Analysing the Potential Impact of Labeling Disagreements for Engineering Sensor Data

Andre Busche and **Ruth Janning** and **Lars Schmidt–Thieme**
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim
{busche, janning, schmidt-thieme}@ismll.uni-hildesheim.de

Abstract

We present the hyperbola recognition problem in Ground Penetrating Radar – GPR – data as an example for pattern recognition in complex engineering sensor data. Traditionally, GPR data are analyzed manually by human experts in a tedious and time-consuming process, e.g., to deduce the positioning of linear object underneath roads just before reconstruction works take place. For supporting this process using Machine Learning methods, one needs to have accurate ground truth data to derive models out of it. As an accurate acquisition of such annotated data is impossible even for a quasi-ideal case, we annotated 700 radargram images manually. This paper presents and discusses the outcomes of this study and concludes, that using just a single evaluation criteria to compare performances of GPR-focused Machine Learning methods might not be enough.

1 Introduction

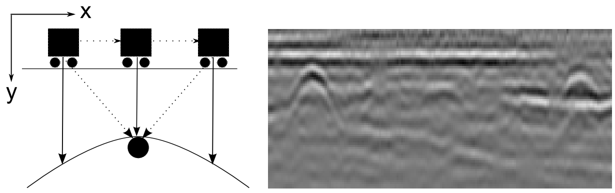
Ground Penetrating Radar (GPR) is used to investigate the shallow surface, e.g., to find buried landmines [Wilson *et al.*, 2007] or pipes and cables underneath (road) surfaces. Our current data is measured using an on-site vehicle equipped with a multi-channel array (multiple ‘channels’ are recorded at different frequencies and relative positions while the vehicle is moving) and illuminates structures in subsoil down to about 3-4 meters in depth. We are aiming at assisting the analysis process by means of probabilistic methods, while a special focus is put on the identification of pipes and cables of various types (e.g., PE, metal, stoneware) which are represented as hyperbola-like structures on measured radargram images (see Figure 1).

One ultimate goal of GPR data analysis is the derivation of supply maps, that is, maps of buried objects of a certain kind. The creation of those maps is required, e.g., as municipalities, according to one of our project partners and at least in Germany, seldomly have a single map of their buried structures. Instead, maps and plans of buried pipes and cables are cluttered and only available in a distributed manner, and can only be partially gathered and combined when requested by all parties owning buried objects, such as water supply companies, power supply companies, and telecommunication companies. In any case, those existing maps may be inaccurate and not recent, causing additional problems when highly accurate maps are required. This paper represents a first step towards to (semi-)automated creation of such supply maps, by means of developing supervised Machine Learning methods for an automated detection of such buried objects. The overall process can be split up into two distinct aspects: (a) the detection of individual objects in radargram images, and (b) the creating of supply maps out of individually detected object locations. This work focuses on the first aspect, whereas the latter aspect (b) can be tackled, e.g., by solutions as presented in [Chen and Cohn, 2011]. The collection of individual radargrams, resp. cross-sections, e.g., of a road, is done as follows: A specialized measurement vehicle drives at a constant speed along the x-axis (cf. Figure 1a) and measures a radargram image. This image visualizes reflected energies / intensities at discrete time points (y-axis). While in theory, one is able to *induce* the appearance of a radargram image out of known subsoil structures (e.g., by means of a numerical simulation software, see [Giannopoulos, 2003]), the reverse action - the *deduction of subsoil structures* out of radargram images - is a highly non-trivial task in real-world situations for a variety of reasons: (a) hyperbola reflections get distorted by supplementary reflections from horizontal layer breaks (see Figure 1, on the top right), (b) the signals’ energy decays with increasing depth, resulting in lower (visual) contrasts (see our preprocessing in Figure 2 at the bottom of area 7), (c) the pipes’ type and fillings (e.g., PE pipes filled with water, or being empty) causes multiple vertical reflections, (d) the depth–dependent energy decay of emitted radar wave requires preprocessing techniques and cause reflections to transition into background clutter at increasing depths.

Supporting the hyperbola recognition task by means of supervised models requires us of having a labeled set of radargram images at hand.

We will take the scenario of applying patch-based image-classification techniques as a running example. This requires us of having an a priori labeled dataset which contains patches (fractions of a radargram image) being labeled positive (patches containing hyperbola shapes) and negative (clutter; background noise).

