machine learning problem. To ensure a high transferability, we select only the base classifiers that yield an classification error less than a threshold value in their respective source type. Additionally, we find a redundancy problem where some base classifiers from multiple sources are practically identical. Thus, we sort the base classifiers on their splitting attributes and then eliminate ones which have a negligible difference to others.

Applying Rules to Target Samples Training the classifier directly from a small number of target samples might lead to over-fitting those samples. Thus, we modify a transfer learning algorithm, TaskTrAdaBoost [12], to evaluate each base classifier in the candidate set using the target samples as positive data and samples from a source type as negative data. At each boosting iteration, a base classifier which minimizes the error over the labeled data is chosen. The error is based on the number of labeled samples which are incorrectly classified. When we apply base classifiers from source types to the target data, it is possible that a base classifier has a classification error exceeding 50%. This happens when the samples of a target and a source type occupies the opposite sides of the decision boundary (refer to the first plot of Figure 4). Thus, we invert the rule of the base classifier to adapt to the target data and the error becomes less than 50%. This training error satisfies the boosting condition described in [4].

Extending to Multi-class So far, we have trained a binary classifier to discriminate a target type against a source type. To extend to a multi-class classification, we use an "one-versus-one" strategy: for K types, a total of $\frac{K(K-1)}{2}$ binary classifiers are trained for all possible pairs. Given the feature vector $\mathbf{x}_i$ of a pollen sample, a binary classifier which distinguishes a pair of pollen types is derived similarly to the boosting framework as $\hat{f} = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}_i)$ where h_t is selected as the base classifier, α_t is computed from the classification error, and T is the boosting iterations. Let $p_l(k|\mathbf{x}_i)$ be the probability output for the boosting algorithm to classify type k from type $l \neq k$. This probability output is directly computed from the classifier $\hat{f}$ using a sigmoid function $p_l(k|\mathbf{x}_i) = \frac{1}{1+\exp(-\hat{f})}$. Assume that all $p_l(.)$ are independent, the posterior probability for type k can be computed as $P(k|\mathbf{x}_i) = \prod_{l \in K} p_l(k|\mathbf{x}_i)$. Finally, the type of the pollen sample $\mathbf{x}_i$ is predicted as

$$\hat{y}_i = \arg\max_{k \in K} P(k|\mathbf{x}_i). \quad (1)$$

2.3. Selection Criterion for Unlabeled Samples

Since the detection and feature extraction are fully automatic, a large pool of *unlabeled* grains can be collected from the pollen images. In Section 2.2, we leverage the classification rules from the source types to handle small training samples of the target type. Based on these classification rules, additional target samples are selected from the unlabeled pool to re-train the classifier. In this section, we further enhance the classification of the target type by selecting valuable target samples from the unlabeled pool. A simple approach would be to randomly select samples from the unlabeled pool and then use only the samples of the target types for training. However, in a classification problem with a large number of pollen types, the random sampling approach obtains only a small portion of target samples from the unlabeled pool. Instead, we formulate a search for the samples which are confusing to the classifier as well as likely to be a target type.

We obtain valuable target samples by modifying a popular active learning approach, margin sampling [7, 10]. The

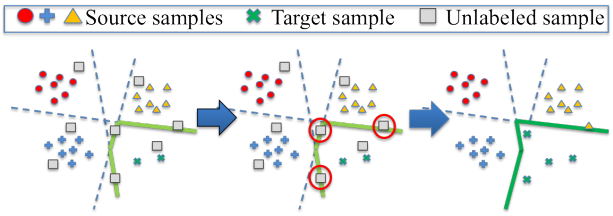


Figure 4. A simple illustration of the decision boundary formation in a training iteration using the transferred rules to select valuable target samples. Left: the classification rules from the source types (in dashed blue lines) are transferred to form the decision boundary of the target type (in green lines). Center: the proposed method actively selects the most confusing unlabeled samples (in red circles) located near the decision boundary between the target type and a source type. Right: knowing the labels of these samples helps improve the decision boundary (now in dark green lines).

