# AN AUTOMATED MACHINE-LEARNING PROCEDURE FOR ROBUST CLASSIFICATION OF SEM IMAGES OF CROSS-LAMINATED SANDSTONES FOR DIGITAL ROCK ANALYSIS

Chen Jin and Jingsheng Ma

Institute of Petroleum Engineering, Heriot-Watt University, UK

*This paper was prepared for presentation at the International Symposium of the Society of Core Analysts held in Avignon, France, 8-11 September, 2014*

## ABSTRACT

Depositional structures such as laminae in sandstones reflect local changes in grain size, shape, orientation and composition. Laminae can occur as a set of parallel or intersecting structures, depending on the depositional processes. Further aided by diagenetic modifications, pore structures and pore surface properties may vary to a large degree, in their topology and geometry as well as physicochemical nature, within each lamina and across a set of laminae. The combination of spatial arrangements and local properties means that laminations can greatly influence the flow of multi-phase fluids. As an example, it is well-known that sandstone laminae can trap a significant amount of hydrocarbon in reservoirs. Most of the previous studies, however, treat each lamina as a uniform continuum without taking into consideration the true grain-pore characteristics associated with them (see [1] and references therein), whereas fewer others did but for tabular lamina only [2].

To gain a fuller understanding of the aforementioned combinational effect on multi-phase fluid flow and to obtain more appropriate estimates of effective properties for cross-laminated reservoir rock samples, we are taking digital rock analysis approach by reconstructing the pore structures of representative samples and then numerically simulating fluid flow through the pore systems. Because a representative sample deems to be much larger than a usual core plug, the reconstruction calls for multi-scale imaging techniques (e.g. using industrial CT scanner, microCT, Scanning Electron Microscopy (SEM)) to capture the spatial arrangements of the laminae and associated diverse pore structures, and deterministic and stochastic integration techniques to fuse obtained images. In integration, the lamina structures, which are captured in low-resolution 3D images, need to be calibrated against those identified in high-resolution 2/3D images, in order to reconstruct fine-scale grains and pores in those coarser structures. However, it is non-trivial to identify those structures in a high-resolution image. In this report, we develop an automated machine-learning procedure for image classification to perform this task. We illustrate this procedure using an SEM image of a cross-laminated tight sandstone sample. This work is an attempt to extend multi-scale data integration for digital rock analysis

beyond what has been proposed for core plugs [3] to larger and structurally more complex samples.

# INTRODUCTION

The purpose of our work is to show how to use image analysis to distinguish sedimentary structures, namely tabular and cross laminations from a high-resolution grey-scale SEM image of a tight sandstone sample. We do so by image classification based on a range of image features, each of which characterises a certain aspect of the grey-scale intensity data of an image. For example, the mean and standard deviation of grey-scale values in a selected window of an image measure the local variation and can be used to compare two windows for their similarity, while the Sobel filter, treated as a feature here, identifies the edges of the sub-regions. Other image features used in this work are mentioned in a later section. A known fact of image classification for a natural porous material is that its true physical characters are rarely related to image features through a simple and linear relationship.

SEM has long been used successfully in petrographic analysis [4, 5]. It can image the whole thin section of a typical size at a resolution down to a few hundred nanometres – sufficient for resolving grains and pores for a tight sandstone sample. However, classifying a high-resolution SEM image is not trivial because: 1) the composition of grain-forming materials does not always allow adjacent pixel values to be distinguished un-ambiguously and robustly; 2) an image may contain too much detailed information that may distract and obscure the identification of large-scale patterns of concern; and 3) a naive interactive classification often presumes a simple and linear relationship between the image features and true classes.

Given an SEM image, simple image filtering techniques are capable of identifying specific textual patterns of the image, but incapable of discovering a complex nonlinear relationship between features and classes. Machine learning is an approach where hidden relationships in a system can be discovered through learning from the data using computer algorithms. It has been shown that many machine-learning algorithms can be trained for a nonlinear system using a small set of samples to make correct predictions where an unknown sample belongs to a certain part of the system. In what follows we develop an automated procedure to classify the grey-scale values of a given SEM image into classes corresponding to physical characters.

