METRICS FOR THE QUALITY AND CONSISTENCY OF ICE LAYER ANNOTATIONS

Naomi Tack, Bayu Adhi Tama, Atefeh Jebeli, Vandana P. Janeja, Don Engel, Rebecca Williams

University of Maryland, Baltimore County (UMBC), Baltimore, Maryland, USA

ABSTRACT

Ice layers in glaciers, such as those covering Greenland and Antarctica, are deformed over time. The deformations of these layers provide a record of climate history and are useful in predicting future ice flow and ice loss. Cross sectional images of the ice can be captured by airborne radar and layers in the images then annotated by glaciologists. Recent advances in semi-automated and automated annotation allow for significantly more annotations, but the validity of these annotations is difficult to determine because ground-truth (GT) data is scarce. In this paper, we (1) propose GT-dependent and GT-independent metrics for layer annotations and (2) present results from our implementation and initial testing of GT-independent metrics, such as layer breakpoints, local layer density, spatial frequency, and layer orientation agreement.

***Index Terms*—** Ice sheet, ice-penetrating radar, quality metrics, auto-annotation

1. INTRODUCTION

Englacial ice layers are deformed and influenced by flow fields, and as these layers are buried, the basal conditions are recorded as described by Holschuh et al. [1]. Thus, englacial ice layers can be used to infer climate history, glacial dynamics, and physical ice properties, among others [2]. Ice penetrating radar can detect these layers [3, 4], which must then be annotated to further understand and describe the flow fields. Semi-automated methods are being developed [3] to reduce the annotation burden, but these methods require substantial ground truth validation in order to train a model.

Ground truth (GT) is defined as manual annotation by domain specialists, which is time-consuming and limited by the quality of the data, which often makes reliable manual annotation impossible. We have developed several quantitative metrics to characterize the quality and agreement of automatic englacial layer detection. Because GT is scarce, we separate our metrics into two groups: those that require GT and those that are separate from GT.

This work is made possible by NSF Award #2118285, "iHARP: NSF HDR Institute for Harnessing Data and Model Revolution in the Polar Regions." We acknowledge the use of the CReSIS toolbox from CReSIS generated with support from the University of Kansas, NASA Operation IceBridge grant NNX16AH54G, and NSF grants ACI-1443054, OPP-1739003, and IIS-1838230.

To aid in understanding these metrics, we produced a method of quantifying and visualizing the local metrics as quality maps. For example, quantification of layer density allows for the identification of image artifacts and glacial phenomena such as lakes or melt [5]. However, it also highlights the need for additional annotations of layers, enhanced image processing, and need for alternative partitioning for the training-testing-validation data incorporating layer density as a factor [5].

2. CONTRIBUTIONS

We investigate and implement concepts from fingerprint identification/quality analysis [6] and multi-target tracking [7, 8, 9] to produce feature and quality maps and representative statistics for englacial layer segmentations that are outputs of an unsupervised algorithm [10], and surface & bedrock annotations output from a two-step deep neural network model [11]. These methods use airborne ice-penetrating radar data hosted by the Center for Remote Sensing of Ice Sheets (CReSIS) [12], collected during several overlapping flight paths.

We investigate an initial proof-of-concept metric suite and analysis framework upon which we will build in future work. We also provide recommendations for standardization and enhancement for metrics with high "labeling utility." To this end, we propose structured families of metrics that evaluate different aspects of englacial layer detection. We consider two conceptual groupings of metrics:

Metrics that require GT vs. do not require GT: When GT is not available or is expensive to curate, it is beneficial to consider quality metrics that don't require the use of reference data. The goal is not to completely negate the use of GT but to decrease the manual burden of annotating it by producing as many *a priori* automated annotations as possible. Attempting to measure layer fidelity/correctness or association errors necessarily require GT. Metrics that do not require GT can be assembled based on their utility as annotated GT labels for training supervised approaches.