For our data being measured on a test-site, GPS measurements exist for all pipes buried therein. Though one intuitively assumes that this GPS information helps for the creation of ground truth data (inducing the positions of hyperbola apexes out of the known subsoil structure), the opposite situation is the case, as (a) the heterogeneity of soil makes a consistent estimation of the actual pixel-*depth* impossible, (b) an unsteady movement of the vehicle needs to be interpolated and aligned to the radar traces, resulting in inaccurate *horizontal* pixel positions, (c) given the pipe identification task, *multiple reflections* occur, e.g., for water-filled pipes, while only the top-most reflection hyperbola can be derived from the GPS ground truth data. What



(a) GPR Data Collection Process (b) Real-world radargram (excerpt)

Figure 1: Hyperbolas are caused by steady movement of a radar vehicle across a buried pipe (a). At each horizontal position, an A-Scan (column) is measured, whose stacked visualization results in so-called radargrams, or B-Scans, (b).

immediately follows is an inaccuracy in pixel position in the ground truth data obtained from the given GPS data. This is the aspect we aim to improve by manually altering the apex positions derived from GPS annotations to match visual phenomena in the underlying radargram image. This paper is only concerned with obtaining and analyzing accurate hyperbola annotations at pixel level; the real-world deduction of pipe positions after identifying 'enough' candidate apex positions is not considered and can be achieved, e.g., utilizing methods as given in [Chen and Cohn, 2011].

Instead, this paper analyzes the outcome of these manual annotation sessions according to the following aspects:

1. **Is there a measurable difference between human annotation behavior for GPR data, and if so, can it be related to a certain aspect of the data?**

   While some reflections are clearly visible in the measured data, other cases (e.g., multiple reflections caused by the material of the pipe) exist, for which it is less intuitive to decide, whether or not those should be annotated. The question is now, to what extend humans agree on the visibility of these visual phenomena.

2. **Is it possible to gain an annotation set for complex engineering data which is inter-humanly agreed, or is the quality of annotations subject to personal taste?**

   If the annotation task can be identified as being related to personal taste, the question arises how the suitability and correctness of automatized methods trained on human tastes should be addressed.

2 Related Work

Relevant work is summarized from both the Machine Learning and Psychological perspective, with special emphasis on applications in real-world scenarios and possibilities for automatization, for which the human factor was identified to influence final results on the applicability of a technique.

2.1 Machine Learning Techniques for GPR Data Interpretation

The ultimate goal of GPR analysis is the derivation of complete and accurate tomographies based on usually just a small set of radargrams [Simi *et al.*, 2008; Chen and Cohn, 2011]. Before radargram images are fed into an automated algorithm, they are usually preprocessed. This process is usually visually [Pasolli *et al.*, 2009; Busche *et al.*, 2012] or methodologically [Chen and Cohn, 2010; 2011; Janning *et al.*, 2012a] driven.

One out of three different approaches for hyperbola detection in radargram images can be distinguished: (a) Estimation from sparse data [Chen and Cohn, 2010; Janning *et al.*, 2012a], (b) Brute-force methods, e.g., the Hough Transform, [Simi *et al.*, 2008], and (c) Supervised machine-learning, e.g., Neural Networks for patch-based classification, for which training data needs to be carefully collected beforehand (as we do here) [Al–Nuaimy*et al.*, 2000; Birkenfeld, 2010]. Our analysis discussed here influences each of these approaches, as those approaches need to be evaluated against some ground truth knowledge which, as we will show, is not obvious to obtain.

2.2 Psychological Aspects covered in this paper

For conducting manual annotation tasks, generally two groups of people can be distinguished: non-specialized humans having weak prior knowledge [Nowak and Rüger, 2010], and domain experts [Mello-Thoms, 2006; McCarley *et al.*, 2004; Volkmer *et al.*, 2005]. Many works were identified for other domains, e.g., videos / keyframes [Volkmer *et al.*, 2005] or texts [Nowak and Rüger, 2010], the one most similar being the one presented in [Klebanov *et al.*, 2008] for the text domain. No such work were identified for partial image annotations in the domain of complex engineering sensor data.