## The Procedure

Figure 1 illustrates our classification procedure. Given an SEM image, sub-domain samples are taken randomly first and then are classified or labelled into classes without supervision, i.e. without knowing ‘true’ labels. This set of samples becomes a training dataset on which the whole image is to be classified using a supervised algorithm that makes use of class labels directly. Finally the success of the classification is evaluated.

To use this procedure, the following need to be decided: 1) the resolution of the thinnest lamina in the input image; 2) the sizes of sub-sampling templates; 3) the numbers of sub-samples to take; and 4) features and algorithms for unsupervised and supervised classifications, respectively.

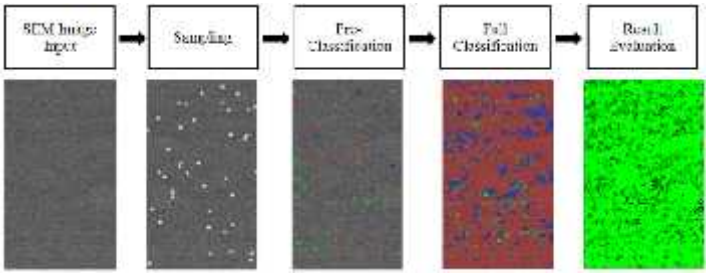


Figure 1 A schematic diagram of a classification procedure for an SEM image of a cross-laminated sandstone. Note that the last image shows the pixel-wise difference of two classified images.

As mentioned above, a high-resolution SEM image, which may be created by tiling images taken on an array of overlapping sub-domains in turn, may not be suited for identifying laminae because there is too much sub-resolution information that can be distractive for classification. Therefore, such sub-resolution information should be suppressed first. Multi-scale spatial patterns in an image may be separated out using techniques such as curvelet transformation [6] and suppressed selectively. Since only laminae are of interest, one may even choose to simply coarsen an image by averaging pixels. An image may be scaled by a factor to ensure that the thinnest lamina can be resolved sufficiently and therefore appropriate templates can be defined.

Given an image, the dimensions of an appropriate template in a certain shape (e.g. square) should be smaller than the thickness of the thinnest lamina and large enough so that the coefficient of variation (Cv) of the mean values for all templates at the same size is small. The mean value of each template is calculated as an arithmetic mean of the grey-scale values in a template. To determine a suitable template, one can carry out a test by positioning templates of a number of sizes at randomly selected sites, and then choose the size that gives a local minimum of the Cv. Note that the Cv is expected to oscillate from one template size to another due to the constraint that a template must be smaller than the thicknesses of the thinnest lamina. The degree of that oscillation seems to be greater if the resolution of the image is lower. To determine how many samples are required, a random sampling can proceed by increasing the number of samples so that the Cv decreases to a stable value.

At this point, one will obtain a set of samples and can label them by an unsupervised algorithm such as K-means or an even simpler one as described in a following section. These labelled templates form a labelled training dataset that is needed for classifying the whole image using supervised algorithms. The number of classes could be chosen heuristically. Since the results of a supervised classification depend on the labels of a training dataset, it is important to label the templates using few generic image features, such as the mean and standard deviation; they measure broad characters of interest and are suitable for unsupervised classification.

There are many supervised classification algorithms available. A suitable algorithm must be able to: 1) take multiple features; 2) discover the hidden nonlinear relationship between features and the true resulting classes; and 3) distinguish weak differences in the image reliably and robustly. Most of the supervised classification algorithms do meet the first two requirements. The third requirement can only be met if one can choose features that can characterise important subtle variations of the image.

This procedure can be automated if an *a-priori* estimate of the thinnest lamina of an image can be determined.

**A Demonstration Case**

To demonstrate the procedure above, we implemented it in ImageJ (http://imagej.nih.gov/ij/), using the ImageJ macro programming language and WEKA plugin [7]. Then we carried out the classification on an SEM image that was obtained from a thin section of a fine-grained tight Triassic red sandstone sample (Figure 2a) in Southwest Scotland, UK.

