Introduction

In the last decades, many works have investigated the use of texture features to support seismic image retrieval applications (Alfarraj et al. 2016). Among the considered texture features are GLCM, LBP, Gabor, wavelet and curvelet, and each of them may be the best option depending on the task and dataset at hand. Our previous work (Ferreira et al. 2019) built on top of these works to investigate the impact of considering a point set distance measure instead of the classical rescaling approach for a retrieval application. The result indicated that for the retrieval of images with different sizes, the point set distance measure outperformed the rescaling approach.

In this work, we turn our attention to a slightly different problem. Instead of a retrieval application, we consider a scenario in which we have many seismic datasets and we want to run a quick in situ analysis to have an idea of the diversity in the datasets and to identify which ones are similar to one another. Point set distance measures are an interesting fit not only because in situ datasets may have arbitrarily different sizes, but also because the rescaling approach cause distortions in the patterns and structures found in seismic images, hindering the analysis.

Methodology

We evaluate 6 point set distance measures for the analysis of 2D seismic datasets. First, we extract texture features, then we compute a point set distance measure to evaluate the distance between the 2D seismic datasets. These distances are then used to cluster the datasets and the resulting grouping is compared to the original groups (surveys) the seismic datasets belong to. We investigate not only the clustering but also the computational performance of the considered measures. In this section, we present the texture features, point set distance measures, clustering and evaluation algorithms used in this work.

Texture Features

In this section we present the two texture extraction algorithms used in our work. Before texture computation, the 2D seismic images are divided in tiles and the algorithms are applied to each tile. The first algorithm, LBP (Local Binary Pattern) (Ojala et al. 2002), is a technique that encodes the information about the neighborhood of a pixel into a binary number. Given a radius $R \geq 1$ and a number of neighbors $P$, the neighborhood for pixel $p$ is defined as $N = \{p_0, \ldots, p_{P-1}\}$ where all pixels $p_i$ are $R$ distant from $p$. The code for pixel $p$ is defined as a binary number where the $i^{th}$ digit is $1$ if $p_i \geq p$ and $0$ otherwise. More formally:

$$LBP_{p,R} = \sum_{i=0}^{P-1} s(p_i - p) \cdot 2^i, \quad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$ \hspace{1cm} (1)

We divide the input image in tiles and apply the LBP algorithm to each tile separately. The final feature vector of each tile is the normalized histogram of the LBP codes with $2^P$ elements. In the end, these histograms are concatenated to form the final descriptor of the image. For this work, we considered $P = 8$ and $R = 1$.

GLCM (Grey Level Co-occurrence Matrix) is a classical class of texture attributes originally proposed by (Haralick et al. 1973). For each analysis window, the algorithm creates a set of matrices $L \times L$, where $L$ is the number of grey levels in the image. Each matrix $C_{\theta \delta} [i,j]$ is defined by two parameters, the angle $\theta$ and the distance $d$. Common values for the angles are $0^\circ$, $45^\circ$, $90^\circ$ and $135^\circ$. Each element $(i,j)$ in these matrices represent the number of occurrences of grey level $i$ adjacent to grey level $j$ separated by distance $d$ in direction $\theta$. These matrices are usually normalized so that the sum of its elements is equal to 1, which can be regarded as joint probability matrices $P_{\theta \delta} [i,j]$. Based on (Ferreira et al. 2019), we consider 4 GLCM-based features (Equations 2-5) and we use $d = 1$ and all 4 angles, producing 16 GLCM-based texture attributes.
\[
\text{entropy} = -\sum_{i} \sum_{j} P_{ij} \log P_{ij} \\
\text{energy} = \left[ \sum_{i} \sum_{j} p_{ij}^{2} \right]^{\frac{1}{2}}
\]

(2) \hspace{2cm} (3)

\[
\text{correlation} = \sum_{i} \sum_{j} \frac{(i - \mu_i)(j - \mu_j)p_{ij}}{\sigma_i \sigma_j}
\]

(4)

\[
\text{maximum probability} = \max(P)
\]

(5)

