

Kerogen Classification Using Image Analysis

Author:



University of Wales Trinity Saint David

3rd May 2022

Research Methods and Professional Development

Research Proposal

MSc International Innovation Management

Supervisor: Graham Howe



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Contents

Introduction	3
Literature Review	4
Key Questions.....	6
Methods	6
Results	9
Timeline	10
Costs	10
Further developments.....	11
References	12
Appendix	14

Introduction

The analysis of the organic debris present in palynological samples, also known as kerogen analysis, corresponds to methodology widely used in the geological community for developing palaeoenvironmental reconstructions, inferring the type of hydrocarbon produced by source rocks and also to help understanding sedimentation settings throughout the geological history of the different basins. Overall, this type of analysis categorises the different types of particles such as algae, pollen and spores, and plant remnants at a microscopic level and the percentages obtain will provide information regarding the parameters described above.

Currently, a substantial amount of progress has been done in the field of Image Analysis and Image Recognition. Different sectors, that span from medical and pharmaceutical to social media and technological gadgets have been leveraging from this technology and have been rapidly evolving to adapt and adopt assisted and/or automated image recognition.

Different studies have attempted to implement Image Analysis in palynological samples (e.g., Hwang et al., 2013; Marcos et al., 2015; Riley et al., 2015). However, these studies have mainly focused on Image Analysis applied to taxonomical identification of pollen and spores to a *Genus* or *species* level. This, of course requires a very high level of accuracy in the recognition of the different images due to the detailed taxonomical features that define each *species*.

Some studies using Image Recognition in kerogen analysis have been conducted (e.g., Kuncheva et al., 2008). However, the results obtained have not been appealing enough for this methodology to be widely adopted. As a result, currently the kerogen classification continues to be carried out using visual observations gathered by an analyst using a transmitted light microscope. However, the published studies have not tested recent technology developed in the field of Image Recognition

This research proposes the use of recently developed and open-source technology such as *Orbit Image Analysis* (Stritt et al., 2020) and *TensorFlow* developed by Google (Abadi et al., 2015) to implement assisted kerogen analysis and classification and evaluate the feasibility of using this technology. The potential impacts of this research will translate to a company-wide adoption of this technology and higher efficiency when conducting this type of analysis, or alternatively, the further steps required to leverage from this technology.

Literature Review

Palynology corresponds to the analysis of acid-resistant microscopic particles such as pollen, spores, algae and acritarch. However, additional acid-resistant debris accompany these particles in a palynological sample. This debris ranges from particles such as the ones mentioned above, to fragments of plants, structured and unstructured algal remains and fungal remains. The analysis that encompasses all this kind of organic particles is known as kerogen analysis, or palynofacies analysis. Combaz (1964) defined the term palynofacies as the total acid-resistant organic content present in the sedimentary rocks obtained in a palynological sample. Powell et al, (1990) defines palynofacies as a specific association of palyno-debris that reflects a specific depositional environment.

The particular usefulness of these acid-resistant particles for geoscientists is that they remain unaffected hundreds of millions of years and therefore provide valuable insights on the past history of the Earth (e.g., Carvalho et al., 2013; Mueller et al., 2014) and with specific groups, such as marine algae being useful chronological and environmental indicators (e.g., Zonneveld et al., 2013; Head & Mantilla-Duran, 2020).

The paleoenvironmental reconstructions and identifying the type of kerogen present in the sedimentary rocks is also of particular interest for companies in the oil and gas sector as it provides information on the understanding of the different petroleum systems that can have implications in the commercial aspect of the different drilling campaigns and projects. Hence, an appetite for more efficient interpretations using state of the art methods is highly needed.

Different schemes for kerogen classification have been proposed over the past decades. These vary from morphological features (structured and unstructured) to source (marine, lacustrine, continental) (e.g., Tyson, 1995; Oboh-Ikuenobe et al., 2005). However, one of the most comprehensive ones and widely used corresponds to the one proposed by Tyson (1995).