Local vs global metrics: Quality measures at a local level preserve spatial/regional information, whereas global measures assign the whole image a single value, which enables ranking and comparing results directly. Metrics can be computed at the layer level to pinpoint instance-level anomalies

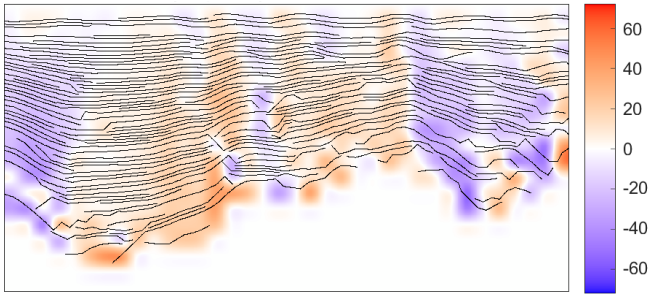

Fig. 1. Local layer orientation map. Red indicates positive slope in degrees and blue indicates negative slope in degrees. These orientations are used to compute orientation agreement map, as shown in Fig. 2

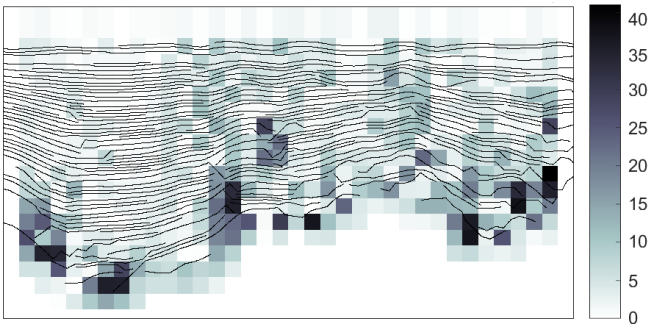

Fig. 2. Quality map of local layer agreement with 8-connected neighbors. Darker areas indicate orientation disagreement, while light areas indicate agreement between neighboring patches. Fairly continuous areas with zero slope have high agreement, while areas near the bedrock with discontinuities or drastic direction changes have low agreement.

and trends or can be windowed and/or gridded to provide full coverage of the image. Local metrics can be assembled into feature heatmaps, quality maps, and histograms. In contrast, global metrics are reported on a per-image basis, typically aggregated from local metrics and then aggregated into a single quality score. The advantages of global metrics include the ability to provide a single number to rank and compare models/algorithms. However, global quality scores can cause aggregation effects that are opaque and may not reflect outputs that are "mostly correct" or "good enough."

Quality values and maps typically involve determining a threshold of acceptance, so we present both the raw local/global feature values and the computed quality values. The visualization goals are to output a variety of quality maps/feature maps wherein anomalous regions appear salient and to display drill-down information such as local and global histogram values. Future work will streamline and extend the visualization for enhanced quality exploration, including in virtual reality [13].

Unsupervised and supervised performances are dependent on both radar image quality and annotation quality. In a two-stage framework using an unsupervised model to produce labels for a supervised approach [14], the quality of the detected/segmented layers affects the supervised model's performance, such that errors are propagated. Therefore, in this work we focus on detected layer quality; future work may incorporate radar image quality measures as well.

For the GT-independent sub-family, we compute and visualize quality maps and local histograms representing layer density/spacing (Figure 1), layer orientation agreement with neighbors (Figure 2), local frequency components, and minutiae detection (breakpoints, branch points, corners). The goal of computing these local metrics is to accumulate a quality feature vector in order to compute a quality score that accurately represents "label utility" of the detected layer mask.

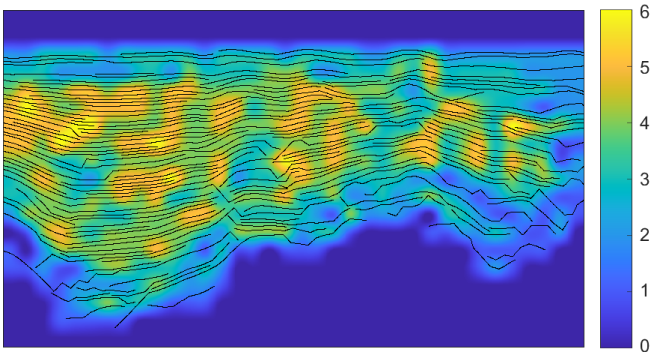

Fig. 3. Local layer density map, where bright yellow areas indicate areas with high layer density, and dark blue represents areas with few or no layers. Layer density maps can be used to identify areas with artifacts and/or incomplete annotation, or to cue in on interesting glaciological morphology (e.g. ice lenses, crevasses, melt ponds, etc.)

3. DATA

We use layers generated by [10, 11] on the ice-penetrating radar data hosted by the Center for Remote Sensing of Ice Sheets (CReSIS) [12]. The radar images were collected during flight, with several intersecting flight paths.

