

Structure Label Prediction Using Similarity-Based Retrieval and Weakly-Supervised Label Mapping

Yazeed Alaudah^{*†}, Motaz Alfarraj^{*†}, and Ghassan AlRegib^{*}

ABSTRACT

Recently, there has been significant interest in various supervised machine learning techniques that can help reduce the time and effort consumed by manual interpretation workflows. However, most successful supervised machine learning algorithms require huge amounts of annotated training data. Obtaining these labels for large seismic volumes is a very time-consuming and laborious task. We address this problem by presenting a weakly-supervised approach for predicting the labels of various seismic structures. By having an interpreter select a very small number of exemplar images for every class of subsurface structures, we use a novel similarity-based retrieval technique to extract thousands of images that contain similar subsurface structures from the seismic volume. By assuming that similar images belong to the same class, we obtain thousands of image-level labels for these images; we validate this assumption in our results section. We then introduce a novel weakly-supervised algorithm for mapping these rough image-level labels into more accurate pixel-level labels that localize the different subsurface structures within the image. This approach dramatically simplifies the process of obtaining labeled data for training supervised machine learning algorithms on seismic interpretation tasks. Using our method we generate thousands of automatically-labeled images from the Netherlands Offshore F3 block with reasonably accurate pixel-level labels. We believe this work will allow for more advances in machine learning-enabled seismic interpretation.

INTRODUCTION

In recent years, there has been a significant interest in machine learning-based techniques for various seismic interpretation applications such as salt body delineation, fault and fracture detection, horizon extraction, and facies classification (e.g., Coléou et al., 2003; Barnes and Laughlin, 2005; Wang et al., 2015a; Guillen et al., 2015; Zhao et al., 2015; Wang et al., 2015b; Figueiredo et al., 2015; Qi et al., 2016; Ramirez et al., 2016; Lin et al., 2017).

Supervised machine learning has proven to be one of the most successful machine learning paradigms. By definition, supervised machine learning techniques require labels to perform training. However obtaining labels for large volumes of seismic data is a very demanding task, especially when the size of the data is in the tens or hundreds of gigabytes. Furthermore, while the amount of data is growing continuously, the ability of human experts to manually label large quantities of data is insufficient. Therefore, the time and effort required to manually annotate large amounts of training data can often exceed the time and effort *saved* by techniques based on automated machine learning workflows. This is where, we believe, *similarity-based retrieval* and *weakly-supervised learning* can play a significant role in overcoming this problem.

Similarity-based retrieval is a widely used technique in content-based image retrieval (CBIR) applications (Zhou et al., 2017). It utilizes metrics designed to measure the similarity of two images, to retrieve visually similar images from a large database of images. On the other hand, weakly-supervised learning is a machine learning paradigm where the model is trained using examples that are only partially annotated or labeled (Torresani, 2014). Figure 1 explains this further. Assume we would like to train a simple machine learning model to classify salt bodies. The model takes an input image similar to the one in the figure and produces the desired output shown, where red denotes salt body, and cyan denotes everything else. To train a *fully*-supervised model, we would need training la-

^{*} {alaudah,motaz,alregib}@gatech.edu

Center for Energy and Geo Processing (CeGP) at Georgia Tech and King Fahd University of Petroleum & Minerals (KFUPM)

[†] Electrical Engineering Department, King Fahd University of Petroleum & Minerals (KFUPM)

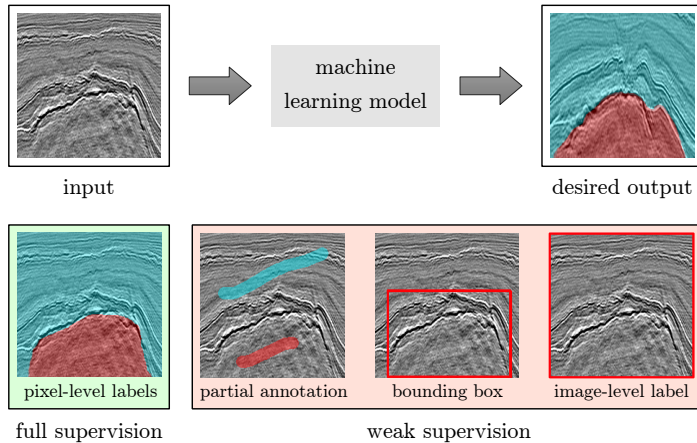


Figure 1: The difference between full supervision and weak supervision. At the bottom row from left to right: pixel-level labels, partial labels, bounding box, and an image level label. Red denotes salt body, and cyan indicates everything else.

bels that are *fully* annotated. In this case, this means we need pixel-level labels for all pixels in the training images. If the interpreter only partially labeled the training images, or provided a bounding box, or just provided an image-level label indicating whether the training image contained a salt body or not, then our trained machine learning model would be a *weakly*-supervised one, and our labels are considered *weak* labels.

In this work, we show that it is possible to *predict* thousands of high-quality pixel-level labels for training supervised machine learning models for seismic interpretation with very minimal input from the interpreter. As few as one or two exemplar images are required for each subsurface structure of interest such as faults or salt domes. This is achieved using an unsupervised similarity-based seismic image retrieval technique to extract thousands of images with visually similar subsurface structures. A weakly-supervised matrix-factorization based technique is then used to learn the common structures and features between these images, and then map their image-level labels into pixel-level labels that can be effectively used to train powerful fully supervised machine learning models.

Preliminary elements of this work were published in (Alfarraj et al., 2016; Alaudah et al., 2017; Alaudah and AlRegib, 2017). To our knowledge, other than this work, no other work attempts to address this issue in the field of seismic interpretation. Nevertheless, we overview the related literature in similarity-based retrieval and weakly-supervised labeling in the following subsections.

Similarity-Based Retrieval

Image similarity measures are functions that quantify the similarity between a pair of images. The definition of image similarity can vary depending on the application. Generic similarity or distance measures, such as peak signal-

to-noise ratio (PSNR) and Euclidean distance, assume a pixel-to-pixel correspondence between images. These measures treat every pixel independently. Structural similarity (SSIM) is a similarity measure that improves upon the pixel-to-pixel metrics by capturing local image structure using low-level local statistics in the spatial domain (S-SSIM) (Wang et al., 2004) or the complex wavelet domain (CW-SSIM) (Wang and Simoncelli, 2005). These measures are often used for applications such as image denoising where the pixel-to-pixel correspondence is justified. Another class of similarity measures is content-based similarity measures which quantify the similarity between the contents of the images without making any assumptions about the correspondence of the locations on the content. Such measures are often used for applications like content-based image retrieval in which the goal is to find images that contain similar visual content to a query image. A particular class of content-based similarity measures is texture-based similarity measures which compare the textured content of images using texture analysis. A well-known example of such measures is the structural texture similarity measure (STSIM) (Zujovic et al., 2013) which uses subband statistics and correlations in a multiscale frequency decomposition, namely, the steerable pyramid. Another example of such multiscale decomposition is the curvelet transform (Candes and Donoho, 2004a). It provides an efficient way of representing images with high directional content. Candes and Donoho (2004a) have shown that images that contain geometrically regular edges are more compactly represented by a curvelet rather than a wavelet decomposition. This is especially true for seismic data, where the wavefronts lie mainly along smooth curves. A few texture-based similarity measures based on curvelet coefficients have been proposed in the literature for texture and seismic image similarity. Al-Marzouqi and AlRegib (2014) proposed a seismic image similarity measure based on adaptive curvelets. Long et al. (2015) proposed combining the texture similarity metric in (Zujovic et al., 2013) with seismic discontinuity maps to evaluate the similarity of seismic images. Alaudah and AlRegib (2015) then proposed a method based on histograms of curvelet coefficients, while Alfarraj et al. (2016) proposed a method based on the singular values of curvelet coefficients, truncated using the effective rank (Roy and Vetterli, 2007).

Weakly-Supervised Labeling

The vast majority of the visual data nowadays is unlabeled or weakly labeled. Lately, there has been considerable interest in weakly-supervised methods for labeling natural images. Several weakly-supervised methods proposed recently are based on matrix factorization techniques. For example, Hong et al. (2016) proposed a framework for clustering images retrieved from a reference dataset using sparse and orthogonal non-negative matrix factorization. However, this method was proposed for image clustering and does not apply to pixel-level label-

ing. Other researchers have considered techniques based on non-negative matrix factorization to infer the pixel-level labels of pre-segmented regions (known as superpixels) within different images (Niu et al., 2015; Lu et al., 2017). Furthermore, Zhang and Gong (2016) proposed a non-negative matrix co-factorization based approach that jointly learns a discriminative dictionary and a linear classifier that classifies features from segmented images into different classes. These methods are all applied on *features* extracted from superpixels and not the actual pixels in the image. This considerably limits the resolution of the resulting labels. Furthermore, most of these approaches assume or require that the image is segmented into semantic regions that represent different classes. This is not the case in our work.

More recently, driven by the success of convolutional neural networks (CNNs) in various computer vision tasks, many researchers have proposed weakly-supervised labeling methods based on CNNs (Papandreou et al., 2015; Pinheiro and Collobert, 2015; Hong et al., 2015; Zhang et al., 2015; Pathak et al., 2015; Hou et al., 2016; Kim and Hwang, 2016; Wei et al., 2017; Kim et al., 2017). These methods, however, require a significant amount of training data to be effective, and usually do not localize various objects accurately without requiring other post-processing steps. Our proposed approach, based on non-negative matrix factorization, does not require any training or post-processing. In fact, very few exemplar images per class are required for our approach to give satisfactory results. As few as one or two exemplars per class can be sufficient to obtain a large number of image-level labeled data using similarity-based retrieval. These image-level labels that we obtain are then directly mapped into pixel-level labels without using any segmentation or any feature extraction technique.

In summary, our main contributions in this work are as follows:

- We introduce a similarity-based image retrieval framework to extract large numbers of images from a 3D seismic volume with similar subsurface structures to exemplar images chosen by an interpreter.
- We propose a weakly-supervised algorithm, based on non-negative matrix factorization, to learn a mapping from image-level labels into pixel-level labels.
- We apply this approach to images extracted from the Netherlands Offshore F3 block and generate thousands of pixel-level labeled images for three classes of subsurface structures: chaotic layers, faults, and salt domes.

