Probabilistic Robust Hyperbola Mixture Model for Interpreting Ground Penetrating Radar Data

Huanhuan Chen, *Member, IEEE* and Anthony G Cohn

Abstract— **This paper proposes a probabilistic robust hyperbola mixture model based on a classification expectation maximization algorithm and applies this algorithm to Ground Penetrating Radar (GPR) spatial data interpretation. Previous work tackling this problem using the Hough transform or neural networks for identifying GPR hyperbolae are unsuitable for on-site applications owing to their computational demands and the difficulties of getting sufficient appropriate training data for neural network based approaches. By incorporating a robust hyperbola fitting algorithm based on orthogonal distance into the probabilistic mixture model, the proposed algorithm can identify the hyperbolae in GPR data in real time and also calculate the depth and the size of the buried utility pipes. The number of the hyperbolae can be determined by conducting model selection using a Bayesian information criterion. The experimental results on both the synthetic/simulated and real GPR data show the effectiveness of this algorithm.**

I. Introduction

With the development of image processing, pattern recognition and computer vision, the fitting of primitive models to image data is an important technique for many industrial applications. Three methods, the moment method [7], the Hough transform [17] and the least-squares method [13], are often employed for this task. The moment method and Hough transform are especially applicable for fitting relatively simple models. Their application to a complex model with a number of parameters involves expensive computation. In this paper, we will consider least squares hyperbola fitting algorithms.

There are several conic fitting algorithms in the literature [3], [13], [20], [19]. However, most of these algorithms can only identify one conic in each image and most are sensitive to outliers. These two shortcomings greatly inhibit practical applications as most real-world image data is contaminated by noise and the data often contain several conics in each image.

To address these two problems and to ensure a fast implementation, this paper proposes a probabilistic hyperbola mixture model using a robust orthogonal distance fitting algorithm and applies the proposed algorithm to an important application area: Ground Penetrating Radar (GPR) data interpretation.

Ground Penetrating Radar has been widely used as a non-destructive tool for the investigation of the shallow subsurface, and is particularly useful in the detection and mapping of subsurface utilities and other solid objects [10]. However, GPR displays are not easily interpreted and only experts can extract significant information from GPR images to make a reliable report after the inspection.

The pattern shapes in the B-scans [4] of GPR data are determined by the propagation of short pulses into a medium with certain electrical properties. Typically, two patterns, hyperbolic curves and linear segments, are often observed in the GPR image: the hyperbolic curves are due to objects with cross-section size of the order of the radar pulse wavelength; the linear segments stem from planar interfaces between layers with different electrical impedances.

As GPR is becoming more and more popular as a shallow subsurface mapping tool, the volume of raw data that needs to be analyzed and interpreted is causing more of a challenge. There is a growing demand for automated subsurface mapping techniques that are both robust and rapid. This paper provides appropriate techniques for this.

The current tools that have been developed to aid in GPR data interpretation are generally computationally expensive, using Hough Transform [17] or neural network based algorithms [2], [1], and inadequate for on-site applications.

In our previous work [8], we have extended a swift conic fitting algorithm for GPR data interpretation. However, the previous algorithm is based on algebraic distance fitting that is sensitive to outliers. Although the proposed probabilistic model [8] can alleviate the problem to some extent, the algorithm is not applicable for GPR data with a relatively large amount noise.

This paper will address this problem by extending a robust conic fitting algorithm based on orthogonal distance fitting in the probabilistic mixture model; the proposed algorithm can be operated in real time. Other benefits of the proposed algorithm include relative robustness to noise compared with previous conic algorithms and automatic determination of the number of hyperbolae by a Bayesian information criterion.

The remaining parts of this paper are organized as follows. Section II will present some relevant works while the algorithm description is described in Section III. The experimental results are reported in Section IV. Finally, conclusions are drawn in Section V.

II. Background

With the development of GPR, there are several published works dealing with the automatic detection of patterns associated with buried objects in GPR data. These algorithms can be grouped into three main categories: 1) Hough transform

Huanhuan Chen is with the School of Computing, University of Leeds, Leeds, LS2 9JT, UK (phone: +44 113 343 5769; email: H.H.Chen@leeds.ac.uk).