Image recognition constitutes an automated recognition and identification of targeted subjects in images and/or videos based on regular features and patterns of the images. The technique typically uses Machine learning, a concept first introduced by Samuel (1959) to refer to self-trained computers. Essentially, this technology uses algorithms to generate models and get self-trained with exponential improvement from accumulative exposure. In general, this technology has already embedded in applications that society uses on daily basis, such as identification of license plates for car parks, tagging in social media, and face recognition for unlocking phones among others.

In science and engineering, image recognition coupled with machine learning has been used in the past decade on different areas. An example of the application of this technology in the medical field for instance, is the use of image recognition and

machine learning to assist on identification of pulmonary tuberculosis on chest radiographs (Lee et al., 2021). Studies on biology have also implemented this technology to classify specific disease- carrying ticks (Omodior et al., 2021).

With regards to palynology, different studies have focused on the use of image recognition mainly for pollen classification. Riley et al. (2015) used fossil and modern pollen grains to test the accuracy of species identification with results of around 13.7% to 16.8% of error. Marcos et al. (2015) focused on the texture of 15 different taxa from 1800 images to perform image identification.

For kerogen analyses, not many studies have been carried out. Kuncheva et al. (2008), being one of the most recent ones, attempted to develop an automatic classification system that identifies two types of particles from palynological samples: vitrinite and inertinite. The results showed accuracies ranging from 86% to 92%. Although the study did not include additional particles, the accuracy values obtained in identifying these two particles, provide a good level of confidence and highlights a promising path towards automatic kerogen analysis. Moreover, when it comes to kerogen classification, there are sometimes disagreements between experts when labelling problematic particles, meaning that there is a tolerable margin of error when conducting kerogen classification. To some extent, the Ternary diagram proposed by Tyson (1995) (Fig. 1), allows an extra degree of flexibility as it lumps the particles into 3 major categories. Hence, in theory, slight misidentifications should not have a major impact on the overall interpretation of the data.

The examples of this type of technology successfully being applied to different areas of science shows the need for building a bridge across these disciplines and bring across this technology for implementing it in kerogen analyses.

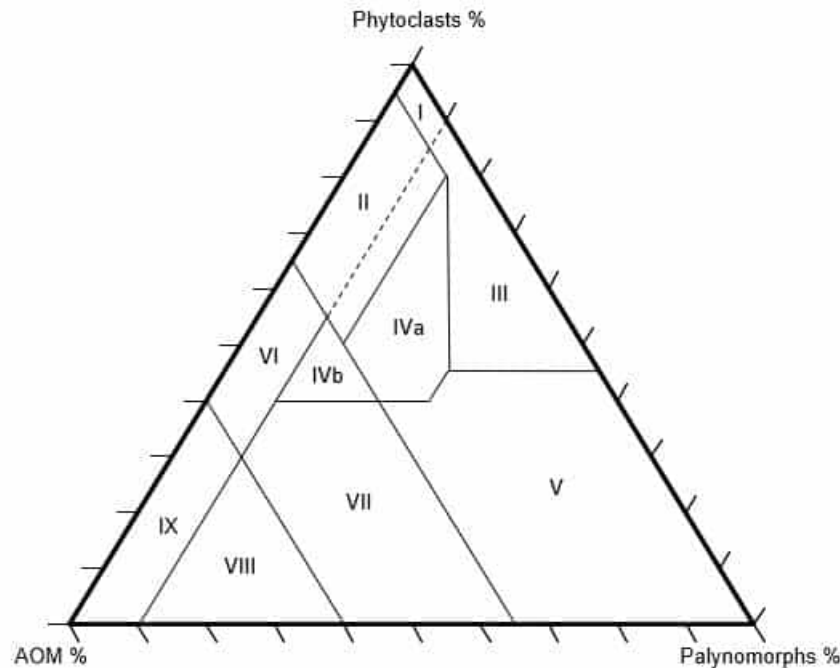


Fig.1: Ternary diagram for environmental interpretations based on palynofacies analysis. From Tyson (1995). **I.** Highly proximal shelf or basin. **II.** Marginal disoxic-oxic basin. **III.** Heterolithic oxic shelf. **IV.** Shelf to basin transition. **V.** Mud-dominated oxic shelf. **VI.** Proximal suboxic-anoxic shelf. **VII.** Distal dysoxic-anoxic shelf. **VIII.** Distal disoxic-anoxic shelf. **IX.** Distal suboxic-anoxic basin.