The rest of the paper is organized as follows: first, we introduce the proposed approach including the similarity-based retrieval and the weakly-supervised label mapping. Then, we show various results that illustrate the effectiveness of our similarity-based retrieval method, as well as the weakly-supervised label mapping. We then discuss

these results and how our method can be improved, and finally, conclude the paper.

PROPOSED METHOD

Overall Workflow

The overall workflow of the proposed method is shown in Figure 2. First, the interpreter hand-selects a few images $\mathbf{X}_e = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_e}]$ such as those in Figure 3, to exemplify each class of subsurface structures. These can be as few as one image per class if the visual features of that class are uncomplicated. A similarity-based retrieval method is then used to extract a very large number of images that contain similar subsurface structures. We assume that most-similar images to the exemplar belong to the same class as that of the exemplar image. We validate this assumption later in the results section. At this stage, all these images have image-level labels. A weakly-supervised mapping, based on non-negative matrix factorization, is then used to map these image-level labels into pixels to obtain the final results. A high-level workflow of the proposed approach is depicted in Figure 2. The remaining parts of this section explain the similarity-based retrieval and the weakly-supervised label mapping modules in detail.

Similarity-Based Retrieval

Image similarity measures are used to evaluate the similarity of the visual content between two images. They take two images as input and return a value, often in the range $[0, 1]$, that indicates the level of similarity between the two images. A higher value indicates higher similarity. These measures are often used to search for images within large visual datasets. Here, we present the similarity measures we use to capture the similarity of seismic images (Alfarraj et al., 2016) and then describe how this metric is used to retrieve images that contain similar subsurface structures. This similarity measure works by computing the similarity between two vectors containing the singular values of the curvelet decomposition of the two images, trimmed adaptively using effective rank approximation. We will explain this in detail in the following subsections.

The Curvelet Transform

The curvelet transform is a directional multiscale decomposition first introduced by Candes *et al.* (Candes and Donoho, 2004b; Candes et al., 2005). It provides an efficient way of representing images with high directional content. Candes and Donoho (2004a) have shown that the curvelet transform provides an optimally sparse representation for curve-like structures, such as seismic reflectors, when compared to wavelets.

The curvelet transform works by taking the 2D fast Fourier transform of an image (2D FFT) and then dividing the plane into multiple scales and orientations as is shown in Figure 4(a). The total number of scales in the

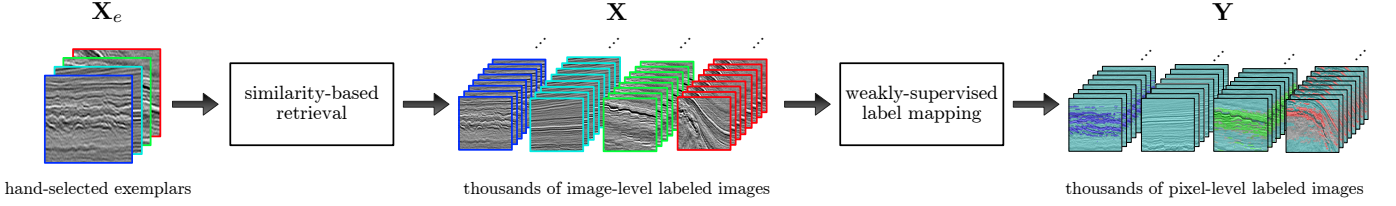


Figure 2: General workflow of the proposed method. Colored image boundaries indicate the image-level labels. Blue, cyan, green, and red denote the **chaotic**, **other**, **faults**, and **salt dome** classes respectively.

curvelet tiling, J , depends on the size on the image, and is given by

$$J = \lceil \log_2 \min(N_1, N_2) - 3 \rceil, \quad (1)$$

where N_1 and N_2 are the number of pixels in vertical and horizontal directions, respectively; and $\lceil \cdot \rceil$ is the ceiling function. The number of orientations at scale $j \geq 1$, $K(j)$, is given by:

$$K(j) = 16 \times 2^{\lceil (j-1)/2 \rceil}. \quad (2)$$

For scale $j = 0$, there is only one orientation. Curvelet coefficients are then generated by taking the inverse FFT for each wedge (such as the one highlighted in Figure 4) after multiplying it by a smooth band pass filter. Since the FFT of real images is symmetric around the origin, only two consecutive quadrants of the Fourier spectrum are necessary for obtaining the curvelet coefficients. Figure 4 shows the spatial and frequency representations of a curvelet wedge.

Feature Extraction and Similarity Measurement

To compute the similarity between two images, we first compute feature vectors for each image and then we arrive at the similarity value by comparing the corresponding feature vectors of two images. The feature vector of a given image is a collection of all effective singular values for all scales and orientations of the curvelet coefficients of an image. The overall workflow for computing the similarity is depicted in Figure 5.

To obtain a feature vector for a grayscale image, $\mathbf{x}_i$, we first apply the curvelet transform, and compute its coefficients for all scales, $j = \{1, 2, \dots, J\}$, and orientations, $k = \{1, 2, \dots, K(j)\}$. Then, the singular values of these curvelet coefficients are calculated as $\boldsymbol{\sigma}_{[j,k]} = [\sigma_1, \dots, \sigma_L]^T$ where $\sigma_1 \geq \sigma_2 \dots \geq \sigma_L$ and L is the smallest dimension of the coefficients matrix.

Ideally, if the rank of a matrix is r , only the first r singular values are non-zero. However, when we consider the singular value decomposition (SVD) on images that are subject to different types of noise, the number of non-zero singular values is greater than r . In most cases, none of the singular values are exactly zero; even for a rank-deficient matrix. Roy and Vetterli (2007) proposed the *effective rank* as a method to estimate the actual rank of a matrix by estimating its effective dimensionality. To calculate the effective rank, we first compute the normalized singular values as

$$p_i = \frac{\sigma_i}{\|\boldsymbol{\sigma}_{[j,k]}\|_1} \text{ for } i = 1, \dots, L, \quad (3)$$

where $\|\cdot\|_1$ is the ℓ_1 norm. Then, the effective rank is calculated as a function of the entropy of the singular value distribution defined in equation 3, that is

$$\text{EffectiveRank} = \exp \left(- \sum_{i=1}^L p_i \log p_i \right), \quad (4)$$

resulting in a real number less than or equal to L with equality if and only if all singular values are equal.

For each set of curvelet coefficients, the **EffectiveRank**, is calculated as in equation 4. A new vector of *effective* singular values is formed by keeping the first $\lfloor \text{EffectiveRank} \rfloor$ singular values, where $\lfloor \cdot \rfloor$ denotes the floor function. The remaining singular values are set to 0. In other words, for scale j and orientation k , we form the vector $\hat{\boldsymbol{\sigma}}_{[j,k]} = [\sigma_1, \dots, \sigma_{\lfloor \text{EffectiveRank} \rfloor}, 0, \dots, 0]$. The overall feature vector of image $\mathbf{x}_i$ is then obtained by concatenating all $\hat{\boldsymbol{\sigma}}_{[j,k]}$ for all scales and half the number of orientations,

$$\mathbf{v}_i = [\hat{\boldsymbol{\sigma}}_{[1,1]}, \hat{\boldsymbol{\sigma}}_{[2,1]}, \dots, \hat{\boldsymbol{\sigma}}_{[2,K(2)/2]}, \hat{\boldsymbol{\sigma}}_{[3,1]}, \dots, \hat{\boldsymbol{\sigma}}_{[J,1]}]. \quad (5)$$

The similarity between two images, $\mathbf{x}_1$ and $\mathbf{x}_2$, is then computed as

$$\text{Similarity}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\|\mathbf{v}_1 - \mathbf{v}_2\|_1}{\|\mathbf{v}_1 + \mathbf{v}_2\|_1}. \quad (6)$$

Where, $\mathbf{v}_1$ and $\mathbf{v}_2$ are the feature vectors corresponding to $\mathbf{x}_1$ and $\mathbf{x}_2$. Since the singular values are non-negative by definition, the resulting similarity value is in the range $[0, 1]$ with a value closer to 1 indicating higher similarity.

Similarity-Based Retrieval

By ranking the retrieved images based on their similarity to a reference image, we can extract thousands of images that resemble exemplar images chosen by an interpreter. In our work, we extract thousands of images extracted from random locations within the Netherlands Offshore F3 block (dGB Earth Sciences, 1987). We then use similarity-based retrieval to retrieve the top M images that are most similar to the exemplar images shown in 3. The value of M is selected by the interpreter such that the M retrieved images can be reasonably assumed to belong to the same class as the reference exemplar image. Later in the results section, we show that our approach is robust to wrongly-retrieved images.

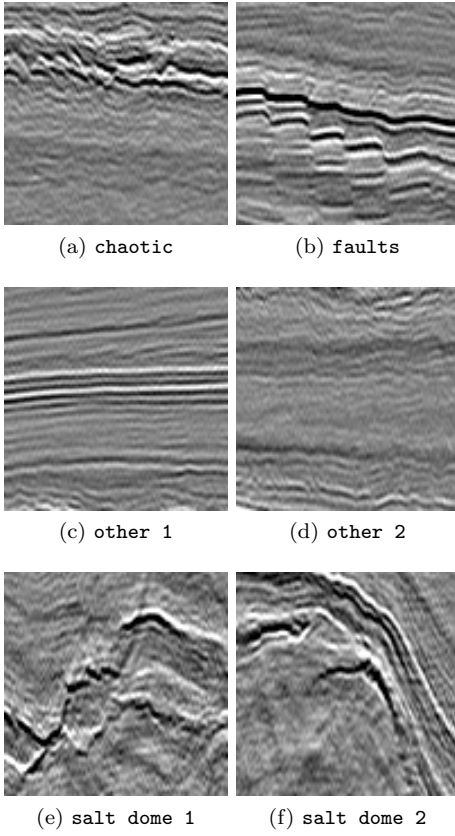


Figure 3: The exemplar images of each class of subsurface structures that were used to retrieve the images in this work. One exemplar image was used for **chaotic** and **fault**, and two exemplars were used for **other** and **salt dome**. These images are of size 99×99 pixels and were obtained from the Netherlands Offshore F3 Block (dGB Earth Sciences, 1987).