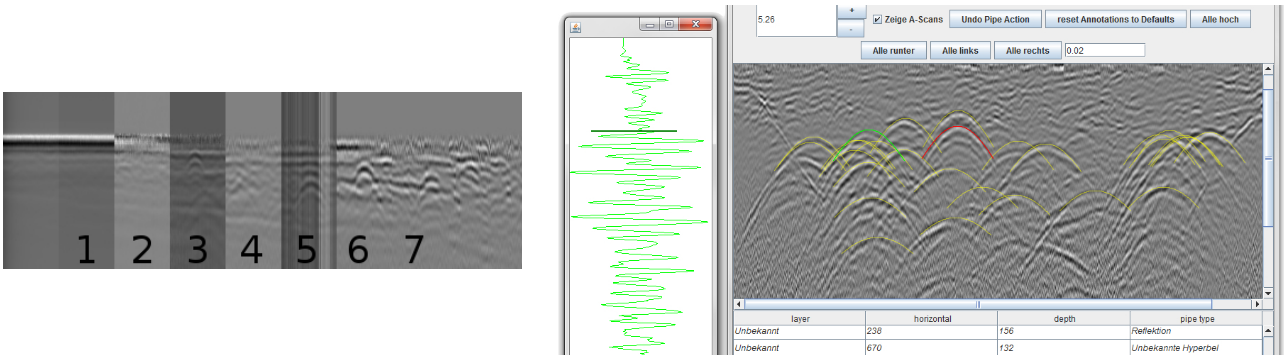
In the image domain, validating and adjusting previously defined and given annotations requires human annotators to establish a best-matching hypothesis, explaining which annotations correspond best to which structure in the raw data [Gregory, 1980]. Making final judgments on the suitability of an interpretation (being ones own or someone else's) is a non-trivial problem [Cavanagh, 2011; Nowak and Rüger, 2010], as mistakes may easily affects human life (e.g., wrong interpretations in medical screening [Mello-Thoms, 2006] or x-ray luggage screening at airports [McCarley *et al.*, 2004]). Multilayer interpretation of image (patches) by means of Neural networks [Sermanet *et al.*, 2009; Birkenfeld, 2010] is well known in the Machine Learning Community, while a thorough survey on their motivation based on the functioning of the brain is presented in [Rolls, 2012].

3 GPR Data Analyzer & Annotator

We first need to discuss GPR data preprocessing steps, as those might have an effect on the later visibility of subsoil structures (cf. Figure 2 on the left). The discussion continues with the presentation of our specialized GPR annotation software as shown in Figure 2 on the right which human annotators used to perform the manual annotation task.

3.1 Data Preprocessing

A multi-stage filter chain (presented in Figure 2 on the left) was used in the following way: the intensity values were modified to (1) have zero mean at each position (A-Scan) and (2) zero mean at each depth. Intensities are increased in a depth-dependent manner (3) to compensate the wave propagation loss. The characteristic reflection pattern was augmented using a pattern correlation filter (4): A sliding window of fixed length was moved along an A-Scan, calculating a correlation score against the sequence $(0_4, 1_5, 0_4)$ (the subscript denotes the number of repetitions), resulting in a new intensity value used for the following analy-



layer	horizontal	depth	pipe type
Unbekannt	238	156	Reflektion
Unbekannt	670	132	Unbekannte Hyperbel

Figure 2: Our GPR Analyzer and Annotator preprocesses data as shown on the left. After preprocessing, a GUI shows basic information on individual pipes / hyperbolas and is used to interactively perform CRUD - **C**reate, **R**etrieve, **U**pdate and **D**elete - operations for hyperbolas.

sis. Intensity values for an A-Scan are adjusted (5) to have variance 1. Step (6) repeats step (1). Finally, (7) a depth-dependent windowed moving average subtraction filter is applied.

All but the pattern correlation filter are non-parametric; the pattern correlation filter corresponds to a rectangular function which approximates the vertical reflection pattern of a pipe [Busche *et al.*, 2012].

3.2 Interactive Controls

Given an initial seed annotation set (a list of pixel positions) and constant soil permittivity, our GPR Data Annotator stacks an interactive layer to create, modify and delete hyperbola annotations (yellow / bright on the Figures) on top of a raw radargram visualization panel. On the top pane, it is both possible to adjust the soil permittivity (using a homogeneity assumption; used to adjust the curvature of hyperbola annotations) and to jointly adjust the initial positions of all annotations (since close-by pipes cause distortions which are hard to distinguish). Hyperbola annotations are highlighted based on the mouse position, can be drag-dropped and contain further information (e.g., its type) shown in tabular form at the bottom. Optionally, A-Scans may be visualized (left).

4 Manually Annotating GPR Data

The annotation process of 350 radargram images in total was designed while having potential inter-human disagreements in mind. We aimed at having at least two annotation sets per radargram at the end, thus finally resulting in 700 images to be annotated. For the annotation process being repetitive, tedious, and long-lasting (an average of 4 minutes per image results in estimated 46 hours) and thus error-prone, we splitted the process into two phases, also allowing to incorporate feedback after round 1.

4.1 Annotation Protocol

The annotation protocol for the human annotators was as follows: Initially, a radargram image along with a seed set of annotations ('annotation set' in the following) obtained by the GPS measurements, were visualized on the user interface. Though an initial inspection of GPS annotation sets showed inaccuracies of up to $15cm$ in both horizontal and depth / vertical image direction ($1cm$ does not scale equally on both axes), those GPS seed annotations were shown, because the test site contains many near-by located pipes being represented by interfering and intersecting reflection patterns whose distinction is challenging even for human experts (see the right area in the radargram in Figure 2). The exact location of those near-by pipe apexes was therefore determined by a 'radargram-wise global best visual match' of all hyperbola annotations being present, as some hyperbola reflections were clearly visible (see the center of the radargram therein). After globally adjusting the annotation locations, individual positions of hyperbola annotations were altered.