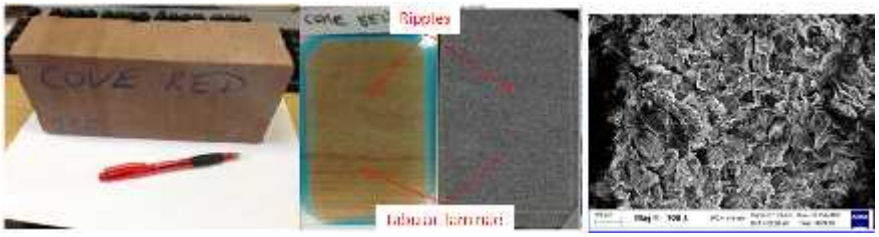


Figure 2 (a) The sample block; (b) 1x1.5 inch impregnated thin section; (c) SEM image (17328x26378 pixels, resolution≈1.1um); and (d) grains, pores and their scales.

The sample contains visible laminae of climbing ripples at the top, and tabular laminae at the bottom, and the former are less distinguishable than the latter due to mineral changes, a character often useful for classifying laminae [8, 9]. Because information contents in the image are different for these two types of laminae, it is important to recognise and distinguish them. Figure 2a shows the sample block, from which an impregnated thin section (Figures 2b) was obtained, and subsequently scanned using SEM (Figure 2c) where the yellow rectangle region was used in this demonstration. Figure 2d shows grain and pore sizes of the sample and they are below 100 microns.

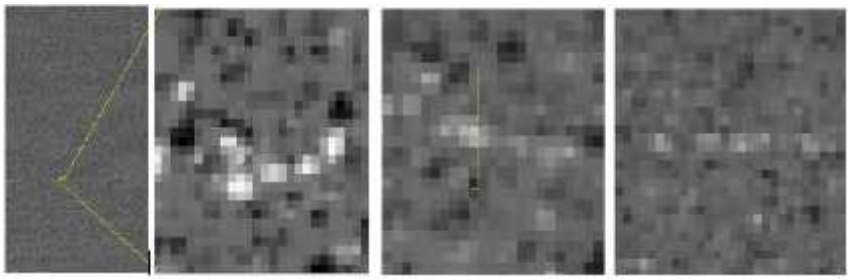

Figure 3 (a) the SEM image and a selected region on a lamina (b) I (c) II (d) III. The heights of the yellow boxes are the thicknesses of that lamina after rescaling, respectively.

Following the procedure above, we first chose to rescale the image, by local averaging, into 3 different sizes (Figure 3a): I - 900x1424, II - 450x712 and III -

265x419 pixels, respectively. As shown in Figure 3b to 3d, the thinnest lamina is approximately 20, 12 or 5 pixels, respectively.

For these three images, random sampling was carried out using a square template at a number of template sizes. The minimum template sizes were determined to be 13x13 12x12 and 9x9 pixels for I, II and III, respectively, according to the Cv in Figure 4a. III was over-scaled because the template dimension was larger than the thinnest lamina and therefore excluded from the further analysis. The number of the samples is chosen to be 100. The remaining two images, I and II are both suitable for further analysis but only II was chosen in this demonstration. Figure 4 plots Cv versus the template size and the number of samples.

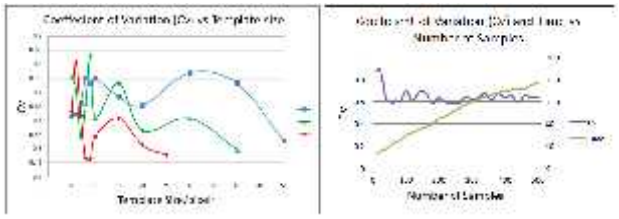

Figure 4 (a) Coefficient of Variation (Cv) vs. Template Size; (b) Cv and Run Time vs. the Number of Samples for II.

One hundred templates of 12x12 pixels in size were selected randomly and then labelled into 3 groups using a simpler scheme as shown in Figure 5a. Figure 5b shows 100 labelled random samples using a template of 12x12 pixels in size. Having had this labelled training dataset, we carried out the supervised classification using WEKA's fast random forest algorithm. We used the following 9 features: **G**aussian **B**lur(GB), **S**obel **F**ilter(SF), **H**essian(H), **D**ifference of **G**aussians(DG), **M**embrane **P**rojections(MP), **M**ean(M), **V**ariance(V), **E**ntropy(E) and **N**eighbours(N) in the supervised classification. Figure 5c shows the resulting classification for II. Note that the red label indicates denser and better sorted laminae, the green colour indicates the lamina interfaces where mineral change emerges, and the blue colour the coarser and poorly sorted laminae. It is evident that this procedure can identify not only high contrast laminae, due to the change of mineralogy, but also weakly contrasting laminae or within-lamina patterns, due to grain size and orientation changes.