Weakly-Supervised Label Mapping

Once we have obtained image-level labels, we use a weakly-supervised learning approach, based on non-negative matrix factorization, to map these image-level labels into pixel-level labels. This mapping is a weakly supervised one since every image $\mathbf{x}_i \in \mathbb{R}_+^{n \times m}$ has only one label as opposed to $n \times m$ labels (i.e., one label per pixel). In the remainder of this section, we will describe this weakly-supervised label mapping in detail.

Given the image-level labeled images, $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_s}\}$, we vectorize them to construct the data matrix $\mathbf{X} \in \mathbb{R}_+^{N_p \times N_s}$ such that each image is a column in $\mathbf{X}$, where $N_p = n \times m$ is the number of pixels in each image, and N_s is the total number of images.

Non-Negative Matrix Factorization

Non-negative Matrix Factorization (NMF) (Paatero and Tapper, 1994; Lee and Seung, 1999) is a commonly used matrix factorization technique that is closely related to many unsupervised machine learning techniques such as

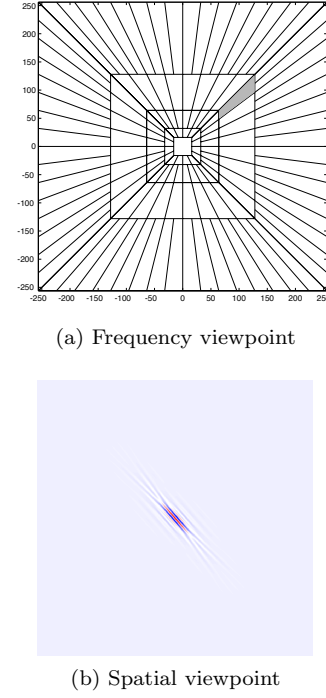


Figure 4: Frequency and spatial viewpoints of a curvelet wedge. Adapted from Candes et al. (2005). Copyright ©2006 Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved.

k -means and spectral clustering (Türkmen, 2015; Ding et al., 2005). NMF decomposes a non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{N_p \times N_s}$ into the product of two lower-rank matrices $\mathbf{W} \in \mathbb{R}_+^{N_p \times N_f}$, and $\mathbf{H} \in \mathbb{R}_+^{N_f \times N_s}$ such that both $\mathbf{W}$ and $\mathbf{H}$ are non-negative, and $N_f < \min(N_p, N_s)$. In other words we have,

$$\mathbf{X} \approx \mathbf{WH}. \quad (7)$$

In our work, the matrix $\mathbf{X}$ represents a data matrix where each column is a single image in vector form. The data matrix $\mathbf{X}$ has N_s such images, each of which is a vector of length N_p . Here, we use N_p , N_s , and N_f to denote the number of pixels, the number of samples, and the number of features (or the *rank* of $\mathbf{X}$) respectively. NMF factorizes the data matrix $\mathbf{X}$ into two non-negative matrices, a basis matrix $\mathbf{W}$ and a coefficient matrix $\mathbf{H}$. In clustering terms, the columns of $\mathbf{W}$ represent N_f number of clusters in the data, whereas the columns of $\mathbf{H}$ represent the memberships of each of the images to the different clusters in the data. Here, the clusters can represent different seismic structures like salt domes, faults, or horizons.

The regular NMF problem does not have a closed-form solution, and is typically solved by minimizing the following objective function

$$\arg \min_{\mathbf{W}, \mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 \quad \text{s.t. } \mathbf{W} \geq 0 \text{ and } \mathbf{H} \geq 0, \quad (8)$$

where, $\|\cdot\|_F$ is the Frobenius norm, and $\geq$ is an element-wise inequality. Lee and Seung (2001) proposed an efficient method of solving this problem using multiplicative

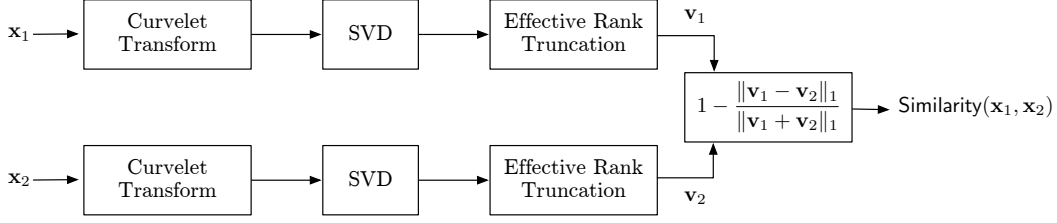


Figure 5: Workflow of the similarity measure proposed in Alfarraj et al. (2016).

update rules, and proved that they converge to a local minima.

Sparsity and Orthogonality Constraints

Lee and Seung (1999) showed that NMF can be used to learn a parts-based representation of the data, where each feature would represent a localized “part” of the data. In practice, this is rarely achieved using the formulation in equation 8. To remedy this, we initialize the feature matrix $\mathbf{W}^0$ using k -means separately on the different classes in the data matrix $\mathbf{X}$. This ensures that each feature $\mathbf{w}_i$ in the matrix $\mathbf{W}$ corresponds to a single class. We then impose a sparsity constraint on these initial features such that the sparsity of every feature $\mathbf{w}_i$ in matrix $\mathbf{W}^0$ satisfies

$$\rho(\mathbf{w}_i) = \frac{\sqrt{N_p} - \frac{\|\mathbf{w}_i\|_1}{\|\mathbf{w}_i\|_2}}{\sqrt{N_p} - 1}, \quad (9)$$

where $\rho(\cdot)$ indicates the sparsity of a vector. This value is always between zero and one, with higher values indicating higher sparsity. To enforce this sparsity constraint, we follow the algorithm proposed by Hoyer (2004). Additionally, we would expect that our features in $\mathbf{W}$ only represent a few images. In other words, it is unlikely that the same feature will be present in a large number of the images across different classes. To enforce this expectation, we impose an orthogonality constraint on the coefficients matrix $\mathbf{H}$. We also add two regularization terms on $\mathbf{W}$ and $\mathbf{H}$. The problem then becomes

$$\begin{aligned} \arg \min_{\mathbf{W}, \mathbf{H}} & \|\mathbf{X} - \mathbf{WH}\|_F^2 + \gamma \|\mathbf{HH}^T - \mathbf{I}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 \\ & + \lambda_2 \|\mathbf{H}\|_F^2 \quad \text{s.t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \text{ and } \rho(\mathbf{w}_i) = \rho_w, \end{aligned} \quad (10)$$

where matrix $\mathbf{I} \in \mathbb{R}_+^{N_f \times N_f}$ is an identity matrix. The values γ , λ_1 , and λ_2 are regularization constants, and ρ_w is the desired sparsity level.

Multiplicative Update Rules

Instead of solving the problem in equation 10 for both $\mathbf{W}$ and $\mathbf{H}$, we decouple this problem into two separate sub-problems. The first,

$$\arg \min_{\mathbf{W}} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 \quad \text{s.t. } \mathbf{W} \geq 0, \rho(\mathbf{w}_i) = \rho_w, \quad (11)$$

is solved for $\mathbf{W}$ while $\mathbf{H}$ is held constant. Then the second,
$$\arg \min_{\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \gamma \|\mathbf{HH}^T - \mathbf{I}\|_F^2 + \lambda_2 \|\mathbf{H}\|_F^2 \quad \text{s.t. } \mathbf{H} \geq 0, \quad (12)$$

is solved for $\mathbf{H}$ while $\mathbf{W}$ is held constant. We use gradient descent to derive the following multiplicative update rules for $\mathbf{W}$:

$$\mathbf{W}^{t+1} = \frac{\mathbf{W}^t \odot (\mathbf{XH}^{tT})_{ij}}{(\mathbf{W}^t \mathbf{H}^t \mathbf{H}^{tT} + \lambda_1 \mathbf{W}^t)_{ij}}, \quad (13)$$

and for $\mathbf{H}$:

$$\mathbf{H}^{t+1} = \frac{\mathbf{H}^t \odot (\mathbf{W}^{t+1T} \mathbf{X} + \gamma \mathbf{H}^t)_{ij}}{(\mathbf{W}^{t+1T} \mathbf{W}^{t+1} \mathbf{H}^t + \lambda_2 \mathbf{H}^t + \gamma \mathbf{H}^t \mathbf{H}^{tT} \mathbf{H}^t)_{ij}}. \quad (14)$$

Here, $\odot$ represents element-wise multiplication, and the superscript indicates the iteration number. These multiplicative update rules (MURs) are applied successively until both $\mathbf{W}$ and $\mathbf{H}$ converge. As we show in Appendix B, these multiplicative update rules are a special case of gradient descent with an automatic step size selection. One advantage of using these MURs is the guaranteed non-negativity of $\mathbf{W}$ and $\mathbf{H}$ when they are initialized with non-negative values. A detailed derivation of these MURs is shown in Appendix B.