Key Questions

This research aims to address the following questions:

- 1) Is the publicly available software and scripts powerful enough to provide a robust classification of kerogen?
- 2) If not robust enough, is the margin of error acceptable for general interpretations using the Tyson ternary diagram?
- 3) If none of the above, evaluate the reasons why the results are not acceptable and discuss whether or not this technology is still far from being routinely used in kerogen analysis.

Methods

For this research, the processing of the palynological samples and image labelling will be conducted in the premises of [REDACTED] a geological consultancy company with the head office in Conwy, North Wales.

Sample processing:

40 rock samples spanning different depths in the stratigraphic column will be selected for processing. It is important to select an interval from an outcrop or sub-surface well that contains variations in the depositional settings to obtain a high diversity of kerogen particles. If an outcrop is selected, the geological description will provide this information. For a sub-surface well, available wireline logs will indicate which interval contains more fluctuations in the depositional settings.

The samples will potentially be obtained from the British Geological Survey (BGS) in Nottingham, with additional options being Iron Mountain in Aberdeen or CGG in North Wales. Once a suitable section/wellbore has been identified, one of these 3 sample repositories will be contacted.

The samples will then be processed using the standard processing for palynological samples as follows:

- Wash with detergent approximately 30 gr of sediment (approx. 100gr if wet sediment from sub-surface wellbore). Subsequently place the sediment in beakers and transfer them to a fume cupboard.
- Add small amounts of 20% conc. cold HCl to the beakers to remove the carbonates from the rock material. It is important to add the acid in small increments to avoid any possibility of bubbling over due to high amounts of carbonates present in the rock samples. Gradually add 150 mL of acid to each beaker and swirl the sediment into suspension and leave it for 3 hours.
- Swirl the sediment to bring it back into suspension and add distilled water to about 500 ml to neutralise any traces of acid. Measure the pH to confirm neutral conditions. If not, repeat this step until a pH of about 6 is obtained. Then filter out the liquid and keep the remaining sediment in the beaker.
- Add 38% conc. cold HF up to 150 mL to remove silicates from the rock material. Due to the hazardous nature of this acid, thorough risk assessments and safety measures must be in place when using it. Allow the acid to react for 3 hours and then decant the liquid in a sodium carbonate solution for neutralising.
- Top up the beaker where the sediment is with distilled water and swirl to bring the sediment back into suspension. Confirm a pH of 6 or repeat this step if not fully neutral.
- Then sieve the remaining sediment using a 10 µm size mesh, pipette out the obtained material and mount in a microscope slide.

Image Acquisition:

Once the microscope slides have been mounted, they will be passed to a specialist who will use a Nikon Eclipse E600 transmitted light microscope coupled with a Leica MC170 HD microscope camera to obtain the images.

The analyst will then photograph 10 different fields of view using a 400x magnification for each microscope slide, this will provide a dataset of 400 photographs for the studied interval.

Following the image acquisition, each photograph will be exported to the open-source software *Orbit Image Analysis* (Stritt et al., 2020). This software is specifically designed to allow labelling and annotation on each image, which constitutes an important step for training automatic image recognition. Then, the specialist will label a minimum of 3 different particles on each photograph to provide a training dataset of 1,200 labels. It is important for the analyst to aim for a high diversity of kerogen particles to ensure a wide range of particles being included in the dataset.