Anthony G Cohn is with the School of Computing, University of Leeds, Leeds, LS2 9JT, UK (phone: +44 113 343 5482; email: A.G.Cohn@leeds.ac.uk).

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

based methods, 2) machine learning based methods and 3) clustering based algorithms.

The Hough transform [17] is a feature extraction technique used in image analysis to find imperfect instances of objects within a certain class of shapes by a voting procedure in the parameter space. The classical Hough transform was concerned with the identification of lines in the image, but later the Hough transform has been extended to identifying positions of arbitrary shapes, most commonly circles or ellipses. Hough transform based methods can identify the four parameters related to the hyperbola, which facilitates subsequent estimation of the pipe size and depth of the buried assets [24], [5]. However, this method often needs to run thousands of times with different combinations of hyperbola parameters (a, b) to search the best fit hyperbola shape and this usually cannot be deployed in real-time applications. How to specify a suitable threshold for the number of votes to determine the number of hyperbolae in the image is another problem with this kind of algorithm.

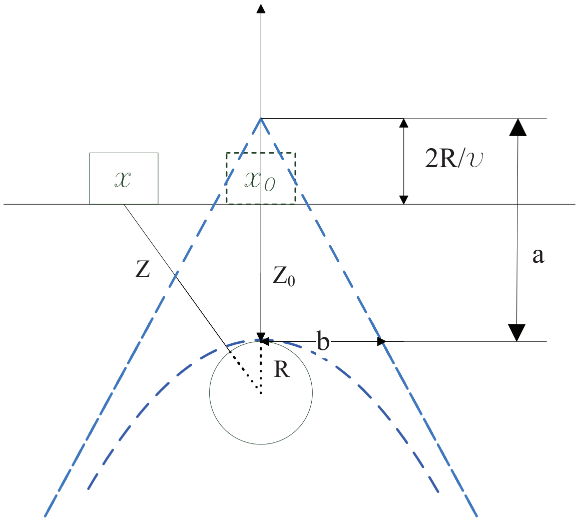
There is some work that uses machine learning methods to estimate the size and the depth of the buried pipes. However, with different mediums, soil types, materials of the pipes, the reflected patterns in GPR data are often different. In the real-world setting, it is very difficult to acquire the training data for different settings. For example, Pasolli et al. only use simulated data to train the neural networks [18] and this method greatly limits the practical applications.

Some work has been done to use a clustering approach to identify the hyperbolae. In [9], the authors applied a wavelet-based procedure to reduce noise and to enhance signatures in GPR images and then used a fuzzy clustering approach to identify hyperbolae. However, this kind of method will not reveal the hyperbola parameters (a, b) and cannot estimate these parameters related to the buried assets using the geometric model.

In order to address the above problems, this paper employs a robust conic fitting algorithm based on orthogonal distance as the basic hyperbola fitting algorithm. A probabilistic hyperbola mixture model is constructed to consider multi-hyperbola in a single image and the noise, including the feature noise around the hyperbolae and the background noise. The model is based on a classification expectation maximization (CEM) algorithm [6]. Since it is fast, the algorithm can be deployed in real-time applications. This algorithm can also be trivially extended to identification of other conic mixtures, such as ellipses and parabolas, thus extending the applicability of the proposed algorithm to other real-world applications.

III. Probabilistic Conic Mixture Model

In this section, we will present some related knowledge on GPR modelling, the robust conic fitting algorithm and the probabilistic conic mixture model. In the following subsections, we will present the GPR model description, conic fitting algorithm, the probabilistic model, the classification EM algorithm and model selection method using a Bayesian information criterion.