Extracting the Labels

Once $\mathbf{W}$ and $\mathbf{H}$ have converged, each column of $\mathbf{H}$, $\mathbf{h}_n$, indicates the features used to construct the n^{th} image. Since every feature in $\mathbf{W}$ should correspond to a single class, we can predict the label of each pixel in the image by knowing which features are used to represent it. In other words, we can map the coefficients in $\mathbf{h}_n$ to the seismic structures that make up the image. Thus for image $\mathbf{x}_n$ we can obtain

$$\mathbf{L}_n = \mathbf{W}(\mathbf{Q} \odot (\mathbf{h}_n \mathbf{1}^T)) \quad \forall n = [1, \dots, N_s], \quad (15)$$

where $\mathbf{1}$ is a column vector of ones of length N_l , and $\mathbf{Q} \in \{0, 1\}^{N_f \times N_l}$ is a cluster membership matrix such that the element $\mathbf{Q}(i, j) = 1$ if the feature $\mathbf{w}_i$ belongs to structure j . The matrix $\mathbf{Q}$ is used to encode our knowledge of the image-level labels, and how the matrix $\mathbf{W}$ was initialized. The resulting matrix, $\mathbf{L}_n \in \mathbb{R}_+^{N_p \times N_l}$ shows the likelihood of each seismic structure for each pixel in the image. Then, the pixel-level labels for image $\mathbf{x}_n$ correspond to the seismic structure given by

$$\mathbf{y}_n(i) = \arg \max_j \mathbf{L}_n(i, j) \quad \forall i = [1, \dots, N_p], \quad (16)$$

where $\mathbf{L}_n(i, j)$ denotes the element in the i^{th} row and j^{th} column of matrix $\mathbf{L}_n$. However, due to the nature of the weakly-supervised mapping of the labels, there is an element of uncertainty in the mapping. Since the features $\mathbf{w}_i$ are sparse, some pixels in an image $\mathbf{x}_n$ may not have a feature that accurately represents all the pixels within it. These pixels typically end up being represented as a weighted sum of a large number of different features, often from different classes and having small coefficients. This leads to noisy labeling results. To remedy this, we introduce a new **uncertain** class that contains pixels with uncertain labels. We define our confidence, $\mathbf{c}_n \in \mathbb{R}^{N_p}$, in the predicted label of every pixel in the image $\mathbf{x}_n$ as

$$\mathbf{c}_n(i) = \max_j \mathbf{L}_n(i, j) \quad \forall i = [1, \dots, N_p]. \quad (17)$$

We can then assign any pixel whose confidence is less than a threshold τ to the **uncertain** class, denoted as class 0

$$\mathbf{y}_n(\mathbf{c} < \tau) = 0. \quad (18)$$

Once we obtain the pixel-level labels $\mathbf{y}$ for each image, we apply a 3×3 median filter to clear any noisy labels and get the final labeling result for that image. We do this for all N_s images and concatenate the results to construct the pixel-level labels matrix $\mathbf{Y} \in \mathbb{Z}^{N_p \times N_s}$ that contains the final pixel-level labels for all the images in the data matrix

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{N_s}]. \quad (19)$$

RESULTS

We present our results in two subsections corresponding to the two main modules of our workflow shown in Figure 2. First, we show results that demonstrate the effectiveness of our similarity measure and validate our assumption that all retrieved images should have the same image-level label as the query image. Then we show sample results of the weakly-supervised label mapping and demonstrate its effectiveness in accurately localizing the various seismic structures we have used in this study.

Similarity-Based Retrieval

To evaluate the performance of our similarity measure compared to other measures, we devise two experiments. Namely, retrieval and clustering. We do both experiments directly on the similarity matrices of the various measures. These matrices contain the similarity values between all pairs of images in a dataset for a specific measure. For example, for a dataset that contains N_s images divided into N_l classes, the size of the similarity matrix $\mathbf{S}$ is $N_s \times N_s$, where $\mathbf{S}(i, j)$ is the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e. $\mathbf{S}(i, j) = \text{Similarity}(\mathbf{x}_i, \mathbf{x}_j)$. The i^{th} row of $\mathbf{S}$ represents the similarity values of all images in the dataset

compared to $\mathbf{x}_i$. Note that in these experiments we convert distance metrics, such as Euclidean distance and the method proposed by [Alaudah and AlRegib (2015)], to similarity measures by normalizing each row of $\mathbf{S}$ by the maximum value of that row and subtracting it from 1. Hence, the diagonal entries of $\mathbf{S}$ are all 1.

Throughout the similarity-based retrieval experiments, we use the LANDMASS-2 dataset* ([Center for Energy and Geo Processing, 2015]) which is comprised of 4000 images of size 99×99 pixels, with their values normalized to be between 0 and 1. These images were extracted from the Netherlands Offshore F3 block and divided equally into four classes according to their dominant structure. The classes are **horizon**, **chaotic**, **fault** and **salt dome**. In the following subsections, we will discuss the two experiments and their results in detail.

In our experiments, we compare the performance of the method we described ([Alfarraj et al., 2016]) to different similarity and distance measures. The following measures were used in the experiments:

1. Euclidean distance
2. CW-SSIM with default parameters ([Wang and Simoncelli, 2005])
3. STSIM-1 and STSIM-2 with 4 scales and 8 orientations ([Zujovic et al., 2013])
4. SeiSIM with 4 scales and 8 orientations ([Long et al., 2015])
5. Curvelet-based distance measure ([Alaudah and AlRegib, 2015])

Figure 6 shows the receiver operating characteristic (ROC) curves of the similarity measures listed above. The retrieval results are summarized in Table 1. The metrics used to evaluate the performance of different methods are detailed in Appendix A. The results show that the similarity measure proposed in ([Alfarraj et al., 2016]) is the best performing in all the different metrics we used.

To further analyze the retrieval results, we show in Figure 7 the Precision @ M curves for each of the four classes in the dataset using the best performing similarity measure, in addition to Precision @ M for the entire dataset. The cyan curve represents the **horizon** class, and it shows that the precision is 1 for all values of M . This is mainly due to the simplicity of structures that appear in **horizon** images. On the other hand, the curves of the other classes drop at different rates depending on the complexity of their structures. For example, images of **fault** class have a different number of faults of different scales and dipping angles, which makes them the most complicated class in the dataset. Hence, the **fault** precision curve drops at a faster rate than those of other classes. We show the curve for the combined images of all classes in black.

To further assess the performance of the similarity measures listed above on seismic data, we set up a clustering experiment using the similarity matrix obtained previously. First, the images in the dataset are projected into a 2-dimensional Euclidean subspace based on their

*<https://ghassanalregib.com/landmass/>

Table 1: Performance of different similarity metrics for similarity-based retrieval and clustering

Metric	Retrieval			Clustering
	RA	MAP	AUC	Rand Index
Euclidean distance	0.3450	0.3944	0.5145	0.3943
CW-SSIM(Wang et al., 2004)	0.7211	0.8062	0.8584	0.8697
STSIM-1(Zujovic et al., 2013)	0.8668	0.9255	0.9655	0.8954
STSIM-2(Zujovic et al., 2013)	0.8548	0.9104	0.9637	0.8772
SeiSIM(Long et al., 2015)	0.8192	0.8860	0.9450	0.8883
Alaudah and AlRegib (2015)	0.8962	0.9494	0.9776	0.9048
Alfarraj et al. (2016)	0.9105	0.9537	0.9830	0.9702

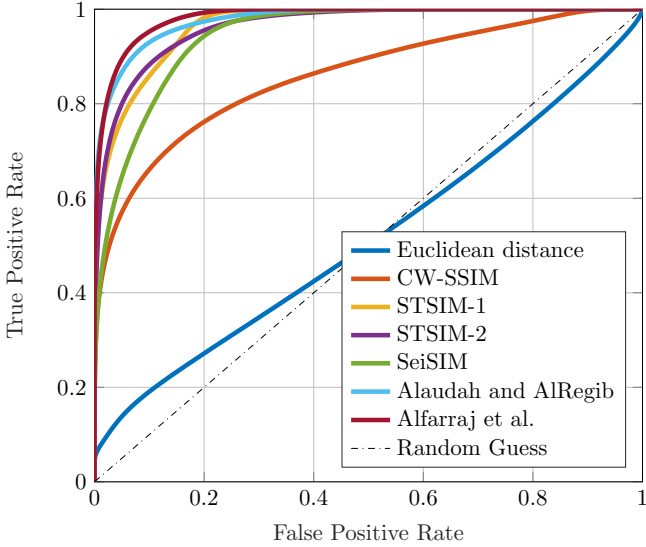
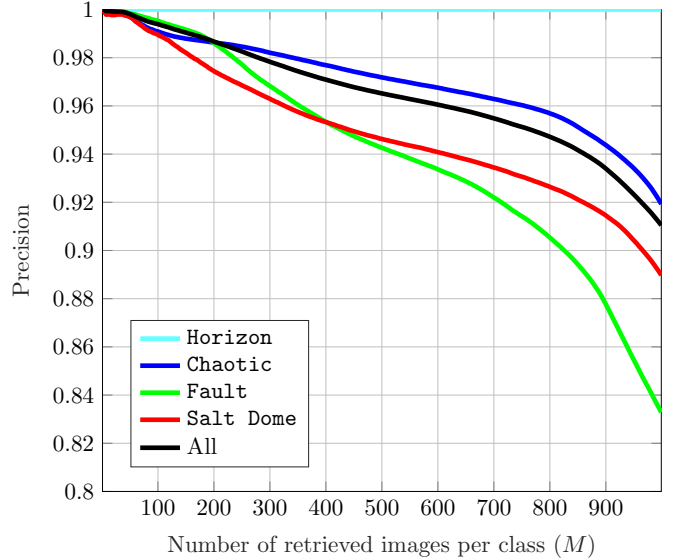


Figure 6: Receiver Operating Characteristics (ROC) curves for the various measures used in the retrieval experiments.

Figure 7: Precision at M curves for different classes of images shown for varying number of retrieved images.

similarity, such that the distance between images in the projection subspace is inversely proportional to their similarity values. The projection is done using classical multidimensional scaling (MDS) (Borg and Groenen, 2005). Then, the projected data points are clustered into four clusters using the k -means algorithm.

Since clustering is an unsupervised method, the clustering algorithm clusters the images according to their measured similarity. Therefore, the clusters do not necessarily correspond to classes; unless the similarity measure has a good discriminative power. Therefore, one can use the clustering results to quantify the goodness of the similarity measure. To evaluate the clustering performance, we compute the rand index (explained in Appendix A) which is a measure of the similarity between two data clusterings. We report the rand index results for different similarity measures in Table 1. The results of the clustering experiment further validate our conclusion from the retrieval experiment that our method is superior to other methods in the literature.