4.2 Round 1

In the first phase, 6 human annotators were annotating 484 radargrams in total, spending 2 days on this task. The group of human annotators was composed of both domain experts (researchers working on the data on a daily basis) as well as semi-professionals (students) working with the data. 142 radargrams were annotated once, while two annotation sets were gained for 171 images. On average, 35 (±3) radargrams per channel (1 out of 14 antenna configurations measured by the vehicle per measurement run) were annotated, while each human annotator saw each channel at least once (on average, 5.9 ±3.8 radargrams / channel). The absolute amount of radargrams per human annotator ranged from 31 to 162.

4.3 Qualitative Results for Round 1

Feedback on the quality of annotation sets was gathered using a structured online feedback questionnaire, showing radargrams row-wise either in a comparison-style view for two existing annotation sets, or in a single radargram view, if just one annotation set existed. Three experts acting as human 'judges', two of which were also annotating the radargrams beforehand, the other one being a highly skilled GPR data analyst working at our project partners company, were visually investigating the quality of the annotation sets as follows. When comparing two annotation sets, (a) a vote could be casted, ranking one set over the other. Marking (b) a pair of annotation sets as being 'problematic' examples triggered a discussion with the human annotators before round 2. For all annotation sets, assigning the *incomplete* label to a radargram marked them as being a candidate for re-annotation in phase 2, if at least two votes ranked that set higher than its competitor. A *wrong* label removed the set from further analysis.

The fact, that an annotation set is 'accurate' given its radargram, was inferred if neither an incomplete nor wrong

Annotators' quality	1	2	3
accurate	26	16	15
incomplete	57	45	37
total annotations	227	250	229

Table 1: Final counts of annotation sets along with their judgement ('accurate', 'incomplete') obtained for radargram images by three human annotators 1, 2 and 3 after both annotation rounds.

label was assigned by any judge (this held true for 39 annotation sets, representing 8% of all annotation sets). We fed all 113 annotation sets (23%) which were marked as being incomplete but superior over the other into the set of radargram images to be annotated in phase 2.

From this first feedback cycle we got first evidence that obtaining a single, consistent and inter-humanly agreed set of annotated radargrams is much more challenging than previously expected. Therefore, we decided to assign annotation quotas, a fixed amount of radargrams to be annotated, to each human.

4.4 Round 2

After further analysis on quality and comparability of the judged results from round 1, we invited only 3 (of all 6) humans to annotate the remaining part, each having a different quota. The set of radargrams contained all remaining and incompletely labeled radargrams from round 1. On average 27 (±3.2) radargrams per channel were annotated, while each human annotator saw at least 2 images per channel (9 on average, ±4).

4.5 Feedback and results from Round 2

We used the same feedback mechanism as used in round 1 for gathering feedback from the same judges, while this time only one of them also annotated the radargrams. 31 annotation sets (8%) were inferred to be 'accurate', while still 69 annotation sets (18%) are marked as being incomplete, but preferable over the other.

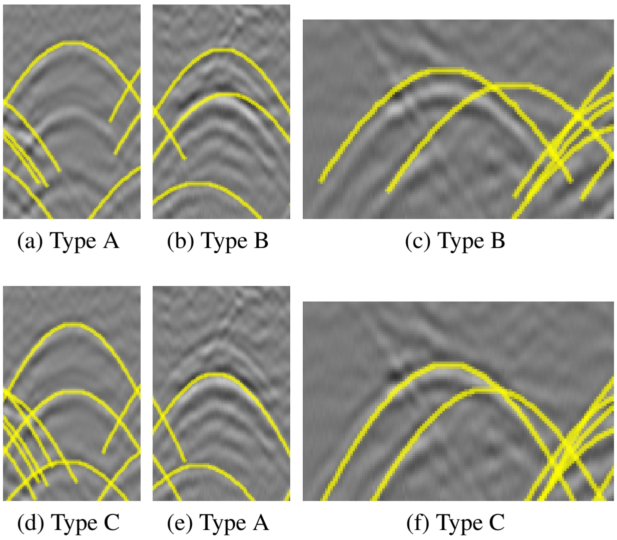
4.6 Result for both Rounds

Taking jointly both annotation rounds for just the three annotators from round 2 into account, we finally ended up with the counts of annotation sets per human annotator as shown in Table 1. As an example, annotator 3 labeled 229 radargram images in total, for which 15 (6.5%) images were judged accurate, 37 (16.2%) were judged incomplete.