This hyperbola is characterised by a and b

Fig. 1. The GPR Geometric Model

A. GPR Model Description

The hyperbolic signatures in GPR data are often formulated as a geometric model [23], which is shown in Figure 1. The relation between the two-way travel time t, the horizontal position x and the velocity of propagation v can be expressed by

$$\left(\frac{t+\frac{2R}{v}}{t_0+\frac{2R}{v}}\right)^2 - \left(\frac{(x-x_0)}{\frac{v}{2}t_0+R}\right)^2 = 1, \tag{1}$$

where (x_0, t_0) are the coordinates of the target, $z = \frac{v}{2}$ and $z_0 = \frac{v_0}{2}$. Equation (1) is the equation of a hyperbola centered around $(x_0, \frac{-2R}{v})$.

Relating Equation (1) with a general hyperbola,

$$\frac{(y-y_0)^2}{a^2} - \frac{(x-x_0)^2}{b^2} = 1, \tag{2}$$

and with some simple derivations, the following relations can be obtained:

$$a = t_0 + \frac{2R}{v}, \tag{3}$$

$$b = \frac{v}{2}(t_0 + \frac{2R}{v}). \tag{4}$$

If the parameters related to the hyperbola (a, b) can be found, the depth and the radius can be obtained by the following equations:

$$R = \frac{b(a-t_0)}{a}, \tag{5}$$

$$depth = \frac{vt_0}{2} = \frac{bt_0}{a}. \tag{6}$$

This model assumes that a long cylinder is buried in a homogenous medium and the movement of the GPR antenna is perpendicular to the cylinder.

Since most of the pipes are long and linear, in practice, the operator of GPR machine always operates in a perpendicular direction to the assumed direction of the cylinder unless it is suspected that there are T-junctions or the pipes change the direction[1]. The other assumption for the homogenous medium can be satisfied if these pipes are located in the shallow subsurface.

B. Hyperbola Fitting based on Orthogonal Distance

In this section, we will introduce a robust hyperbola fitting algorithm based on orthogonal distance fitting.

The orthogonal distance is invariant to transformations in Euclidean space and it exhibits a more robust behavior than the algebraic distance. This algorithm is based on a minor revision of the work [14]. As we know, the parametric form commonly used for a south opening hyperbola can be presented as

$$x = c_1 + a\sinh\varphi \tag{7}$$

$$y = c_2 - b\cosh\varphi \tag{8}$$

Given a set of data points $(x_i, y_i)_{i=1}^m$, the distance d_i of a point $P_i = (x_i, y_i)$, which is not on the hyperbola, can be expressed by

$$d_i^2 = \min_{\varphi_i}\left[(x_i - x(\varphi_i))^2 + (y_i - y(\varphi_i))^2\right], \tag{9}$$

where the point $(x(\varphi_i), y(\varphi_i))$ is the nearest corresponding point of P_i on the hyperbola.

Now we want to determine c_1, c_2, a and b for this hyperbola by minimizing

$$\min \sum_{i=1}^{m} d_i^2. \tag{10}$$

In practice, we can simultaneously minimize $\varphi_1, \cdots, \varphi_m, a, b, c_1, c_2$ to find the minimum of the quadratic function

$$\begin{aligned} & Q(\varphi_1, \cdots, \varphi_m, a, b, c_1, c_2) \\ = & \sum_{i=1}^{m}\left[(x_i - x(\varphi_i))^2 + (y_i - y(\varphi_i))^2\right]. \end{aligned} \tag{11}$$

This is equivalent to solving the nonlinear least squares problem

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} a\sinh\varphi_i \\ b\cosh\varphi_i \end{pmatrix} \approx 0, \quad \text{for } i = 1, \cdots, m.$$

Minimizing the sum of squares of the distances of the given points to the best hyperbola is equivalent to solving the nonlinear least squares problem. Then we have $2m$ nonlinear equations for $m+4$ unknowns: $\varphi_1, \cdots, \varphi_m, a, b, c_1, c_2$.

Then the Gauss-Newton iteration will be employed to solve this minimization problem. If a good initialization is given to the Gauss-Newton method, the algorithm will converge in a few iterations (usually less than 10). In practice, the hyperbola solution obtained from the algebraic distance fitting is used as the initialization. The experiments confirm that this initialization is appropriate and is robust against large noise (We will illustrate the comparisons in the experimental section).