Also, we show the two-dimensional projection of the data using our method in Figure 8. The figure shows that

using the similarity values to project the dataset into a lower dimensional subspace produces clusters that are almost linearly separable. **Horizon** and **salt dome** classes are separated well from all other classes. However, the **fault** and **chaotic** classes slightly overlap. It is important to mention that this is only a 2D projection of the data and that the data is more easily separated in a higher-dimensional space. These results suggest that the similarity measure can be used to discriminate the different classes of seismic images with high accuracy.

Weakly-Supervised Label Mapping

We apply our similarity measure on the Netherlands Offshore F3 block (dGB Earth Sciences, 1987) to retrieve $M = 500$ images for each of the four classes, **chaotic**, **other**, **fault**, and **salt dome**. We put these images in vector form and use them to construct matrix $\mathbf{X}$. We then apply the k -means clustering algorithm on each class separately and use the results to initialize matrix $\mathbf{W}^0$ after we impose the sparsity constraint in equation 9. The coefficients matrix $\mathbf{H}^0$ is initialized with uniform random

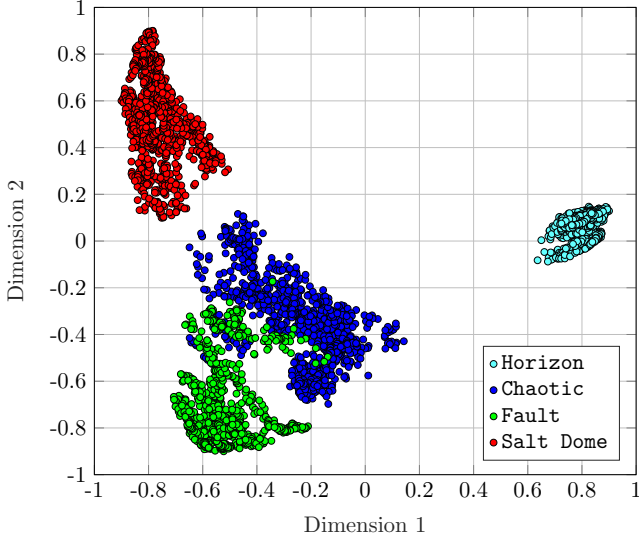


Figure 8: A two dimensional projection of the different images in the LANDMASS-2 dataset using multidimensional scaling with the similarity measure we proposed in (Alfarraj et al., 2016).

numbers in the range $[0, 1]$. We empirically select the values of λ_1, λ_2 and γ as 0.1, 0.5, and 5 respectively. The sparsity of the initial features ρ_w is set to 0.4. Additionally, the confidence threshold τ is set to 0.001.

Since our similarity-based retrieval workflow might produce a few images that do not belong to the same class as the reference image, we might end up with a few wrong image-level labels. However, the k -means initialization step of $\mathbf{W}^0$ greatly enhances the robustness of our label mapping algorithm to mislabeled images. To validate this claim, we examine the effect of wrongly retrieved images on the final pixel-level labels and analyze the robustness of our label mapping algorithm. We achieve this by artificially replacing images in $\mathbf{X}$ with wrongly-labeled images, and then computing the final pixel-level labels and comparing the performance of our label mapping algorithm relative to the base case where no images are replaced. The performance is evaluated using a metric called *pixel accuracy* that measures the percentage of all pixels that are correctly classified. Pixels with low confidence in the base case are ignored. Fig. 9 shows the drop in relative performance as the percentage of wrongly-labeled images in $\mathbf{X}$ increases for varying numbers of feature clusters per class, k . Fig. 10 shows a similar drop for different values of the feature sparsity ρ_w . Overall, the larger the number of clusters, and the higher the sparsity of the initial features, the more robust the label mapping algorithm is to wrongly retrieved images.

We then apply the MURs in equations 13 and 14 successively until both $\mathbf{W}$ and $\mathbf{H}$ converge. Figure 11 shows the convergence curves for the $\mathbf{W}$ objective function in equation 11, the $\mathbf{H}$ objective function in equation 12, and the overall objective function defined in equation 10. We see that although we did not attempt to solve equation 10

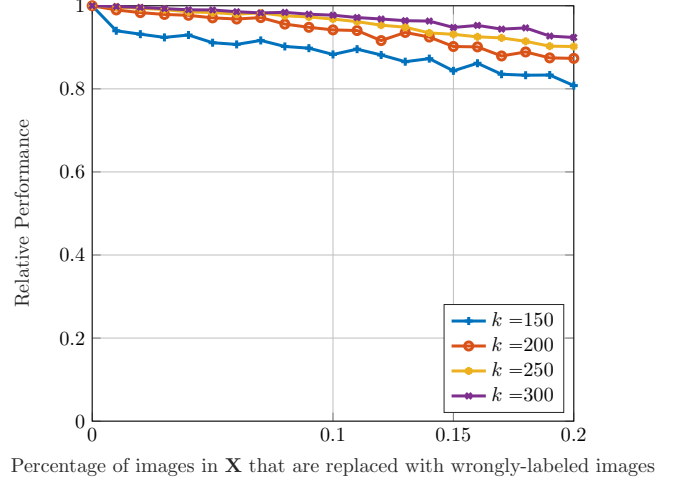


Figure 9: The robustness of our label mapping algorithm to mislabeled images for various numbers of feature clusters per class, k .

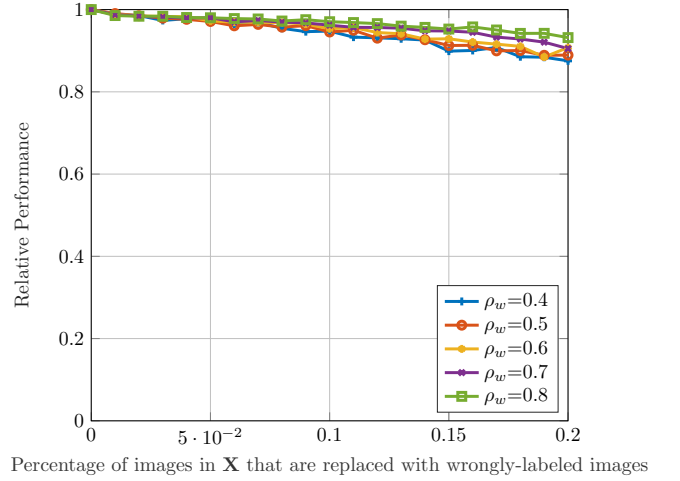


Figure 10: The robustness of our label mapping algorithm to mislabeled images for various feature sparsity levels, ρ_w .

directly, the two MURs in equations 13 and 14 effectively minimize the overall objective function. We terminate our optimization at the 200th iteration. Performed on a GPU, this process takes around 4 seconds.

Figure 12 shows the initial labels for four different images from the four different classes, as well as the labels for various iterations in the optimization process. We note that since the coefficient matrix $\mathbf{H}$ was initialized with random values in the range $[0, 1]$, our “initial” confidence computed using equation 17 is very high, and consequently, very few pixels in the initial labels had coefficient smaller than τ and therefore belonged to the **uncertain** class). However, immediately after we start applying the MURs the confidence value of the labels drastically drops, to the degree that all the pixels in images during the first iteration are labeled as **uncertain**. However, as the opti-

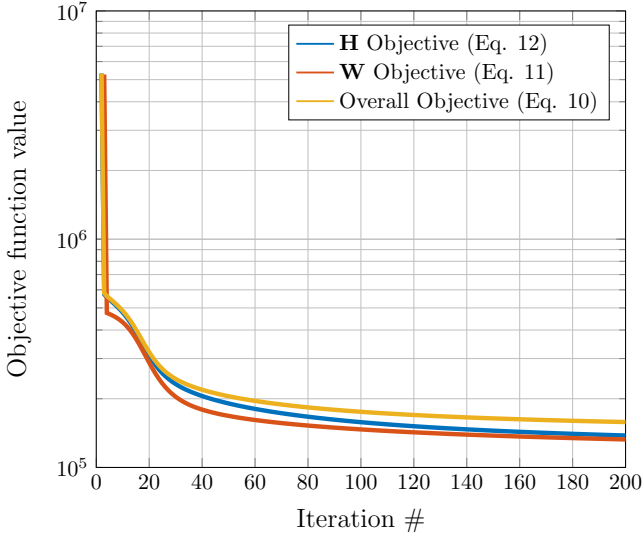


Figure 11: A plot showing the convergence curves for the multiplicative update rules for $\mathbf{W}$ and $\mathbf{H}$ as well as the overall objective function in equation 10.

mization progresses, confidence in various predicted labels gradually increases. Towards the end of the optimization, the orthogonality term in equation 10 plays a more prominent role in ensuring that most features in $\mathbf{W}$ represent only a few images in $\mathbf{X}$, this significantly reduces the number of noisy labels.

Figure 13 displays several examples selected at random from the final results that we obtain for various classes of subsurface structures. These results are shown for the baseline case using NMF only, the case where we use sparse initial features in $\mathbf{W}^0$, and our proposed formulation. We observe that the NMF results are very noisy. The results with sparse initial features are better, but they contain bands of misclassified pixels, typically in the center of the image. However, the results of our proposed formulation are far better at localizing the different subsurface structures, and they do not exhibit the same misclassified bands. Our results show a very good match between the labeled subsurface structures and the structures in the original seismic image. Furthermore, it is important to note that this method is not limited to these particular classes of subsurface structures and can be easily applied to any other structure as long as a sufficient number of similar images are retrieved for each class.

Finally, as with any other workflow, there are several areas where this approach can be improved. First, different classes of subsurface structures can often have different scales, whereas the method we have currently proposed uses a fixed size image for every class. It is worth investigating methods to alleviate this issue. Also, the final pixel-level labels are sensitive to the initial features, $\mathbf{W}^0$. While we have showed that k -means can easily be used to initialize $\mathbf{W}^0$, it is worth investigating other more promising methods such as convolutional neural networks. In addition, if the data matrix $\mathbf{X}$ has a wrong sparsity

structure, applying the sparsity constraint in equation 9 to form the feature matrix $\mathbf{W}$ might not lead to representative features of the different classes in $\mathbf{X}$. In that case, other techniques should be used to initialize $\mathbf{W}$. Finally, there are a few parameters such as the sparsity level ρ_w , the number of retrieved images per class M , and the regularization constants that need to be set by the interpreter based on her empirical assessment of the results.