improvement of a simple decision boundary using the modified margin sampling to select training samples is illustrated in Figure 4. An unlabeled sample is said to be confusing if the difference between its probabilities of belonging to two types is minimal. Consequently, we select the unlabeled sample which has minimal difference between its highest probability type $\hat{y}_1 = \arg\max_k P(k|\mathbf{x}_i)$ and second highest probability type $\hat{y}_2 = \arg\max_{k,\hat{y}_2 \neq \hat{y}_1} P(k|\mathbf{x}_i)$ where $\mathbf{x}_i$ is the unlabeled sample (similar to Equation 1). As we aim to improve the target classifier, the unlabeled samples with a small difference between the target type and a source type are more likely to be valuable. To encourage the classifier to search for the target samples, we introduce a new selection criterion:

$$\begin{aligned} \arg\min_{\mathbf{x}_i} \quad & P(\hat{y}_1|\mathbf{x}_i) - P(\hat{y}_2|\mathbf{x}_i) \\ \text{satisfies} \quad & (\hat{y}_1 - \tau)(\hat{y}_2 - \tau) = 0 \end{aligned} \tag{2}$$

where $\hat{y}_1$ and $\hat{y}_2$ are the predicted labels which return the highest and second highest posterior probability $P(.)$, and τ is the target type. In the above equation, we employ the minimization constraint $(\hat{y}_1 - \tau)(\hat{y}_2 - \tau) = 0$ to focus the search on the target samples. Constrained by the above equation, samples with a small margin between the target type and a source type are more likely to be selected. While the traditional margin sampling would have focused on selecting a minimal margin samples from *any* type, this selection criterion favors the confusing samples that are most likely to belong to a *target* type. Using this selection criterion, we rank all unlabeled samples into an ordered set. Finally, the top samples of this ordered set are selected to be labeled by the human annotator for the next iteration of training.

3. Experiment

We conduct a series of experiments on pollen images to demonstrate the performance of the proposed method against two boosting-based classification algorithms: AdaBoost [4] and TaskTrAdaBoost [12]. The AdaBoost algorithm does not employ either the transferred knowledge from the source types or any selection strategy for new training samples. A recent transfer learning approach, TaskTrAdaBoost, exploits the knowledge from source types to be re-used on the target types. Since TaskTrAdaBoost does not explicitly provide a strategy to select new samples, its new training samples are selected randomly from the unlabeled data. When the target training data is small, TaskTrAdaBoost is demonstrated to be effective in exploiting the existing knowledge to learn the new data in [12].

Data The pollen images used to test the proposed method were taken from an experiment done on domestic honey bees. On a microscope with a motorized stage, non-overlapping regions of the microscope slide are scanned at 40x magnification into a digital image using the software NIS-Elements (Nikon). Each image covers approximately 1mm^2 at a resolution of 0.23 μm/pixel. Previous classification by a human expert indicates that these samples contain a various mixtures of pollen types. We collect a total of 768 grains of 9 pollen types as shown in Figure 1.

Detection Evaluation We use a standard metric to evaluate the detection performance: precision and recall. By definition, precision = TP / (TP + FP) and recall = TP / (TP + FN) where TP is the number of true positives, FN is false negatives, and FP is false positives. A detected grain from the algorithm is a true positive if the distance to the closest manually labeled grain is within the diameter of the smallest pollen type; otherwise it is a false positive. Any undetected grain is considered a false negative. Overall, the detection method performs at 93.8% recall and 89.5% precision.

Classification Procedure To evaluate the accuracy of a classification method, we select each pollen type as the target and randomly choose 6 other types from the training set to be used as the source types. All classification methods are provided initially with only 3 target samples which are selected randomly, and the rest of samples form the unlabeled pool. At each iteration of training, 5 additional unlabeled samples are selected for a human expert to label; then the classifier is re-trained based on the newly labeled samples. We repeat the training for 50 iterations. The classification accuracy on the target type is evaluated by the ratio between the number of correctly classified target samples and the total number of target samples in the test set. We record the classification accuracy of the target type per each iteration. The overall accuracy is computed as the average over all types. To eliminate the bias of selecting the initial target training samples, the experiment is replicated 30 times and the average performance of each method is recorded.

Table 1. The classification accuracy of the proposed method (in %) with respect to the spike count feature. The spike count improves the classification accuracy in many types (bold faced).