**Point Set Distance Measures**

In our methodology, we investigate 6 point set distance measures (Equations 6-11) (Sherif and Ngomo 2018). The input for the distances consists of two point sets \(s = (s_1, ..., s_n)\) and \(t = (t_1, ..., t_m)\), where \(n\) and \(m\) represent the number of points in each set. The points in a set represent the tiles in an image, where the tiles are described by their corresponding texture feature vectors. Hence, when we discuss the comparison of point sets in this section, we are ultimately referring to tile-based image comparison. For all measures we use the symbol \(\delta\) to represent the Euclidean distance. As a baseline, we also consider the approach of resizing the images to a common size and computing the Euclidean distance of the point sets (pair-wise).

\[
D_{\text{mean}}(s, t) = \delta \left( \frac{\sum_{i \in s} s_i}{n}, \frac{\sum_{i \in t} t_i}{m} \right)
\]

Mean distance

(6)

\[
D_{\text{max}}(s, t) = \max_{s_i \in s, t_j \in t} \delta(s_i, t_j)
\]

Max distance

(7)

\[
D_{\text{min}}(s, t) = \min_{s_i \in s, t_j \in t} \delta(s_i, t_j)
\]

Min distance

(8)

\[
D_{\text{avg}}(s, t) = \frac{1}{nm} \sum_{s_i \in s, t_j \in t} \delta(s_i, t_j)
\]

Average distance

(9)

\[
D_{\text{som}}(s, t) = \frac{1}{2} \left( \sum_{s_i \in s} \min_{t_j \in t} \delta(s_i, t_j) + \sum_{t_j \in t} \min_{s_i \in s} \delta(t_j, s_i) \right)
\]

Sum of minimums

(10)

\[
D_{\text{hausdorff}}(s, t) = \max \left( \frac{1}{n} \sum_{s_i \in s} \min_{t_j \in t} \delta(s_i, t_j), \frac{1}{m} \sum_{t_j \in t} \min_{s_i \in s} \delta(t_j, s_i) \right)
\]

Hausdorff distance

(11)

**Clustering and Evaluation Algorithms**

We use a hierarchical clustering method that seeks to build a hierarchy of clusters (Müllner 2011). As opposed to the divisive type which has a top-down approach, in the agglomerative type each observation starts in its own cluster and pairs of clusters are merged as we move up in the hierarchy (bottom-up). Instead of working on the feature vectors, **Agglomerative Clustering** can also take a distance matrix as input. This way, we pre-compute the distances using the considered distance measures to feed the algorithm. **V-Measure** is an external entropy-based cluster evaluation measure (Rosenberg and Hirschberg 2007). It represents the harmonic mean between two cluster aspects: homogeneity and completeness. **Homogeneity** measures whether each cluster has only data points which are member of a single reference class. **Completeness** measures if all data points that belong to a given class are elements of the same cluster.

\[
v = \frac{2 \times (\text{homogeneity} \times \text{completeness})}{(\text{homogeneity} + \text{completeness})}
\]

(12)

**Databases**

Our experiments were performed on 12 2D seismic surveys acquired in the Taranaki Basin, New Zealand\(^1\). The Taranaki Basin presents an area of about 330,000 km\(^2\), located mostly offshore but also onshore the New Zealand North Island west coast. The Taranaki Basin was developed in a complex tectonic setting since the Cretaceous, combining structural styles related to passive and active margins. The surveys are presented in Table I, which comprise 304 2D seismic datasets. The dimensions of the seismic images range from 350 to 3500 pixels, both vertically and horizontally.

\(^1\)https://www.nzpam.govt.nz/maps-geoscience/petroleum-datapack/

---

82nd EAGE Conference & Exhibition 2020
8-11 June 2020, Amsterdam, The Netherlands
Table I. 2D seismic surveys.