CONCLUSION

In conclusion, we have proposed a method for predicting the labels of various seismic structures to enable the use of fully-supervised machine learning techniques for seismic interpretation. Our method uses a few exemplar images that are assigned image-level labels by an interpreter. Thousands of visually similar images that contain the same structures are then automatically retrieved. Furthermore, our weakly-supervised label mapping algorithm learns the joint structures in all these retrieved images and maps their image-level labels into pixel-level labels that are more suitable for fully supervised machine learning applications. We applied our method to data extracted from the Netherlands Offshore F3 block. Our results show that this approach accurately labels the locations of subsurface structures such as faults, chaotic horizons, and salt dome boundaries using at most two exemplar images for each class.

This approach can also be used to predict labels for facies classification and similar problems within seismic interpretation. We believe this approach can significantly reduce the time and effort needed to obtain quality labeled data for training supervised machine learning models for seismic interpretation tasks. We consider that this work, and others in the future, will help open the way for more machine learning advances in seismic interpretation, and structural interpretation in particular.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the Center for Energy and Geo Processing (CeGP) at Georgia Institute of Technology in Atlanta, USA, and King Fahd University of Petroleum and Minerals (KFUPM) in Dhahran, Saudi Arabia.

APPENDIX A

PERFORMANCE EVALUATION METRICS

We use several metrics to evaluate the performance of our method in the retrieval and clustering experiments mentioned in the results section. Here, we will explain these metrics in detail.

Retrieval Evaluation Metrics

To assess the retrieval performance of our method compared to other methods in the literature, we set each image $\mathbf{x}_i$ in the dataset as a query and retrieve the top images

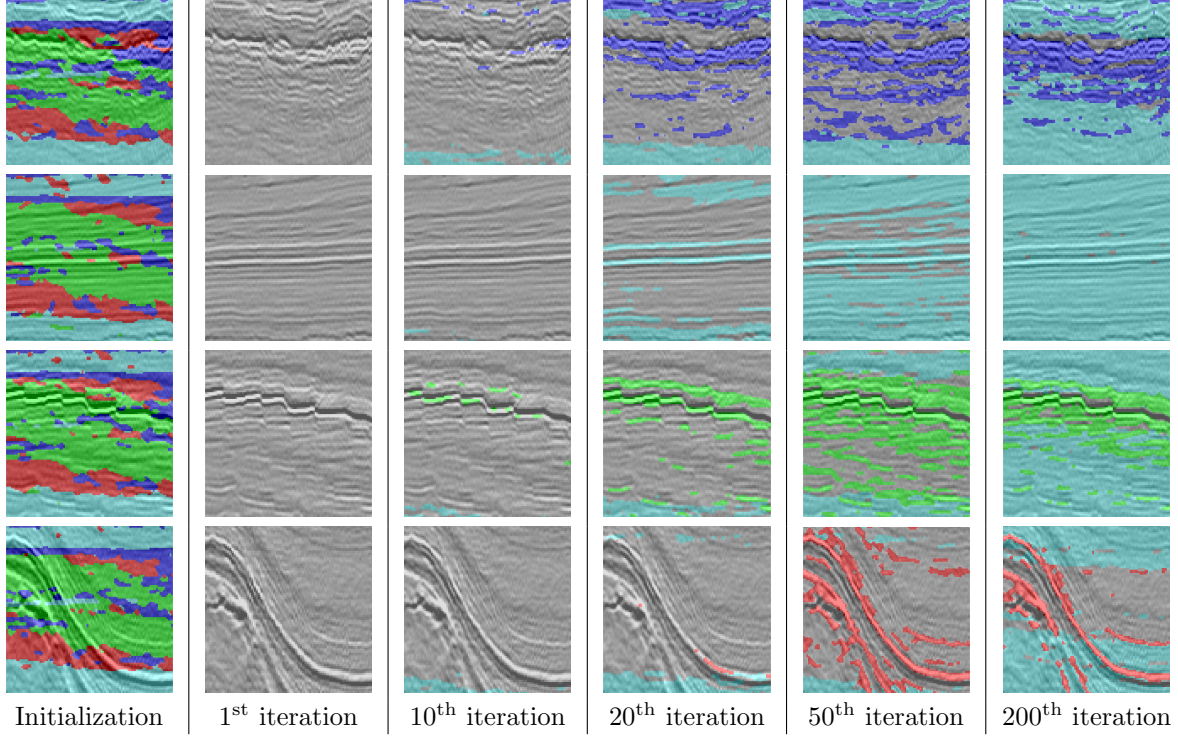


Figure 12: Results of our weakly supervised label mapping approach for sample images from each class for various iterations. The initial labels (i.e. with randomly initialized coefficients) are also shown in the first column.

in terms of their similarity values. The performance of a similarity measure is quantified using information retrieval metrics. In order to present these metrics, let us first define the following sets:

- $\mathcal{R}_i^{(j)} = \{\mathbf{r}_i^{(1)}, \mathbf{r}_i^{(2)}, \dots, \mathbf{r}_i^{(j)}\}$ is the set of the first j retrieved images for $\mathbf{x}_i$. Note that the elements of $\mathcal{R}_i^{(j)}$ are sorted according to their similarity to $\mathbf{x}_i$ such that $\text{Similarity}(\mathbf{x}_i, \mathbf{r}_i^{(k)}) \geq \text{Similarity}(\mathbf{x}_i, \mathbf{r}_i^{(k+1)})$.
- $\mathcal{C}_i$ is the set of all images that are of the same class as $\mathbf{x}_i$; excluding the image itself.
- $\mathcal{R}_i^{(j)} \cap \mathcal{C}_i$ is the intersection set of $\mathcal{R}_i^{(j)}$ and $\mathcal{C}_i$. It contains images that are of the same class as the query image $\mathbf{x}_i$ in the set of retrieved images $\mathcal{R}_i^{(j)}$.

Next, we define information retrieval metrics that were used to assess the performance of the similarity measures.

- **Precision at M (P@ M)** is the average percentage of the correctly retrieved images when M images are retrieved. Formally,

$$\text{P@}M = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{|\mathcal{R}_i^{(M)} \cap \mathcal{C}_i|}{|\mathcal{R}_i^{(M)}|}, \quad (\text{A-1})$$

where $|\cdot|$ is the number of elements in the set.

- **Retrieval Accuracy (RA)** is the P@ n when n is equal to the number of elements that are of the same class the query images, i.e. $n = |\mathcal{C}_i|$.

$$\text{RA} = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{|\mathcal{R}_i^{(|\mathcal{C}_i|)} \cap \mathcal{C}_i|}{|\mathcal{R}_i^{(|\mathcal{C}_i|)}|}. \quad (\text{A-2})$$

- **Average Precision (AP)** for query image $\mathbf{x}_i$ is a measure of precision that takes into account the order of which the correct images are retrieved. It is defined as:

$$\text{AP}_i = \frac{1}{|\mathcal{C}_i|} \sum_{j=1}^{N_s-1} \frac{|\mathcal{R}_i^{(j)} \cap \mathcal{C}_i|}{|\mathcal{R}_i^{(j)}|} \times 1_{\{r_i^{(j)} \in \mathcal{C}_i\}}, \quad (\text{A-3})$$

where $1_{\{r_i^{(j)} \in \mathcal{C}_i\}}$ is the indicator function and it is equal to 1 if and only if $r_i^{(j)} \in \mathcal{C}_i$, and 0 otherwise. **Mean Average Precision (MAP)** is the mean value of AP for all images in the dataset.

- **Receiver Operating Characteristics (ROC)** is a plot of the True Positive Rate (TPR) versus False Positive Rate (FPR) for different similarity thresholds. TPR is the percentage of pairs of images that are correctly identified as similar by the similarity measure. FPR is the percentage of pairs of images that are not similar but were identified as similar