Because we want to reduce dependence on GT for evaluation while still enabling useful evaluation assessments, we focus on evaluation/quality metrics that can be computed without GT. For the purposes of this paper, we consider "GT" to include layer instance ids, layer pixel locations, and any image artifacts or ice anomalies. Radar parameters and flight paths are considered supporting metadata available for computing both GT-dependent and GT-independent layer quality metrics.

4. RESULTS

We compute local and global metrics and visualize them using feature maps and quality maps, and accumulate a histogram and quality feature vector that can be used to determine layer quality effects on the supervised stage of Jebeli *et al*. [14].

4.1. Ground-truth Independent Metrics

Our first metric suite encompasses global and local measures that can be computed without requiring accurate GT, as defined in Section 3. In the GT-independent sub-family, we further discuss the computed quality maps and features as discussed in Section 2. The advantages of these measures stem from the retention of spatial information that can be used to produce a map of high vs. low-quality areas based on each metric. We use the aforementioned average layer density to compute an appropriate window size for measuring orientations.

Layer breakpoints - Layer breakpoints are identified when a layer ends before the end of the image. Layers are generally continuous throughout the radar images, and identifying the "dropped" layers can help identify either layers that have actually "dropped" from the image or if they are a product of the layer detection model.

Local layer density - The local layer density is calculated using a sliding window average with a 50% overlap (Fig. 3). By counting the number of connected components per window, we can identify where dense layers areas are calculated in the AI approach. These metrics may prove useful when evaluating the overall performance of the approach for annotation as they may indicate areas where layers are easier to automatically annotate and areas the model fails to capture the complexity of the layers.

Spatial frequency - The identified layers were modeled mathematically using cubic splines to generate their equations and interpolate the normals along evenly spaced intervals. By testing the intersections of the normal with the nearest layer above, we then calculate the Euclidean distance between layers. This allows us to generate mean distances across several columns of the image and collapse these representative distances into means for the column. These were compared to the 2D FFT spatial frequency map computed, as shown in Fig. 4. Areas with low spatial frequency can be seen near incomplete layers.

Orientation and orientation agreement - Fig. 2 shows the orientation agreement quality map. By calculating the layer normals we can understand the overall orientation of the layers (shown in Fig. 1) and allows us to take a sliding window average and see how orientations agree with the neighboring layers. This could indicate the flow of the ice or bias in the model to produce layers that are "going in the same direction."

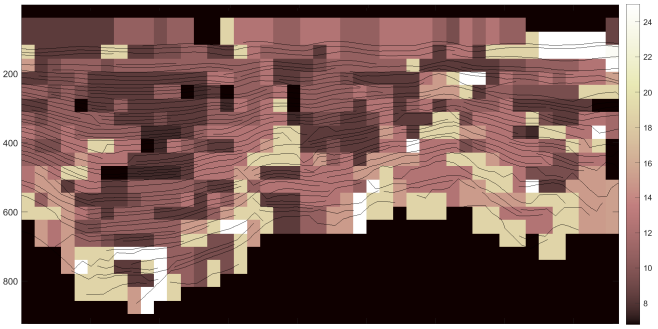

Fig. 4. Local spatial frequency map. Bright areas correspond to lower spatial frequency (units of px/cycle), darker areas indicate higher spatial frequency (i.e. decreased distance between layers). The maximum frequency recorded is the half the window size (in this case, 30 px/cycle for a window size of 75 px).

5. FUTURE WORK

In this paper, we focus mainly on ground-truth independent metrics to avoid over-reliance on the availability and quality of hand-picked/annotated layers. In our ongoing work, we are planning to also compute several GT-dependent metrics, from both the local and global families of metrics. These include inter-annotator agreement for GT (per available annotated layer) and layer completeness for GT (per image).

For both the GT dependent and independent metric sets, we plan to assemble a quality vector and examine the effects of both high and low-quality layers (i.e., their predictive utility) using the U-net framework in development by [11].

Our current work aims to establish a standardized framework for evaluating the performance of models and techniques that enable englacial layer detection, localization, and association. In the near-term, we will also perform a sensitivity analysis of our quality metrics and supervised algorithm performance. We will compare these metrics with domain-expert quality assessment and investigate feature importance using dimensionality reduction and other techniques.