The peak of accurate annotations of annotator 1 is in line with Gregory's [Gregory, 1980] hypothesis who argues that prior experience strongly influences the perception and, in direct consequence, the quality of annotations (annotator 1 is more familiar with diverse sets of radargram images than annotators 2 and 3).

5 Analyzing and Discussing Human Annotation Behavior

To answer our questions stated in the beginning, we will now have a closer look at the results of the visual judgments with respect to its potential later automatization. The following error types are qualitatively introduced after having investigated both, all obtained annotation sets, and prior GPS annotations, and shall guide both our current analysis and future ideas and developments while developing algorithms for automatic GPR image data interpretation.


(a) Type A (b) Type B (c) Type B

(d) Type C (e) Type A (f) Type C

Figure 3: Analysis of humans' annotation behavior and qualitative categorization to the proposed error types.

1. **Type A Errors** are defined as being *non-annotated hyperbolas* in radargram images. Their counter phenomena, annotated hyperbolas without visual evidence, also falls in this category.
2. **Type B Errors** denote annotated hyperbolas having a *locational apex error* of a few pixels compared to the underlying radargram image.
3. **Type C Errors** correspond to a *wrongly assigned curvature* (estimated soil permittivity) information.

Most related work refers to Type A errors by measuring accuracy [Wilson *et al.*, 2007; Chen and Cohn, 2010], while Type B errors relate to RMSE – Root Mean Squared Error – scores on the apex positions (for which a prior matching of identified apex positions to the existing ground truth data is required) [Janning *et al.*, 2012a; Pasolli *et al.*, 2009; Janning *et al.*, 2012b]. Type C errors correspond to soil permittivity estimation [Simi *et al.*, 2008]. We presented a viable solution for its estimation once the apex is found [Busche *et al.*, 2012] and will not further focus on this error type / task here.

Our current, GPS-based annotation sets contain all three kinds of errors: (a) Type A errors are present through reflections, (b) Type B and Type C errors correspond to soil heterogeneity.

Figure 3 demonstrates a qualitative categorization of errors which were still contained in our human annotation sets after phase 1. In the Figure, each row corresponds to the same human annotator, while the same radargram patches are each used to visualize different annotation phenomena.

5.1 Assessing Differences in Annotation Behavior

To answer our first question, *whether or not a measurable difference between human annotation behavior exist*, we will first have a look at the quantity of pipes annotated (Type A errors), while thereafter having a closer look at Type B errors.

As our human annotations were derived from seed GPS-based annotation sets (annotations denoting previously known objects), we are able to measure two characteristic

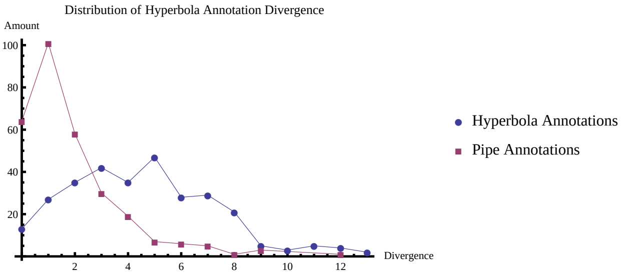


Figure 4: Inter human analysis of annotation behavior: The absolute differences exceed 12 hyperbolas (not shown here), resulting in diverging model performances at a later stage.

Visibility Loss	PE	Stoneware	Steel
Round 1	83%	41%	57%
Round 2	72%	36%	50%

Table 2: Percentage of pipe (type) annotations which are still present after the individual annotation rounds in the manual annotation sets (humans were able to identify visual evidence for this pipe type).

quantities for a given radargram, if two annotation sets are present: (a) For previously *existing* annotations, calculating the difference between the amount of removed pipes in the both annotation sets (by two different human annotators) indicates how well the final annotation sets match the data / the visual presence of reflection phenomenas. As an example, if both humans agreed on removing *the same* 5 annotations (denoting the same pipe annotations) from their annotation sets, the absolute difference of both annotation set sizes is 0. One can deduce that they both agreed on the fact that for 5 pipes, no visual evidence exists, whereas for all other pipe annotations, a corresponding visual evidence existed in the radargram. (b) For previously *unknown* hyperbolic shapes that have been *added* to the annotation sets, e.g., to mark reflections, measuring their absolute difference gives insights in whether or not these are identifiable by humans (distinguishable for background clutter). Take, as another example, a pair of annotation sets whose additional hyperbola annotations differ by 3: Then, one annotator was able to identify 3 more hyperbolic shapes on the radargram image.

Figure 4 visualizes the counts of differences between the size of two annotations sets for the same radargram image. Therein, *Pipe Annotation* denotes case (a) from above, and *Reflection Annotation* denotes case (b). We deduce from the high counts for low absolute differences in the *pipe reflections* case that *direct* pipe reflections are quite noticeable for humans. Contrary, the rather high counts for larger differences for *Reflection Annotations* (cf. Figure 3 (a) and (d)) indicate that the annotation of reflections is either more likely subject to personal taste, or is subject to ambiguity (e.g., the contrast is too low for humans to reliably distinguish them from the background).