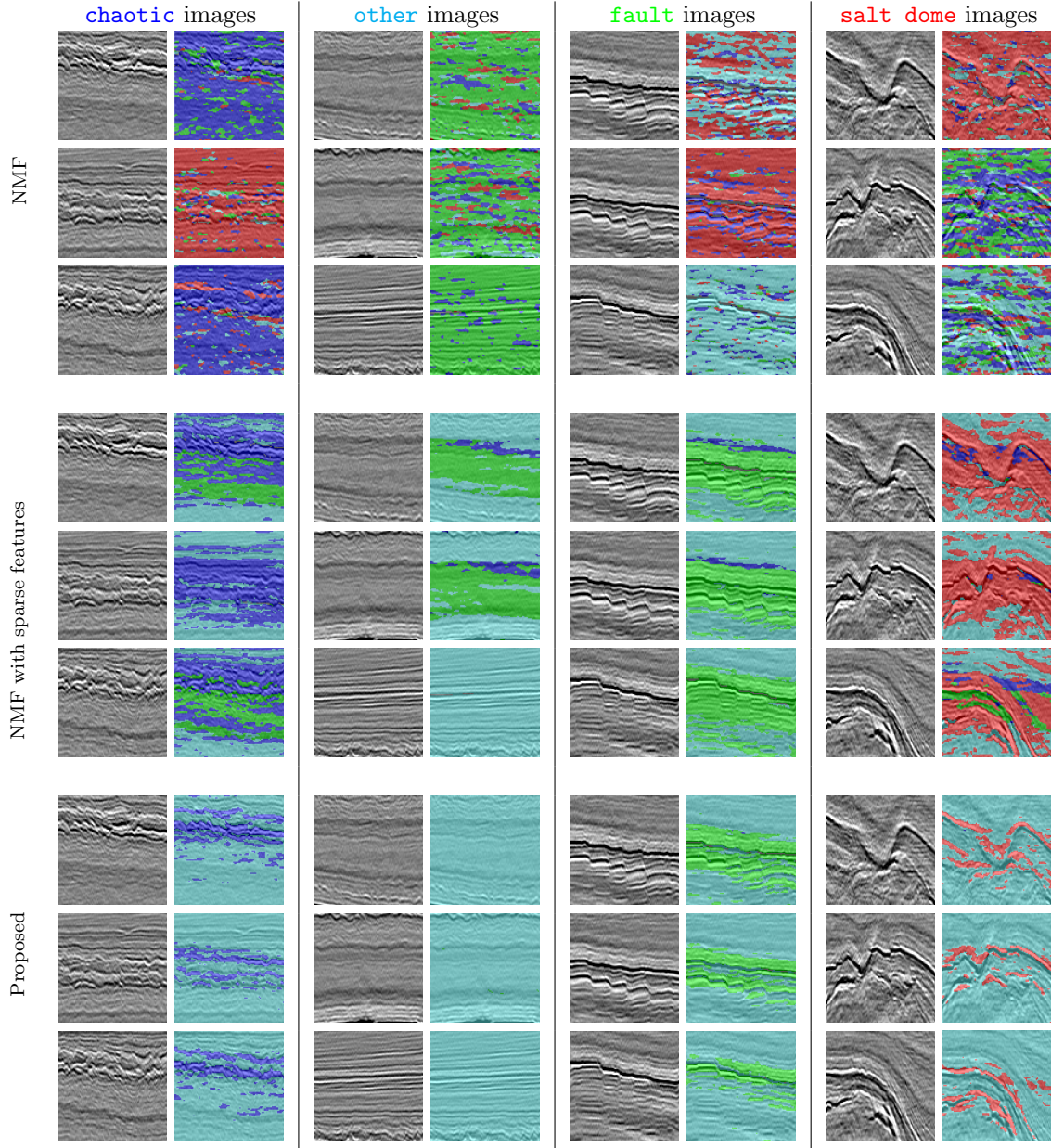


Figure 13: A small sample of the final results of the weakly-supervised label mapping algorithm using NMF (top), NMF with sparse features in $\mathbf{W}^0$ (middle), and our proposed formulation (bottom). The first two columns show images containing **chaotic** structures, and the corresponding **chaotic** pixel-level labels generated by our method in blue. The middle two columns show images that contain **fault** structures, and **fault** pixel-level labels generated by our method in green. The last two columns show images that contain **salt dome** boundaries, and **salt dome** pixel-level labels generated by our method in red. Cyan corresponds to the **other** class.

by the similarity measure. The area under the ROC curve, denoted as **AUC**, is used as a measure of the discriminative power of a similarity measure. The ideal ROC curve would have perfect TPR (TPR=1) for all values of FPR, and in this case, the area under the curve would be maximum AUC = 1.

Clustering Evaluation Metric

We use the **rand index** to evaluate our performance in the clustering experiments. The rand index is defined as follows. For each pair of images, $\mathbf{x}_i$ and $\mathbf{x}_j$, in the dataset, we compare the results obtained by k -means clustering with the ground truth which are the image labels. Then we count the number of correctly clustered pairs. A pair is said to be correctly clustered if:

- $\mathbf{x}_i$ and $\mathbf{x}_j$ are of the same class and are in the same cluster in the similarity-based clustering.
- $\mathbf{x}_i$ and $\mathbf{x}_j$ are of different classes and are in different clusters in the similarity-based clustering.

If p_{correct} is the total number of correctly clustered pairs and $p_{\text{total}} = \binom{N_s}{2}$ is the total number of possible pairs in the dataset, The Rand Index is defined as the ratio of the two numbers,

$$\text{Rand Index} = \frac{p_{\text{correct}}}{p_{\text{all}}} = \frac{2p_{\text{correct}}}{N_s(N_s - 1)}. \quad (\text{A-4})$$

All of the metrics defined above are in the range $[0, 1]$ with 1 being a perfect score.

APPENDIX B

DERIVATION OF THE MULTIPLICATIVE UPDATE RULES

We would like to derive the multiplicative update rules shown in equations 13 and 14. These multiplicative update rules solve the optimization problems introduced in equations 11 and 12. We adopt an approach similar to that proposed by Lee and Seung (2001). Namely, we will derive the gradient descent updates to solve the problem for $\mathbf{W}$ and $\mathbf{H}$ separately. Then, we solve the problem in equation 10 by alternately updating $\mathbf{W}$ and $\mathbf{H}$ successively until they converge. We will derive the multiplicative update rules using the objective functions in equations 11 and 12. For the sake of the simplicity of the derivation, we drop all the constraints in equations 11 and 12, and later show that the derived multiplicative update rules for the non-constrained problem also solve the constrained optimization problem under the conditions that we have. Additionally, we have shown in Figure 11 that solving these two problems iteratively also solves the problem in equation 10. Therefore, for matrix $\mathbf{W}$ we have

$$\arg \min_{\mathbf{W}} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2, \quad (\text{B-1})$$

and for $\mathbf{H}$, we have

$$\arg \min_{\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \gamma \|\mathbf{HH}^T - \mathbf{I}\|_F^2 + \lambda_2 \|\mathbf{H}\|_F^2. \quad (\text{B-2})$$

We derive the multiplicative update rules for $\mathbf{W}$ and $\mathbf{H}$ respectively in the following two subsections.

Multiplicative Update Rule for $\mathbf{W}$

If we denote the objective function defined in B-1 as $\mathcal{F}_{\mathbf{W}}$, we can rewrite $\mathcal{F}_{\mathbf{W}}$ as

$$\mathcal{F}_{\mathbf{W}} = \text{Tr}((\mathbf{X} - \mathbf{WH})^T(\mathbf{X} - \mathbf{WH})) + \lambda_1 \text{Tr}(\mathbf{W}^T \mathbf{W}), \quad (\text{B-3})$$

where $\text{Tr}(\cdot)$ denotes the trace of a matrix. Simplifying the expression further, and employing the property that $\text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B})$, we obtain

$$\begin{aligned} \mathcal{F}_{\mathbf{W}} = & \text{Tr}(\mathbf{X}^T \mathbf{X}) - \text{Tr}(\mathbf{X}^T \mathbf{WH}) - \text{Tr}(\mathbf{H}^T \mathbf{W}^T \mathbf{X}) \\ & + \text{Tr}(\mathbf{H}^T \mathbf{W}^T \mathbf{WH}) + \lambda_1 \text{Tr}(\mathbf{W}^T \mathbf{W}). \end{aligned} \quad (\text{B-4})$$

Taking the partial derivative of $\mathcal{F}_{\mathbf{W}}$ with respect to $\mathbf{W}$ we get

$$\begin{aligned} \frac{\partial \mathcal{F}_{\mathbf{W}}}{\partial \mathbf{W}} = & -2(\mathbf{XH}^T) + 2(\mathbf{WHH}^T) + 2\lambda_1 \mathbf{W} \\ \propto & -\mathbf{XH}^T + \mathbf{WHH}^T + \lambda_1 \mathbf{W} \end{aligned} \quad (\text{B-5})$$

The gradient descent update for $\mathbf{W}$ will then be a step in the direction of the negative gradient. In other words,

$$\mathbf{W}^{t+1} = \mathbf{W}^t + \eta(\mathbf{XH}^T - \mathbf{W}^t \mathbf{H}^t \mathbf{H}^{tT} - \lambda_1 \mathbf{W}^t), \quad (\text{B-6})$$

where η is the step size. Note that this is an additive update rule. The negative signs indicate that even if the values in $\mathbf{X}$, $\mathbf{W}^0$ and $\mathbf{H}^0$ are non-negative, we are not guaranteed to arrive at a non-negative final solution. However, by selecting our step size as

$$\eta = \frac{\mathbf{W}^t}{\mathbf{W}^t \mathbf{H}^t \mathbf{H}^{tT} + \lambda_1 \mathbf{W}^t}, \quad (\text{B-7})$$

and substituting in the gradient descent update in equation B-6, the additive rule becomes a multiplicative update rule:

$$\mathbf{W}^{t+1} = \mathbf{W}^t \odot \frac{(\mathbf{XH}^T + \epsilon)_{ij}}{(\mathbf{W}^t \mathbf{H}^t \mathbf{H}^{tT} + \lambda_1 \mathbf{W}^t + \epsilon)_{ij}}. \quad (\text{B-8})$$

We add a small positive real number ϵ to avoid division by zero. This result is identical to the result in equation 13.

Multiplicative Update Rule for $\mathbf{H}$

Similarly for $\mathbf{H}$, we write the objective function in equation B-2 as

$$\begin{aligned} \mathcal{F}_{\mathbf{H}} = & \text{Tr}((\mathbf{X} - \mathbf{WH})^T(\mathbf{X} - \mathbf{WH})) + \lambda_2 \text{Tr}(\mathbf{H}^T \mathbf{H}) \\ & + \gamma \text{Tr}((\mathbf{HH}^T - \mathbf{I})^T(\mathbf{HH}^T - \mathbf{I})). \end{aligned} \quad (\text{B-9})$$

Taking the partial derivative of $\mathcal{F}_{\mathbf{H}}$ with respect to $\mathbf{H}$, and simplifying the expression further,

$$\begin{aligned} \frac{\partial \mathcal{F}_{\mathbf{H}}}{\partial \mathbf{H}} = & -2\mathbf{W}^T \mathbf{X} + 2(\mathbf{W}^T \mathbf{W} \mathbf{H}) + 2\lambda_2 \mathbf{H} \\ & + 4\gamma \mathbf{H}^T \mathbf{H} \mathbf{H}^T - 2\gamma \mathbf{H} \\ & \alpha(\mathbf{W}^T \mathbf{W} \mathbf{H}) + \lambda_2 \mathbf{H} + \gamma \mathbf{H}^T \mathbf{H} \mathbf{H}^T \\ & - (\mathbf{W}^T \mathbf{X} + \gamma \mathbf{H}). \end{aligned} \quad (\text{B-10})$$

The gradient descent update step then becomes

$$\begin{aligned} \mathbf{H}^{t+1} = & \mathbf{H}^t + \eta \left(\mathbf{W}^{t+1T} \mathbf{X} + \gamma \mathbf{H}^t \right. \\ & \left. - (\mathbf{W}^{t+1T} \mathbf{W}^{t+1} \mathbf{H}^t + \lambda_2 \mathbf{H}^t + \gamma \mathbf{H}^{tT} \mathbf{H}^t \mathbf{H}^{tT}) \right). \end{aligned} \quad (\text{B-11})$$

If we select the step size to be

$$\eta = \frac{\mathbf{H}^t}{\mathbf{W}^{t+1T} \mathbf{W}^{t+1} \mathbf{H}^t + \lambda_2 \mathbf{H}^t + \gamma \mathbf{H}^{tT} \mathbf{H}^t \mathbf{H}^{tT}}, \quad (\text{B-12})$$

and substitute this value in equation B-11 and simplify, we arrive at the multiplicative update rule for $\mathbf{H}$

$$\mathbf{H}^{t+1} = \frac{\mathbf{H}^t \odot (\mathbf{W}^{t+1T} \mathbf{X} + \gamma \mathbf{H}^t + \epsilon)_{ij}}{(\mathbf{W}^{t+1T} \mathbf{W}^{t+1} \mathbf{H}^t + \lambda_2 \mathbf{H}^t + \gamma \mathbf{H}^{tT} \mathbf{H}^t \mathbf{H}^{tT} + \epsilon)_{ij}} \quad (\text{B-13})$$

This is identical to the update rule shown in equation 14.