Longer-term, we intend to explore the agreement of annotations in multiple images captured in proximity to each other, such as images captured by parallel flights of aircraft as well as images captured by intersecting flight paths.

Our ultimate goal is to develop and implement a standardized evaluation, visualization, and annotation tool for ice layer detection. This will help users employ the quality metrics to aid in correction/new annotation, provide quality assurance and quality control (QA/QC) for hand-picked layers, and could ultimately support training a supervised algorithm to correct them automatically.

6. REFERENCES

[1] N. Holschuh, B. R. Parizek, R. B. Alley, and S. Anandakrishnan, "Decoding ice sheet behavior using englacial layer slopes," *Geophysical Research Letters*, vol. 44, no. 11, pp. 5561–5570, 2017.

[2] J. A. MacGregor, W. T. Colgan, M. A. Fahnestock, M. Morlighem, G. A. Catania, J. D. Paden, and S. P. Gogineni, "Holocene deceleration of the greenland ice sheet," *Science*, vol. 351, no. 6273, pp. 590–593, 2016.

[3] J. A. MacGregor, M. A. Fahnestock, G. A. Catania, J. D. Paden, S. Prasad Gogineni, S. K. Young, S. C. Rybarski, A. N. Mabrey, B. M. Wagman, and M. Morlighem, "Radiostratigraphy and age structure of the greenland ice sheet," *Journal of Geophysical Research: Earth Surface*, vol. 120, no. 2, pp. 212–241, 2015.

[4] J. P. Briner, J. K. Cuzzone, J. A. Badgeley, N. E. Young, E. J. Steig, M. Morlighem, N.-J. Schlegel, G. J. Hakim, J. M. Schaefer, J. V. Johnson, A. J. Lesnek, E. K. Thomas, E. Allan, O. Bennike, A. A. Cluett, B. Csatho, A. de Vernal, J. Downs, E. Larour, and S. Nowicki, "Rate of mass loss from the greenland ice sheet will exceed holocene values this century," *Nature*, vol. 586, no. 7827, pp. 70–74, Oct 2020.

[5] D. West, J. T. Harper, N. F. Humphrey, and W. T. Pfeffer, "Measurement and Modeling of Firn Densification in the Percolation Zone of the Greenland Ice Sheet," in *AGU Fall Meeting Abstracts*, Dec. 2009, vol. 2009, pp. C31E–0477.

[6] E. Tabassi, M. Olsen, O. Bausinger, C. Busch, A. Figlarz, G. Fiumara, O. Henniger, J. Merkle, T. Ruhland, C. Schiel, and M. Schwaiger, "NIST fingerprint image quality 2," 2021-07-13 04:07:00 2021.

[7] Y. Song, Z. Hu, T. Li, and H. Fan, "Performance evaluation metrics and approaches for target tracking: A survey," *Sensors*, vol. 22, no. 3, pp. 793, Jan 2022.

[8] E. Drelie Gelasca, J. Byun, B. Obara, and B. Manjunath, "Evaluation and benchmark for biological image segmentation," in *2008 15th IEEE International Conference on Image Processing*, 2008, pp. 1816–1819.

[9] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 246309, May 2008.

[10] S. Xiong, J.-P. Muller, and R. C. Carretero, "A new method for automatically tracing englacial layers from MCoRDS data in NW Greenland," *Remote Sensing*, vol. 10, no. 1, pp. 43, 2017.

[11] A. Jebeli, B. A. Tama, V. Janeja, N. Holschuh, C. Jensen, M. Morlighem, J. A. MacGregor, and M. Fahnestock, "TSSA: Two-step semi-supervised annotation for englacial radargrams on the greenland ice sheet," in *International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2023, In press.

[12] CReSIS, "Radar depth sounder (RDS) data," http://data.cresis.ku.edu/, 2021, Lawrence, Kansas, USA.

[13] N. Tack, N. Holschuh, S. Sharma, R. Williams, and D. Engel, "Development and initial testing of XR-based fence diagrams for polar science," in *International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2023, In press.

[14] A. Jebeli, B. A. Tama, V. Janeja, N. Holschuh, C. Jensen, M. Morlighem, J. A. MacGregor, and M. Fahnestock, "TSSA: Two-step semi-supervised annotation for englacial radargrams on the greenland ice sheet," submitted for publication.