Based on the formulation in this section, this robust fitting algorithm can fit other conic functions and the combinations of different conic functions, such as elliptic and hyperbolic mixture model. In this case, the proposed probabilistic conic mixture model can be used for other conics.

C. Probabilistic Model

In practical applications, GPR images are often contaminated with noise. Although various kinds of pre-processing techniques have been proposed to reduce the noise level, it is impossible to guarantee that the processed GPR data is free from noise. In order to take noisy spatial points into consideration, we model two kinds of spatial noise in the proposed probabilistic algorithm. These two kinds of noise include background noise, in the form of observed points which are not part of the hyperbolae and feature noise, which is the deviation of the observed hyperbolic points.

Suppose that X is a set of observation points, and M is a partition consisting of hyperbolae, $M_0, M_1, \cdots, M_K$, where partition M_k contains N_k points. The background noise is denoted by M_0.

In the proposed model, we assume that the background noise is uniformly distributed over the region of the image, which is equivalent to Poisson background noise, and the hyperbolic points are distributed uniformly along the true underlying hyperbola; that is, their orthogonal distances follow a normal distribution, with mean zero and variance σ_j^2.

The resulting model becomes a hyperbolic mixture model with the mixing probability π_k $(0 < \pi_k < 1, k = 0, 1, \cdots, K$, and $\sum_{k=0}^{K} \pi_k = 1)$. Then the likelihood can be presented by

$$L(X|\pi, \sigma) = \prod_{i=1}^{N} L(x_i|\pi, \sigma), \tag{12}$$

where $L(x_i|\pi, \sigma) = \sum_{k=0}^{K} \pi_k L(x_i|\pi_k, \sigma_k, x_i \in M_k)$ and

$$L(x_i|\pi_k, \sigma_k, x_i \in M_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{\|f_k(x_i)\|^2}{2\sigma_k^2}\right),$$

where $f_k(x_i)$ is the orthogonal distance from the point (x_i, y_i) to the kth hyperbola.

For background noise, the likelihood can be expressed by

$$L(x_i|\pi_0, \sigma_0, x_i \in M_0) = \frac{1}{Area},$$

where $Area$ is the area of the image.

[1] Utility map records, which although notoriously inaccurate, at least in the UK, generally give the rough direction of the line of the buried apparatus (which is typically along the line of the road).

D. Classification Expectation Maximization Algorithm

The classification expectation maximization (CEM) algorithm is a classification version of the well-known EM algorithm [11]: it incorporates a classification step between the E-step and the M-step of the EM algorithm using a maximum a posteriori (MAP) principle. We now present the CEM algorithm as applied to the classification problem for points as described above.

Firstly, we give the number of hyperbolae in the GPR data and start with an initial partition using the k-means algorithm.

1) Begin with an initial partition.
2) (M-step) With the configuration of the current partitions, fit a hyperbola to each partition and then compute the maximum likelihood estimates (π_k^m, σ_k^2) for $k = 1, \cdots, K$.

$$\pi_k^m = \frac{\#\pi_k^{m-1}}{N},$$

and

$$\sigma_k^2 = \frac{1}{\#\pi_k^{m-1}} \sum_{x_i \in \pi_k^{m-1}} (f_k(x_i) - \bar{f}_k(x_i))^2,$$

where $f_k(x_i)$ is the orthogonal distance from the point (x_i, y_i) to the kth hyperbola, and π_k^m is the mixing probability of kth hyperbola in iteration m. π_0^m can be estimated by $\frac{\#\pi_0^{m-1}}{N}$.
3) (E-step) Based on the current hyperbolae and parameter estimates, calculate the likelihood of each point being in each partition.

$$t_k^m(x_i) = \frac{\pi_k^m L_k(x_i)}{\sum_{k=0}^{K} \pi_k^m L_k(x_i)},$$

where $L_k(x_i) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\|f_k(x_i)\|^2}{2\sigma_k^2}\right)$, $k = 1, \cdots, K$ and $L_0(x_i) = \frac{1}{Area}$.
4) (Classification step) Assign each point x_i to the partition which provides the maximum posterior probability $t_k^m(x_i)$, $0 \leq k \leq K$, (if the maximum posterior probability is not unique, we choose the partition with the smallest index).
5) Check for convergence: end or return to Step 2.