Pollen Type	1	2	3	4	5	6	7	8	9	Average
Accuracy (%) without Spike Count	96	100	97	89	72	93	92	93	70	89
Accuracy (%) with Spike Count	**100**	100	97	89	**77**	93	89	**98**	**83**	**92**

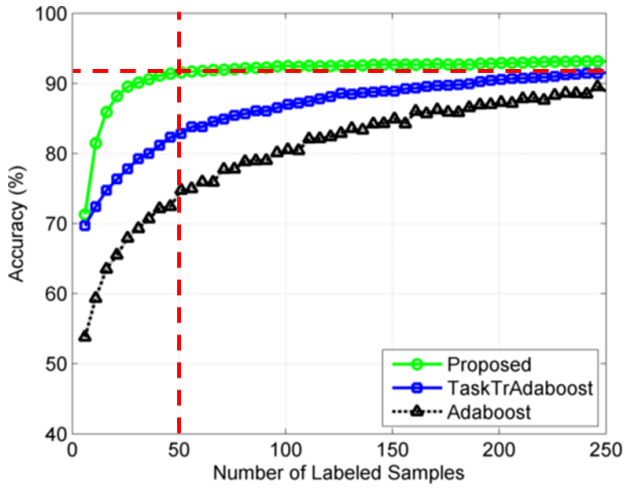


Figure 5. Comparison of the classification methods with respect to the number labeled samples. The proposed method achieve a comparative performance to other method while requires 80% less number of labeled samples.

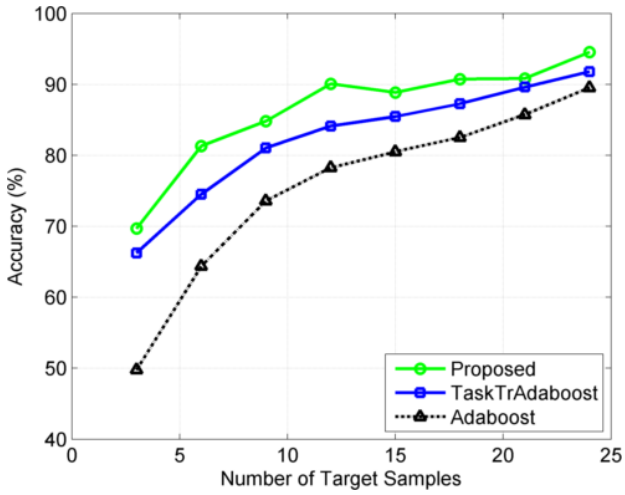


Figure 6. Comparison of classification methods with the same number of *target* samples. Higher accuracy is observed in the proposed method implying that it selects more valuable target samples to improve the classification.

4. Results

We assess the effectiveness of the proposed method on three different aspects: (1) the accuracy improvement with respect to the spike count feature; (2) the reduction of the number of labeled samples to achieve the performance of previous classification methods; and (3) the consistency to a human expert in a biological application.

4.1. Improvement in Classification Accuracy

We first measure the improvement in classification accuracy by integrating the spike count feature (described in Section 2.1). Table 1 provides the accuracy of the proposed method with and without the spike count. Without the spike count, type 5 and 9 are confused with type 1 with the errors rates of 13% and 18% respectively. The spike count feature improves the accuracy by 5% for type 5 and 13% for type 9 since it discriminates them with type 1 which has many spikes. Additionally, the accuracy improves in type 8 by 5% since it has a smooth boundary with no spike. We also observe a slight increase in error rate for type 7 due to its confusion to type 6 which has a similar number of spike. Overall, the error rate is reduced over 3% in all types resulting in 92% accuracy.