As we have now identified a general difference depending on the type of annotation, a closer look at the fraction of remaining annotations per pipe type (the amount / kind of annotations *not* being removed) reveals certain notable characteristics, as shown in Table 2.

As can be seen, e.g., only 41% of the stoneware annotations being present in the GPS-based annotation sets were still existing after the first annotation round. Since humans were asked to retain only visible annotations, one may either deduce that (a) identifying these types is more challenging, (b) the current preprocessing techniques are not optimal for those pipe types, or (c) their absolute depth and filling or surrounding material causes masking effects, which are, compared with Figure 4, rather consistent for the individual pipe types.

	GPS vs. human	human vs. human	round
PE	18.3 ± 11.1	8.2 ± 11.0	1
	18.7 ± 13.7	7.3 ± 8.8	2
Stone-ware	21.6 ± 16.6	6.6 ± 4.9	1
	20.5 ± 19.0	7.3 ± 6.4	2
Steel	23.2 ± 15.7	9.7 ± 14.6	1
	22.7 ± 13.8	12.2 ± 16.8	2

Table 3: Averaged locational distances between pipe annotations in different annotation sets. Clearly visible is an inter-human agreement that pipes are not located at the GPS positions (distances to GPS are large), but at other locations (inter-human distances are small)

For sure, some of these differences are also influenced by experience and familiarity of the human annotators while working on GPR data. Before answering the second question, we are having a closer look at the individual locational differences when manually adjusting the apex positions for two cases (shown in Table 3): (a) The *GPS vs. human* comparison computes for all humanly created annotation sets for all annotations therein the RMSE / euclidean distance against the apex positions from the ground truth derived from the GPS measurements. (b) The *human vs. human* comparison computes RMSE differences on retained pipe annotations to compare how the individual manipulation of hyperbola apexes matches between humans. In other words, for both cases, we compare the distances of hyperbola apexes (type B errors) by assuming that lower distances correspond to more accurate annotation qualities, as annotations more closely match visual phenomena. This already gives first insights in inter-human agreements, that is, how close individual apex positions in two annotation sets created by humans for the same radargram are.

Even though these numbers are biased against the smaller quantity of pairs of annotation sets contained in the *human vs. human* comparison, a clear trend towards more consistent and accurate apex estimations for the inter-human comparison case is visible.

5.2 Gaining Inter-Human Agreed Annotation Sets

To answer our second question, *whether it is possible to get an inter-humanly agreed annotation set for complex engineering data*, we may both refer to table 3, indicating that there is a rather low average pixel-distance between human annotation sets, and present an indicative result when comparing the rankings of judges on their own annotation sets, that is, having a look at whether a judge favourizes his own annotation set over an annotation set of another human annotator.

Even tough we have only limited data (there are just three cases (2 annotators have also been judges in phase 1, whereas only 1 annotator was a judge in phase 2), we may take the following results as an indication: Ones own annotations in round 1 were preferred by the first annotator in $77/96$ (80%), the other in $46/66$ (70%) of all cases. Less indicative is round 2, for which $16/28$ (57%) annotations were preferred.

Even though we are able to derive that humans tend to favourize their own annotation sets, relating these to the appropriateness or suitability of the annotation sets to the *hyperbola identification task* is not easily possible: As we have seen above, a rather high deletion rate of annotations, esp. for certain pipe types, could be observed, even though it was well known to all participants that those pipes actually exist. The only valid conclusion which may be drawn here is that ones own *interpretation* on how to solve the pipe annotation task differs in a constant way between humans.

6 Impact of the Labeling Accuracy for Machine Learning Algorithms

We evaluated the quality of our annotation sets by performing a simple classification experiment using a state of the art classifier for patch-based classification, namely a Convolutional Neural Network as implemented in the eblearn library [Sermanet *et al.*, 2009]. The network structure is the well-known Lenet-5 network. We note that we are not primarily seeking for an optimal classification result, but aim at validating the suitability of the annotation sets, that is, whether the annotated apex locations obey an underlying structure in the radargram image which is easy to generalize.

Using the set of radargram images for which we obtained 'accurate' annotations, we created a dataset and splitted it in a leave-one-out fashion per individual radargram image. One radargram image was used for validating the classifier, whereas performance scores are reported on a test radargram image.