After calculating the probability of each point being in each partition, we assign each point into the partition for which it has the highest likelihood. Note that at the end of each iteration, the likelihood of the model will be calculated. Since the classification expectation maximization iterations sometimes decrease the likelihood, the process is executed for a predetermined number of iterations, and we choose the model with the highest overall likelihood as the final result.

E. Bayesian Information Criterion for Model Selection

Similarly to other mixture models, the hyperbolic mixture model needs to specify the number of hyperbolae at the beginning. The usual strategy is to search a range for the number of hyperbolae k and select the best one based on proper model selection methods.

In this paper, we propose to use a Bayesian information criterion (BIC) [22] for model selection among a class of parametric models with different numbers of parameters.

The model selection based on BIC can be seen as a form of regularization since it is possible to increase the likelihood by adding additional parameters in the maximum likelihood estimation, which may lead to overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model.

Model selection based on BIC provides (asymptotically) consistent estimators of the probability distribution given a data set [21]. This approach works well in practice for mixture models and other model-based clustering problems [15], [21]. The BIC for a model with K hyperbolae and background noise is defined by:

$$BIC = 2\log(L) - M\log(N),$$

where $M = K(DF + 2) + K + 1$ is the number of parameters, DF is the degrees of freedom used in fitting a hyperbola; there are four degrees of freedom in each hyperbola, i.e. $DF = 4$. The number of hyperbolae is K; for each hyperbola, we need to estimate σ_j, and we fit a hyperbola using four degrees of freedom. There are K parameters associated with the mixing proportions[2] and one more parameter is used for image area estimation. The larger the BIC, the more the model is favoured by the data.

In this paper, we utilize our previous algorithm based on algebraic distance fitting as the initialization. In the experimental sections, we emphasize that the number of data points in the primary hyperbola is significantly more than the number of noise points.

IV. Experimental Study

In order to examine the proposed algorithm, this section conducts several experiments on simulated and real GPR images. The precision and the computational cost are also analyzed in this section.

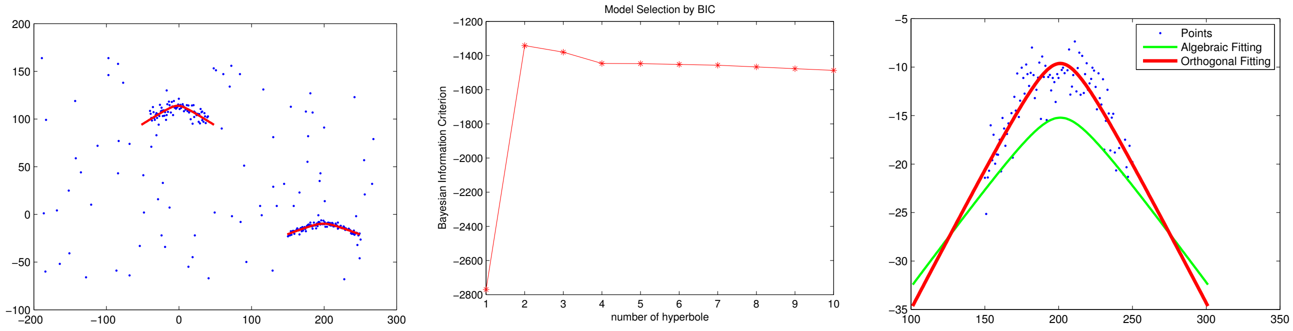
A. Synthetic Data

In this subsection, two synthetic data sets with two and eight hyperbolae respectively are generated with Gaussian white noise, respectively. These hyperbolae are positioned in different locations and have similar shape to reflect the case with real GPR data. Since in GPR B-scan images, the shape of the hyperbola is only determined by the medium where objects are buried [12] and if we assume that the medium does not change dramatically over a small neighbourhood, then the reflected hyperbolae should have similar shapes.

Figures 2 and 3 illustrate the results. From these figures, it can be seen that the proposed algorithm successfully identifies these hyperbolae and ignores the noise points.