4.2. Reduction in Training Effort

Figure 5 displays the average testing accuracy on all target types with respect to the number of labeled samples. Note that the accuracy improves with the proposed method at any number of labeled samples. The improvement with a small number of labeled samples is especially large: 11% over TaskTrAdaBoost and 23% over AdaBoost when trained with 10 samples. While TaskTrAdaBoost and AdaBoost require more than 250 labeled samples to achieve a 91% accuracy, the proposed method only requires 20% of the training effort, or 50 samples, to achieve a similar performance (as shown in the red dotted lines in Figure 5). This result suggests that the proposed method reduces the workload of the human annotator as high as 80% and still achieves a comparative performance to other methods. The reduction in training effort is caused by two reasons. First, the proposed method selects more target samples, on average 23 target samples out of the first 50 unlabeled samples while TaskTrAdaBoost and AdaBoost select only 11 target samples. Second, the selection criterion (Equation 2) selects more useful training samples. Figure 6 compares the results of all three methods with the same number of *target* samples. The accuracy of each classification method is

Table 2. Comparison between the automatic and manual methods in the application of pollen count. Since the experiment is replicated 30 times, the automatic count is provided as Mean±Standard Deviation. The manual count is done only once by a palynologist. Using t-tests at 90% confidence level, there is no significant difference found between the automatic and manual methods.

Pollen Type	1	2	3	4	5	6	7	8	9
Automatic Count	286±2.6	123±0.7	147±3.1	38±1.7	32±3.0	28±1.0	20±0.7	62±2.6	32±2.8
Manual Count	289	120	145	39	31	26	22	63	33
Count Error (%)	1.0	2.6	1.5	1.0	3.9	6.6	9.8	2.4	3.4

plotted according to the number of target samples is presented in the selected unlabeled samples. The performance of TaskTrAdaBoost is higher AdaBoost due to the transfer of rules which is consistent with [12]. Our method shows an additional improvement over TaskTrAdaBoost supporting that it selects more effective ones.

4.3. Application in Pollen Counting

Counting the grains of each pollen type on the microscope slides is a slow laborious process that must be performed by highly skilled palynologists. We compute the pollen distribution from the classified grains and compare to the ground truth established by a palynologist on the data set (described in Section 3). Using t-tests with 90% confidence level, there is no statistically significant difference between manual and automated methods; and the average error is as small as 3.6% (refer to Table 2). Higher error rates are observed on type 6 (6.6%) and 7 (9.8%) due to the low sample frequency of such types. Compared to other automated pollen counters such as in [5], the difference in count in our case is indeed small (on average ≤ 2 grains per type). Thus, we believe the method has enabled a reliable pollen counter for the biology community.

5. Conclusions

In this paper, we propose an automatic classification method to discriminate pollen grains with a variety of taxonomic types. First, we develop a new feature that captures the spikes of pollen to improve the classification accuracy. Second, we take advantage of the classification rules extracted from existing pollen types and apply them to the training samples of the new types. Third, we introduce a new selection criterion to obtain the most confusing samples of the target types from the unlabeled data and therefore improve the classification performance. The proposed method achieves 92% accuracy in pollen classification while reducing the training effort up to 80% compared to other classification methods. We believe the proposed method shows great potential toward the automation of pollen identification and counting which is commonly done by palynologists. As a future work, we would like to investigate on improving the classification performance as the number of pollen types increases over time.

References

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

[2] A. Boucher, P. Hidalgo, M. Thonnat, J. Belmonte, C. Galan, P. Bonton, and R. Tomczak. Development of a semi-automatic system for pollen recognition. *Aerobiologia*, 18(3):195–201, 2002.

[3] W. Dai, Q. Yang, G. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, page 200. ACM, 2007.

[4] Y. Freund and R. Schapire. A short introduction to boosting. *Japonese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[5] K. Holt, G. Allen, R. Hodgson, S. Marsland, and J. Flenley. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology*, 167(34):175–183, 2011.

[6] P. Li, W. Treloar, J. Flenley, and L. Empson. Towards automation of palynology 2: the use of texture measures and neural network analysis for automated identification of optical images of pollen grains. *Journal of quaternary science*, 19(8):755–762, 2004.

[7] T. Luo, K. Kramer, D. Goldgof, and L. Hall. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6:589–613, 2005.

[8] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345 –1359, Oct 2010.

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

[10] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[11] C. Xu and J. Prince. Snakes, shapes, and gradient vector flow. *Image Processing, IEEE Transactions on*, 7(3):359 – 369, Mar 1998.

[12] Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *Computer Vision and Pattern Recognition (CVPR) IEEE Conference on*, pages 1855 –1862, Jun 2010.