We used a grid search to determine an optimal hyperparameter combination for patches of size 32×32 as follows: The set of learning rates was set to $\{5, 1, 0.1, 0.5, 0.05\} \cdot 10^{-3}$, while different $l1$ and $l2$ regularizations, each being set to $\{0, 10^{-2}, 10^{-3}, 10^{-4}\}$, were tested as well. Positive training instances were created by using all patches being centered at annotated hyperbola apex positions, as well as using their neighbourhood, given that the amount of neighboring pixels in those patches were overlapping by 95%. Patches with an overlap between 95% and 30% were discarded to not introduce class boundary ambiguities. Negative training instances were randomly sampled from the remainder of the radargram image at a 2% rate, resulting in a class imbalance of approx. $1 : 7$, that is, seven times as much negative training instances than positive training instances, resulting in a baseline accuracy for a constant classifier being about 87.5%.

Table 4 shows the performance assessments for two dataset variants for two techniques to derive an optimal model given the performance scores on the validation set. The patches used to create the dataset may either be normalized, resulting in a pixel contrast range per individual patch over the whole greyscale from $[0, 255]$, or not normalized, for which the raw patches as present in a preprocessed radargram images are used. The algorithm as implemented in eblearn is trained by minimizing an 'energy' value. We determine a model to be used for evaluating its performance on the test set on both, the minimal energy value on the validation set (Accuracy (energy)), as well as on the maximal accuracy score on the validation set (Accuracy (correct)). Performance scores in Table 4 are both showing accuracy scores for comparability.

Our main aim is to compare the lift of the accurate annotations over those ones obtained from the GPS measurements. What can be seen is that for all four combinations, when combining both dataset variants with both performance assessment scores, our manual annotations increase the classifier performance. We conclude that our manually obtained annotation sets more closely match characteristic patterns within the radargram images, compared to those ones as obtained by the GPS measurement.

7 Conclusion and Future Work

This paper presented our methodology in annotating 700 GPR images, representing one example for the annotation task of complex engineering sensor data. Based upon the initial finding that the derivation of an accurate ground truth from a priori measured (GPS-) data is impossible for our current task at hand, we successfully showed that humans are able to improve the overall annotation quality. Anyhow, due to the still large fraction of 'inaccurately' annotated radargram images, we need to note that this overall process is costly, so that the question on the generalizability of this approach to other domains arises.

For our specific use case, we are now able to define certain subsets of the data, representing different 'agreement levels' between humans, that is, proportions of annotation sets for which a majority of the jugdes agree on their quality, as follows:

1. **Validation Dataset**: 57 'accurate' radargram images being well balanced between 3 human annotators compose a small dataset for which inter-human agreement exists on a high-quality annotation
2. **Scale-Up Dataset**: 139 incomplete, but not wrongly, annotated radargram images allow us to simulate automated analysis in semi-observed scenarios for evaluating performances under presence of noisy and partially ambiguous annotations.
3. **Inter- / Intra-Human Consistency checks**: For either set, models can be tested on their ability to generalize over either human annotators, or radargram images, in a controlled environment.

Besides of having an accurately labeled GPR data corpus, we showed that the annotated proportions within the underlying radargram images are more easily to generalize compared to the annotations derived from the GPS data. For reaching our goal of deriving supply maps, we are now able to proceed with improving state of the art Machine Learning Models for the detection of patches containing hyperbolic structures.

For similar use cases (in other application domains), the two main outcomes of this study are as follows: (a) Annotation quality improves at the cost of introducing inter-human disagreements to the annotations, and (b) Multiple evaluation metrices are desired to assess model performance. Though the quality and thus the suitability increases in general, the 'human factor' introduces ambiguities in the ground truth data. These need to be taken into account by designing and using evaluation measures which consider these aspects, e.g., by using an accuracy score that considers a locational displacement of a few pixels still as being correct.

normalized?	annotation	Accuracy (correct)	Accuracy (energy)
true	GPS	88.58 (3.72)	91.20 (1.80)
	accurate	89.55 (2.90)	92.57 (1.27)
false	GPS	87.00 (3.93)	90.70 (2.03)
	accurate	90.21 (2.35)	92.41 (2.19)

Table 4: The high quality of our manually obtained annotation sets is validated by performing a simple classification experiment comparing the annotations obtained by the GPS measurements with the ones obtained during the manual annotation rounds. Figures show accuracy scores and their variances in brackets.

Acknowledgments

Special thanks go to Mr. Marek Naser acting as our external expert to judge the annotations. This work is co-funded by the European Regional Development Fund project AcoGPR (Adaptive Contactless Ground Penetrating Radar) under the grant agreement no. WA3 80122470 (`http://acogpr.ismll.de`)

References

[Al–Nuaimy*et al.*, 2000] W. Al–Nuaimy, Y. Huang, M. Nakhkash, M.T.C Fang, V.T. Nguyen, and A. Eriksen. *Automatic detection of buried utilities and solid objects with GPR using neural networks and pattern recognition.* Journal of Applied Geophysics, Vol 43 Nr. 24, pp. 157–165, 2000.