Constrained Optimization

The update rules in equations B-8 and B-13 solve the non-constrained problems in equation B-1 and B-2. However, our original problem in equation 10 is a constrained one. Since we initialize the matrices $\mathbf{W}$ and $\mathbf{H}$ with non-negative values, it is trivial to see that the multiplicative update rules in equations B-8 and B-13 will always give non-negative results, thus satisfying the non-negativity constraint. Furthermore, since the sparsity constraint on the features $\mathbf{w}_i$ is applied to the initial features, $\mathbf{W}^0$, any further application of the update rule in equation B-8 will not modify the zero elements in the matrix $\mathbf{W}$, and hence, the initial feature sparsity is preserved. Therefore, although we solved for the non-constrained problem in equations B-1 and B-2, our solution is still valid for the constrained problem in equation 10. It is important to note that the final solution is dependant on the initial features $\mathbf{W}^0$. If $\mathbf{W}^0$ had a wrong sparsity structure, then the final results would not be accurate. Therefore, it is essential to initialize $\mathbf{W}^0$ properly to obtain accurate results.

REFERENCES

- Al-Marzouqi, H., and G. AlRegib, 2014, *in* Similarity index for seismic data sets using adaptive curvelets: 1470–1474.
- Alaudah, Y., and G. AlRegib, 2015, A curvelet-based distance measure for seismic images: Image Processing (ICIP), 2015 IEEE International Conference on, 4200–4204.
- , 2017, *in* A weakly supervised approach to seismic structure labeling: 2158–2163.
- Alaudah, Y., H. Di, and G. AlRegib, 2017, *in* Weakly Supervised Seismic Structure Labeling via Orthogonal Non-Negative Matrix Factorization.
- Alfarraj, M., Y. Alaudah, and G. AlRegib, 2016, Content-adaptive non-parametric texture similarity measure: 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSp), 1–6.
- Barnes, A. E., and K. J. Laughlin, 2005, *in* Investigation of methods for unsupervised classification of seismic data: 2221–2224.
- Borg, I., and P. J. Groenen, 2005, Modern multidimensional scaling: Theory and applications: Springer Science & Business Media.
- Candes, E., L. Demanet, D. Donoho, and L. Ying, 2005, Fast discrete curvelet transforms: Multiscale Modelling and Simulations, **5**, 861–899.
- Candes, E. J., and D. L. Donoho, 2004a, New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities: Communications on Pure and Applied Mathematics, **57**, 219–266.
- , 2004b, New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities: Communications on Pure and Applied Mathematics, **57**, 219–266.
- Center for Energy, and Geo Processing, 2015, LAND-MASS: Large North-Sea Dataset of Migrated Aggregated Seismic Structures.
- Coléou, T., M. Poupon, and K. Azbel, 2003, Unsupervised seismic facies classification: A review and comparison of techniques and implementation: The Leading Edge, **22**, 942–953.
- dGB Earth Sciences, B., 1987, The Netherlands Offshore, The North Sea, F3 Block - Complete.
- Ding, C., X. He, and H. D. Simon, 2005, On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering: Proceedings of the fifth SIAM International Conference on Data Mining (SDM), 606–610.
- Figueiredo, A., F. B. Silva, P. Silva, L. de O. Martins, R. L. Milidiú, and M. Gattass, 2015, *in* A Clustering-based Approach to Map 3D Seismic Horizons: 1166–1170.
- Guillen, P., G. Larrazabal*, G. Gonzalez, D. Bumber, and R. Vilalta, 2015, *in* Supervised learning to detect salt body: 1826–1829.
- Hong, S., J. Choi, J. Feyereisl, B. Han, and L. S. Davis, 2016, Joint image clustering and labeling by matrix factorization: IEEE Transactions on Pattern Analysis and Machine Intelligence, **38**, 1411–1424.
- Hong, S., H. Noh, and B. Han, 2015, Decoupled deep neural network for semi-supervised semantic segmentation: Advances in Neural Information Processing Systems, 1495–1503.
- Hou, Q., P. K. Dokania, D. Massiceti, Y. Wei, M.-M. Cheng, and P. Torr, 2016, Mining pixels: Weakly supervised semantic segmentation using image labels: arXiv

- preprint arXiv:1612.02101.
- Hoyer, P. O., 2004, Non-negative Matrix Factorization with Sparseness Constraints: The Journal of Machine Learning Research, **5**, 1457–1469.
- Kim, D., D. Yoo, I. S. Kweon, et al., 2017, Two-phase learning for weakly supervised object localization: arXiv preprint arXiv:1708.02108.
- Kim, H.-E., and S. Hwang, 2016, Deconvolutional feature stacking for weakly-supervised semantic segmentation: arXiv preprint arXiv:1602.04984.
- Lee, D., and H. Seung, 2001, Algorithms for non-negative matrix factorization: Advances in neural information processing systems, 556–562.
- Lee, D. D., and H. S. Seung, 1999, Learning the parts of objects by non-negative matrix factorization.: Nature, **401**, 788–91.
- Lin, Y., G. Guthrie, D. Coblenz, S. Wang, and J. Thiagarajan, 2017, *in* Towards real-time geologic feature detection from seismic measurements using a randomized machine-learning algorithm: 2143–2148.
- Long, Z., Z. Wang, and G. AlRegib, 2015, Seisim: structural similarity evaluation for seismic data retrieval: Presented at the 2015 IEEE ICCSP: Presented at the , IEEE.
- Lu, Z., Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, 2017, Learning from weak and noisy labels for semantic segmentation: IEEE transactions on pattern analysis and machine intelligence, **39**, 486–500.
- Niu, Y., Z. Lu, S. Huang, P. Han, and J.-R. Wen, 2015, Weakly supervised matrix factorization for noisily tagged image parsing.: IJCAI, 3749–3755.
- Paatero, P., and U. Tapper, 1994, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values: Environmetrics, **5**, 111–126.
- Papandreou, G., L.-C. Chen, K. P. Murphy, and A. L. Yuille, 2015, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation: Proceedings of the IEEE International Conference on Computer Vision, 1742–1750.
- Pathak, D., P. Krahenbuhl, and T. Darrell, 2015, Constrained convolutional neural networks for weakly supervised segmentation: Proceedings of the IEEE International Conference on Computer Vision, 1796–1804.
- Pinheiro, P. O., and R. Collobert, 2015, From image-level to pixel-level labeling with convolutional networks: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1713–1721.
- Qi, J., T. Lin, T. Zhao, F. Li, and K. Marfurt, 2016, Semisupervised multiattribute seismic facies analysis: Interpretation, **4**, SB91–SB106.
- Ramirez, C., G. Larrazabal, and G. Gonzalez, 2016, Salt body detection from seismic data via sparse representation: Geophysical Prospecting, **64**, 335–347.
- Roy, O., and M. Vetterli, 2007, The effective rank: A measure of effective dimensionality: European signal processing conference (EUSIPCO), 606–610.
- Torresani, L., 2014, *in* Weakly Supervised Learning: Springer US, 883–885.
- Türkmen, A. C., 2015, A Review of Nonnegative Matrix Factorization Methods for Clustering: 1–23.
- Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, 2004, Image quality assessment: from error visibility to structural similarity: Image Processing, IEEE Transactions on, **13**, 600–612.
- Wang, Z., T. Hegazy, Z. Long, and G. AlRegib, 2015a, Noise-robust detection and tracking of salt domes in postmigrated volumes using texture, tensors, and subspace learning: GEOPHYSICS, **80**, WD101–WD116.
- , 2015b, Noise-robust detection and tracking of salt domes in postmigrated volumes using texture, tensors, and subspace learning: Geophysics, **80**, WD101–WD116.
- Wang, Z., and E. P. Simoncelli, 2005, Translation insensitive image similarity in complex wavelet domain: Presented at the In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP05). IEEE International Conference on, Citeseer.
- Wei, Y., X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, 2017, Stc: A simple to complex framework for weakly-supervised semantic segmentation: IEEE transactions on pattern analysis and machine intelligence, **39**, 2314–2320.
- Zhang, G., and X. Gong, 2016, Nonnegative matrix cofactorization for weakly supervised image parsing: IEEE Signal Processing Letters, **23**, 1682–1686.
- Zhang, W., S. Zeng, D. Wang, and X. Xue, 2015, Weakly supervised semantic segmentation for social images: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2718–2726.
- Zhao, T., V. Jayaram, A. Roy, and K. J. Marfurt, 2015, A comparison of classification techniques for seismic facies recognition: Interpretation, **3**, SAE29–SAE58.
- Zhou, W., H. Li, and Q. Tian, 2017, Recent advance in content-based image retrieval: A literature survey: arXiv preprint arXiv:1706.06064.
- Zujovic, J., T. Pappas, and D. Neuhoff, 2013, Structural texture similarity metrics for image analysis and retrieval: Image Processing, IEEE Transactions on, **22**, 2545–2558.