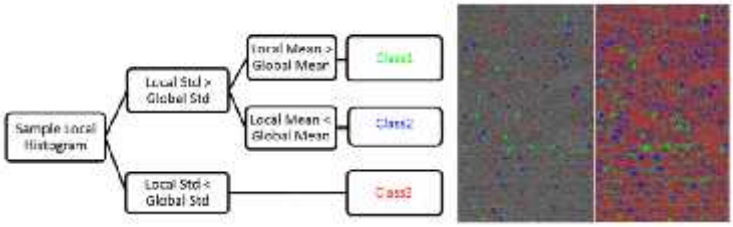


Figure 5 (a) A scheme for labelling randomly sampled templates used in this work. (b) Samples labelled (c) WEKA classification result.

To explore the contribution of each feature, in addition to M and V, to the classification, we calculate the similarity for each selected pair of features for 6 combinations as

Table 1 Comparison of Classification of Pixel-wise Matched Classes (%).

| | M-V-GB | M-V-SF | M-V-H | M-V-DG | M-V-MP | M-V-E |
|---|---|---|---|---|---|---|
| M-V-SF | 76.24 | | | | | |
| M-V-H | 71.92 | 72.31 | | | | |
| M-V-DG | 78.92 | 76.86 | 73.31 | | | |
| M-V-MP | 77.60 | 75.93 | 72.79 | 78.12 | | |
| M-V-E | 72.52 | 72.77 | 70.99 | 73.42 | 72.83 | |
| M-V-N | 74.11 | 74.45 | 73.90 | 75.57 | 74.77 | 72.50 |

shown in Table 1. The similarity is calculated as the percentage of pixels being identical (i.e. same classes) between a selected pair of classified images. Clearly the difference between any pair is around 25-30% and this suggests that a better classification could be achieved if using more than 3 or more suitable features.

## CONCLUSIONS

We described a machine-learning procedure for classification of SEM images of cross-laminated sandstones for digital rock analysis. This procedure can be automated to achieve a robust classification and to address major issues that hinder the use of high-resolution SEM images for calibrating low resolution images as needed in the multi-scale image integration. We showed that the procedure can lead to a successful classification for a tight-sand SEM image where lamina patterns can be reasonably distinguished on a thin section. This work shows that the multi-scale data integration of digital rock analysis could be extended to larger and structurally more complex samples.

## ACKNOWLEDGEMENTS

We thank Jim O. Buckman, Andy R. Gardiner, and Gary D. Couples, at Institute of Petroleum Engineering, for carrying out SEM imaging, providing geological guidance and helping shape this manuscript, respectively, and James J. Howard, at ConocoPhillips, for his useful comments that helped improve the quality of this manuscript.

## REFERENCES

[1] G. Pickup, K. Stephen, J. Ma, P. Zhang, and J. Clark, "Multi-stage upscaling: Selection of suitable methods," *Transport in porous media,* vol. 58, pp. 191-216, 2005.

[2] S. R. McDougall and K. S. Sorbie, "The Combined Effect of Capillary and Viscous Forces on Waterflood Displacement Efficiency in Finely Laminated Porous Media," 1993.

[3] S. J. Latham, A. P. Sheppard, M. A. Knackstedt, R. M. Sok, and M. Kumar, "Rock Typing Across Disciplines," 2010.

[4] E. D. Pittman, "Diagenesis of quartz in sandstones as revealed by scanning electron microscopy," *Journal of Sedimentary Research,* vol. 42, 1972.

[5] C. S. Hutchison, "Laboratory handbook of petrographic techniques," 1974.

[6] E. J. Candes and D. L. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," DTIC Document2000.

[7] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, and B. Schmid, "Fiji: an open-source platform for biological-image analysis," *Nature methods,* vol. 9, pp. 676-682, 2012.

[8] A. Bjorlykke and D. Sangster, "An overview of sandstone lead deposits and their relation to red-bed copper and carbonate-hosted lead-zinc deposits," *Econ. Geol,* vol. 75, pp. 179-213, 1981.

[9] D. C. Greig and J. Pringle, *British regional geology: the south of Scotland*: HMSO, 1971.