To select the most appropriate number of hyperbolae for this data set, we run the algorithm with different k, which is the number of hyperbolae in the data set, and record the

[2] Although there are $M + 1$ mixing coefficients, the constraint $\sum_{k=0}^{K} \pi_k = 1$ reduces one degree of freedom.

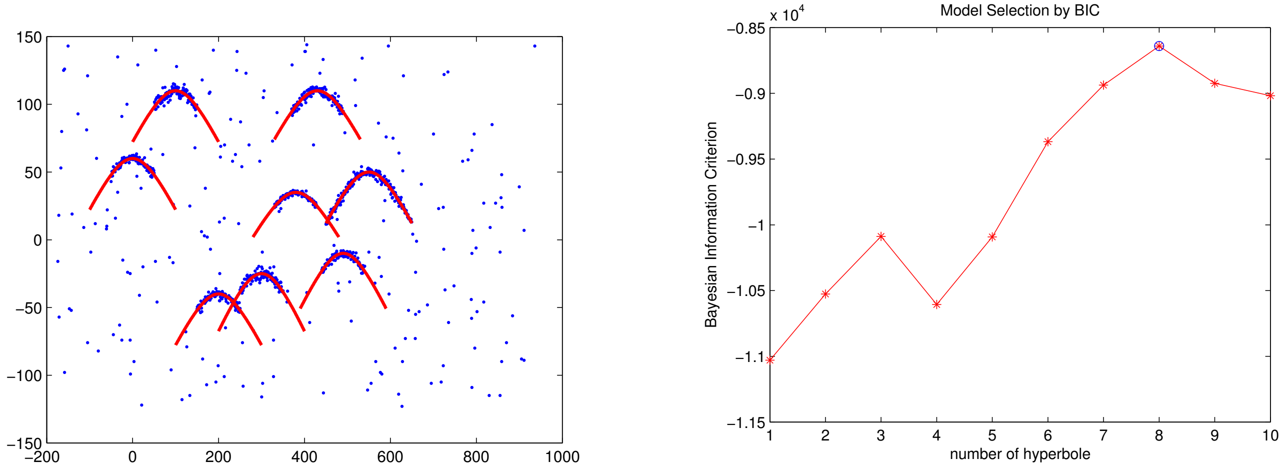


(a) Toy Data

(b) Bayesian Information Criterion

(c) Comparisons between algebraic and orthogonal distance

Fig. 2. The Toy Problem and Model Selection by BIC



(a) Toy Data

(b) Bayesian Information Criterion

Fig. 3. Hyperbola Identification for Synthetic Data

BIC value. The result confirms that two and eight hyperbolae are the appropriate number of hyperbolae for these two data sets, respectively.

The BIC value may not be monotone increasing before it reaches the maximum as some simplified models with few parameters could be better than some relatively complicated models due to the data distribution. This explains the situation that the BIC value with 4 hyperbolae is smaller than that with 3 hyperbolae in Figure 3(b).