Data training:

The labelled files will then be used as a training library available open-source scripts for classifying images based on TensorFlow developed by Google (Abadi et al., 2015). This technology can be used for in a wide range of scenarios, but it particularly focuses on allowing the user to implement neural networks for image recognition and classification. From these images, 80% will be used as a training dataset and the 20% remaining will be left out to be used as a verifier for the trained data. In other words, to test that the trained data is correctly classifying particles not used in the training dataset but with known labels for the analyst. This 20% of labelled data will be assigned as *known dataset*.

Cross-checking:

After the data-training and verification, the analyst will be asked to plot her/his interpreted data from the *known dataset* into the Tyson ternary diagram and provide a palaeoenvironmental interpretation based on this. The best format for this interpretation will be a depth vs palaeoenvironments curve plotted vertically together with any additional available data such as wireline logs, core descriptions, lithological descriptions, lithostratigraphy and chrono/biostratigraphy.

Likewise, the results from the automated classification obtained from the *known dataset* using the TensorFlow-based script will be formatted for plotting them in the Tyson ternary diagram. Similarly, the obtained data will be used to provide an interpretation of palaeoenvironments, and the data will be plotted replicating the format provided by the analyst for easy comparison.

Results

Image analysis and recognition has been rapidly evolving. However, the adoption of this in palynological and kerogen analysis has been to some extent lagging. With more and more available scripts for data recognition available, it is hypothesized that the use of this in kerogen classification should provide acceptable results.

There are 3 potential outcomes from this study:

- 1) Accurate results. Providing that the amount of data used as a training library is sufficient and diverse enough, and the TensorFlow-based scripts implemented are robust and generate accurate models, a successful outcome will be obtained whereby the computer will accurately identify and classify the different particles in a kerogen analysis. This will then be used as the raw data for interpreting palaeoenvironments in the stratigraphic column. The impact of this outcome will be substantially higher efficiencies when conducting kerogen analysis and a more cutting-edge technology being used in the analysis that will ultimately be more appealing to the clients. This can also be used as a marketing topic for the company.
- 2) Reasonable results. If the image identification is somewhat inaccurate, the results will be evaluated to see if the overall interpretation is still usable. When compiling the data for plotting it in the Tyson ternary diagram, it is assumed that the misidentification of some particles should not skew the data too much to substantially compromise the overall trends in the data. For instance, a pollen being misidentified as a spore will not bear any negative impact to the data. Hence, the basic idea with these results will be to provide a general overview of the kerogen present that can be of use when combining it with other proxies, such as wireline logs, core descriptions and biostratigraphy. A direct comparison between the digitally obtained data and the one generated by the analyst should clearly provide information on when the two datasets substantially differ. This will then be used to indagate on the reasons why this is happening in specific intervals.
- 3) Inaccurate results. The results obtained from the image identification show high levels of inaccuracy, and direct comparisons with the data generated by the analyst show major discrepancies. This will trigger a series of discussions on why the obtained data shows a substantial margin of error and evaluate the feasibility of further steps to improve the use of this type of technology in kerogen analysis.

Timeline

The estimated timeline for this project is shown in the Table 1 below:

	Month 1				Month 2				Month 3				Month 4				Month 5				Month 6				Month 7				Month 8			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
*Project set-up	■	■	■																													
Sample selection		■	■	■																												
Sample processing					■	■	■																									
**Identification of TensorFlow script					■	■	■	■	■	■	■																					
Image acquisition									■	■	■	■	■																			
Image labelling													■	■	■	■	■															
Training dataset														■	■	■	■	■														
test of <i>known dataset</i>															■	■	■	■	■													
Analyst data plotting																■	■		■	■	■											
Analyst interpretation																	■	■		■	■	■										
Computer data plotting																		■	■		■	■	■									
Computer interpretation																				■	■		■	■	■							
comparison																					■	■		■	■	■						
Discussion and write-up																						■	■	■	■	■	■	■	■	■	■	

Table 1: Estimated timeline for the development of the research project

*The project set-up will capture all the components required prior to starting the research, such as authorisation for sample collection, resource allocation in the labs for sample processing, and configuring the equipment and systems prior to commencement of the study.