[Birkenfeld, 2010] S. Birkenfeld *Automatic Detection of Reflexion Hyperbolas in GPR Data with Neural Networks*. World Automation Congress, pp. 1–6, 2010.

[Busche *et al.*, 2012] A. Busche, R. Janning, T. Horvath, L. Schmidt-Thieme *A Unifying Framework for GPR Image Reconstruction*. Proc. of the 36nd Annual Conference of the Gesellschaft fr Klassifikation (GfKl 2012), 2012.

[Cavanagh, 2011] P. Cavanagh *Visual cognition*. Vision Res. 51(13), pp. 1538-1551, 2011.

[Gregory, 1980] R.L. Gregory *Perceptions as Hypotheses*. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, Vol. 290, No. 1038, The Psychology of Vision, pp. 181–197, 1980

[Janning *et al.*, 2012a] R. Janning, T. Horvath, A. Busche, and L. Schmidt-Thieme *GamRec: a Clustering Method Using Geometrical Background Knowledge for GPR Data Preprocessing*. 8th AIAI, 2012

[Janning *et al.*, 2012b] R. Janning, T. Horvath, A. Busche, and L. Schmidt-Thieme *Pipe Localization by Apex Detection*. Proc. of the IET Intl. Conf. on radar systems (Radar), Glasgow, Scotland, 2012.

[Klebanov *et al.*, 2008] B.B. Klebanov, E. Beigman, and D. Diermeier *Analyzing disagreements*. Proc. of the WS on Human Judgements in Computational Linguistics (HumanJudge), pp. 2–7, 2008.

[Chen and Cohn, 2010] H. Chen, and A.G. Cohn *Probabilistic robust hyperbola mixture model for interpreting ground penetrating radar data*. IJCNN IEEE, pp. 1–8, 2010.

[Chen and Cohn, 2011] H. Chen, and A.G. Cohn *Buried Utility Pipeline Mapping Based on Multiple Spatial Data Sources: A Bayesian Data Fusion Approach.* IJCNN IEEE, 2411-2417, 2011.

[Giannopoulos, 2003] A. Giannopoulos *Modeling ground penetrating radar by GprMax.* Non Destructive Testing: Selected papers from Structural Faults and Repair, 755-762, 2003.

[McCarley *et al.*, 2004] J.S. McCarley, A.F. Kramer, C.D. Wickens, E.D. Vidoni, and W.R. Boot *Visual Skills in Airport-Security Screening*. Psychological Science, vol. 15 no. 5 pp. 302–306, 2004

[Mello-Thoms, 2006] C. Mello-Thoms *The problem of image interpretation in mammography: effects of lesion conspicuity on the visual search strategy of radiologists*. The British Journal of Radiology, 79, pp. 111-116, 2006

[Nowak and Rüger, 2010] S. Nowak, and S. Rüger *How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation*. In: Proc. of the Intl. Conf. on Multimedia Information Retrieval (MIR '10), pp. 557–566, 2010.

[Pasolli *et al.*, 2009] E. Pasolli, F. Melgani, and W. Donelli *Automatic Analysis of GPR Images: A Pattern-Recognition Approach.* IEEE Transactions Geoscience and Remote Sensing. Col 46, No. 7, pp. 2206–2217, 2009.

[Rolls, 2012] E. Rolls *Invariant visual object and face recognition: neural and computational bases, and a model, VisNet.* Front. Comput. Neurosci., 19, 2012

[Sermanet *et al.*, 2009] P. Sermanet, K. Kavukcuoglu, and Y. LeCun *EBLearn: Open-Source Energy-Based Learning in C++*. Proc. Intl. Conf. on Tools with Artificial Intelligence, IEEE, 2009

[Simi *et al.*, 2008] A. Simi, A. Bracciali, and G. Manacorda *Hough transform based automatic pipe detection for array GPR: Algorithm development and on-site tests.* Proc. of Radar, pp. 1–6, 2008.

[Volkmer *et al.*, 2005] T. Volkmer, J.R. Smith, and A. Natsev *A web-based system for collaborative annotation of large image and video collections: an evaluation and user study.* Proc. of the 13th annual ACM intl. conf. on Multimedia (MULTIMEDIA '05), pp. 892–901, 2005.

[Wilson *et al.*, 2007] J.N. Wilson, P. Gader, L. Wen-Hsiung H. Frigui, and K.C. Ho *A Large-Scale Systematic Evaluation of Algorithms Using Ground-Penetrating Radar for Landmine Detection and Discrimination.* IEEE Transactions on Geoscience and Remote Sensing, vol.45, no.8, pp. 2560–2572, 2007.