<table>
<thead>
<tr>
<th>Survey</th>
<th>#Datasets</th>
<th>Survey</th>
<th>#Datasets</th>
<th>Survey</th>
<th>#Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar89</td>
<td>24</td>
<td>malvern</td>
<td>18</td>
<td>sunz-91</td>
<td>66</td>
</tr>
<tr>
<td>cen05</td>
<td>20</td>
<td>og04</td>
<td>26</td>
<td>tikati</td>
<td>16</td>
</tr>
<tr>
<td>hzt82a</td>
<td>46</td>
<td>omv05</td>
<td>25</td>
<td>tu89</td>
<td>19</td>
</tr>
<tr>
<td>kaheru</td>
<td>12</td>
<td>hf</td>
<td>16</td>
<td>west cape</td>
<td>16</td>
</tr>
</tbody>
</table>

Figure 1. Subsets of seismic images in the surveys.

Figure 1 shows subsets of the seismic images in each seismic survey. The surveys can be classified by the tectonic structures presented, for example, as normal and thrust faults formed during the compressional and extensional phases of the basin. According to the tectonic aspect, there is a distinct division between the surveys less tectonically disturbed (cen05, kaheru, malvern, omv05, sunz-91, tikati and west cape) and the ones more tectonically disturbed (ar89, hf, hzt82a, og04 and tu89). In addition, most surveys are characterized by the predominance of seismic reflectors with moderate amplitude, although, some are highlighted by an expressive amount of noise near higher depths (ar89, kaheru, og04) or in different regions of the record (malvern and tu89). We expect the texture features, combined with point set distance measures, to be able to discriminate the different surveys based on their visual characteristics.

Experiments

In this section we discuss the results obtained for each texture feature and point set distance measure. To extract the texture features we used a tile size of 300x300 pixels. This tile size was chosen to prevent the generation of too many tiles which could slow down the point set computations. Besides that, it allowed the seismic images to be kept in their original sizes. For the baseline, we resized the seismic images to 900x900 pixels, considering a 3x3 matrix of tiles. The images were rescaled, clipping the original intensities in 1% in both sides of the histogram. After that, they were rescaled to 64 grey levels.

We computed the distance matrices using the baseline and the point set distance measures and fed them to the Agglomerative Clustering algorithm. After that we compared the original groups (surveys) to the ones generated by the clustering algorithm using the V-Measure. Figure 2 (a) shows that the best result was obtained with the min distance for the LBP texture feature with a V-Measure of 0.81. The worst two results were obtained with max and average distance measures which did not seem to help differentiate the seismic datasets. The best execution times (Figure 2 (b)) were obtained by the baseline and mean distance measures (<150s) which do not involve many-to-many computations. The min, max and average distance measures took much more time (~30min), since they involve comparing the point sets unidirectionally. Lastly, sum of minimums and Hausdorff distance measures were the slowest ones (~1h) due to reciprocal point set computations.

The fact that the LBP texture features performed better than GLCM in our experiments is not conclusive since a parameter selection was not carried out. Our main objective was to evaluate the point set distance measures and a fixed parametrization for the texture features was used. Nevertheless, the fact that LBP
has fewer parameters and that the classical parametrization produces such an interesting result is a factor that could favor LBP over GLCM.

![Figure 2](image.png)

**Figure 2.** Clustering (a) and execution (b) performances for all texture features and point set distance measures.

Taking into consideration the tradeoff between clustering and execution performances, the min distance measure seems to be an interesting choice if one wants to run a quick analysis on a recently acquired database to have an idea of the diversity in the datasets and to identify which ones are similar to one another. The experiments also show that taking a point set distance approach, we were able to outperform the approach of resizing the images to a common size for comparison, which distorts the patterns and structures found in a seismic dataset.

**Conclusions**

In this work we investigated 6 point set distance measures for 2D seismic analysis. The results obtained with 12 seismic surveys comprising 304 seismic datasets indicate that point set distance measures may be more suitable for the comparison of *in situ* datasets, since they appear in different resolutions and sizes. The experiments show a gain of 53% of the min distance measure in comparison to the classical rescaling approach. Although the results are promising, the authors would like to evaluate the proposed methodology on other datasets and to perform a more detailed assessment of the impact of the different parameters.

**References**