** Time allocated for finding an adequate and suitable script will vary and will depend on background research on the different options available to identify the right candidate(s).

Costs

The costs associated to the development of this project are rather low. Most of the requirement will correspond to man hours as the main components of the study will be data gathering (namely image acquisition and image labelling). However, this component will be covered by the student during downtime and/or free time. In addition, the software and scripts to be used for image labelling and dataset training are open source and therefore will not incur in any costs. The additional costs associated to sampling selection and processing described below will be covered by [REDACTED] as part of an internal Research and Development project.

The table below (Table 2) outlines the costs associated to this project.

Costing Table

	Units	Unit Cost	Total
Sample selection and shipment	1	£120	£120
Sample processing	40	£125	£5,500
TOTAL			£5,620

Table 2: Main costs associated to this project

Further developments

Depending on the results, this study could provide the first building blocks for making the use of this technology a default practice in kerogen analysis. Specific workflows will then need to be put in place to secure consistency in the interpretations obtained using computer-based image classification. Furthermore, this type of technology has successfully been applied to track and identify objects in real time (e.g., Talele et al., 2019). Hence, if a positive outcome is achieved by this study, further developments can be investigated to attempt to implement real-time particle identification in kerogen analysis to assist the analyst when analysing the sample for a different purpose

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X. (2015) 'TensorFlow: Large-scale machine learning on heterogeneous systems' *Software available* from tensorflow.org.
- Carvalho, M. de A., Cabral Ramos, R.R., Crud, M.B., Witovisk, L., Kellner, A.W.A., Silva, H. de P., Grillo, O.N., Riff, D. & Romano, P.S. R. (2013) 'Palynofacies as indicators of palaeoenvironmental changes in a Cretaceous succession from the Larsen Basin, James Ross Island, Antarctica', *Sedimentary Geology*, 295, pp. 53-66.
- Combaz, A. (1964) 'Les Palynofacies', *Revue de Micropaleontologie*, Vol. 7, pp.205-218.
- Martin J. Head, M. J. & Mantilla-Duran, F. (2020) 'Atlanticodinium striaticonulum n. gen., n. sp., a widespread extant dinoflagellate cyst from the late Cenozoic, and its comparison with Atlanticodinium janduchenei (Head et al., 1989) n. comb. *Marine Micropaleontology*, Vol.159. <https://doi.org/10.1016/j.marmicro.2019.101774>.
- G. M. Hwang, G. M., Riley, K. C., Christou, C. T., Jacyna, G. M., Woodard, J. P., Ryan, R. M., Bush, M. B., Valencia, B. G., McMichael, C. N. H., Punyasena, S. W., Masters, D. L. (2013) 'Automated pollen identification system for forensic geo-historical location applications', *IEEE International Conference on Technologies for Homeland Security (HST)*, pp. 297-303, doi: 10.1109/THS.2013.6699018.
- Kuncheva, L.I., Charles, J.J., Miles, N., Collins, A., Wells, B., Lim, I.S. (2007) 'Automated Kerogen Classification in Microscope Images of Dispersed Kerogen Preparation' *Mathematical Geosciences* Vol. 40, pp. 639-652.
- Lee, J.H., Park, S., Hwang, E.J., Goo, J.M., Lee, W.Y., Lee, S., Kim, H., Andrews, J.R., Park, C.M. (2021) 'Deep learning-based automated detection algorithm for active pulmonary tuberculosis on chest radiographs: diagnostic performance in systematic screening of asymptomatic individuals' *European Radiology*, Vol. 31, pp. 1069-1080.
- Marcos, J.V., Nava, R., Cristobal, G., Redondo, R., Escalante-Ramirez, B., Bueno, G., Deniz, O., Gonzalez-Porto, A., Pardo, C., Chung, F., Rodriguez, T. (2015) 'Automated Pollen Identification Using Microscopic Imaging and Texture Analysis' *Micron*, Vol. 68, pp. 36-46.

Mueller, S., Veld, H., Nagy, J. & Kurschner, W.M. (2014) 'Depositional History of the Upper Triassic Kapp Toscana Group on Svalbard, Norway, Inferred from Palynofacies Analysis and Organic Geochemistry', *Sedimentary Geology*, Vol. 310, pp. 16-29.

Omodior, O., Saeedpour-Parizi, M. R., Rahman, M. d. K., Azad, A., Clay, K. (2021) 'Using convolutional neural networks for tick image recognition—a preliminary exploration' *Experimental and Applied Acarology*, Vol. 84(3): pp. 607–622.

Oboh-Ikuenobe, F.E., Obi, C.G., Jaramillo, C.A. (2005) 'Lithofacies, Palynofacies and Sequence Stratigraphy of the Paleogene Strata in Southeastern Nigeria', *Journal of African Earth Sciences*, Vol. 41, pp.79-101.

Powell, A.J., Dodge, J.D., Lewis, J. (1990) 'Late Neogene to Pleistocene Palynological Facies of the Peruvian Continental Margin Upwelling, Leg 112', *Proceedings of the Ocean Drilling Program, Scientific Results*, Vol. 112, pp. 297-331.

Riley, K.C., Woodard, J.P., Hwang, G.M., Punyasena, S.W. (2015) 'Progress towards establishing collection standards for semi-automated pollen classification in forensic geo-historical location applications', *Review of Palaeobotany and Palynology*, Vol. 221, pp. 117-127.

Samuel, A.L. (1959) 'Some Studies in Machine Learning Using the Game of Checkers' *IBM Journal of Research and Development*, Vol. 3, no. 3, pp. 210-229

Stritt, M., Stalder, A.K., Vezzali, E. (2020) 'Orbit Image Analysis: An Open-Source Whole Slide Image Analysis Tool' *PLOS Computational Biology*, Vol. 16(2) doi.org/10.1371/journal.pcbi.1007313

Talele, A., Patil, A., & Barse, B. (2019). 'Detection of Real Time Objects Using TensorFlow and OpenCV. *Asian Journal For Convergence In Technology (AJCT)* ISSN -2350-1146. Retrieved from <https://asianssr.org/index.php/ajct/article/view/783>

Tyson, R.V. (1995) 'Sedimentary Organic Matter: Organic Facies and Palynofacies' *Chapman and Hall*: 615pp.

Wang, S., Yang, D.M., Rong, R., Zhan, X., Xiao, G. (2019) 'Pathology Image Analysis Using Segmentation Deep Learning Algorithms' *The American Journal of Pathology*, Vol.189, pp. 1686-1698. doi.org/10.1016/j.ajpath.2019.05.007

Zonneveld, K.A.F., Marret, F., Versteegh, G.J.M., Bogus, K., Bonnet, S., Bouimetarhan, I., Crouch, E., de Vernal, A., Elshanawany, R., Edwards, L., Esper, O., Forke, S., Grøsfjeld, K., Henry, M., Holzwarth, U., Kieft, J.-F., So-Young, K., Ladouceur, S., Ledu, D., Chen, L., Limoges, A., Londeix, L., Lu, S.-H., Mahmoud, M.S., Marino, G., Matsouka[sic], K., Matthiessen, J., Mildenhall[sic], D.C., Mudie, P., Neil, H.L., Pospelova, V., Qi, Y., Radi, T., Richerol, T., Rochon, A., Sangiorgi, F., Solignac, S., Turon, J.-L., Verleye, T., Wang, Y., Wang, Z., and Young, M. (2013) 'Atlas of modern dinoflagellate cyst distribution based on 2405 datapoints'. *Review of Paleobotany and Palynology*, Vol. 191, p. 1–197.

Appendix

Example of two fields of view from two different palynological samples under the microscope showing some of the particles analysed for kerogen analysis.