B. Real GPR Data

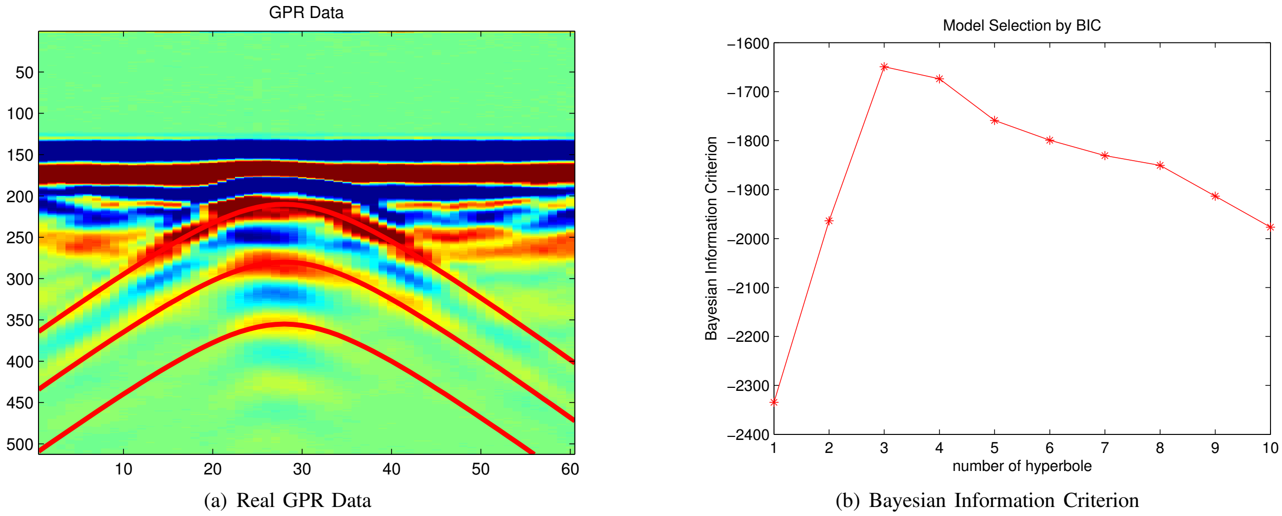
In this subsection, we utilize a real GPR data set to validate the proposed algorithm. The B-scan image is illustrated in Figure 4. In this figure it can be seen that the data set is challenging since it contains significant noise and has two secondary hyperbolae. The secondary hyperbolae are often generated by the reflection of the bottom part of the buried assets. The intensity of secondary hyperbolae are usually determined by the size, material and the depth of buried assets and other factors of the medium. In order to process this data, we preprocess the GPR data to reduce the noise using wavelet, remove the background to delete the linear reflection of ground (upper part in Figure 4), and reduce the clutters.

After the preprocessing step, the proposed algorithm is applied to this data. Figure 5 illustrates the results. From this figure, it can be seen that the proposed algorithm successfully identifies these hyperbolae. Based on the BIC figure, we also notice that the proposed algorithm with over-estimated k often generates a higher BIC value than the mode with under-estimated k. This is due to the existence of substantial noise. We will select the model with relatively small k from some candidate models whose BIC values are similar in the subsequent processing[3].

C. Simulated GPR Data

In practice, at least in the UK, utility records rarely contain depth information (not withstanding its potential usefulness), so evaluating the depth estimate from our algorithm without

[3] In practice, in the utility sector since buried apparatus is typically linear and relatively long in length. Further evidence for the number of hyperbolae/buried objects will come from repeated GPR measurements at regular intervals (typically at least three scans are taken 1m apart along the length of the suspected apparatus.) By integrating the evidence over these multiple scans, the estimate on k can be further improved.



(a) Real GPR Data

(b) Bayesian Information Criterion

Fig. 5. Hyperbola Identification for GPR Data

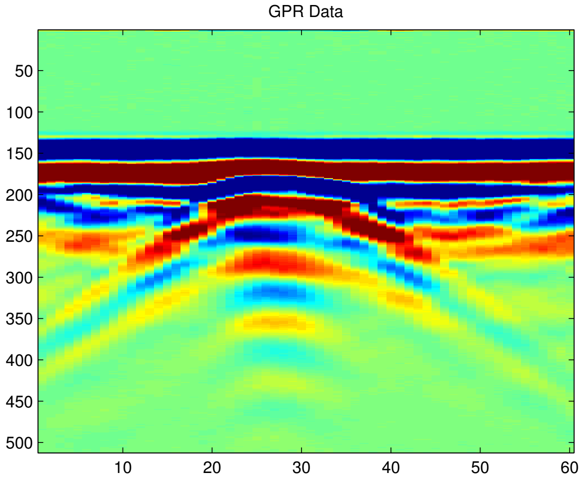


Fig. 4. B-scan GPR Data

physical excavation is difficult. Size information is usually present in the statutory records, but can not be completely relied upon for accuracy. To resolve this problem, in this section we employ simulated GPR data to estimate the accuracy for estimation of the depth and size of buried assets.

The simulated GPR data is generated by means of the electromagnetic simulator GprMax [16]. GprMax was developed on the basis of the finite-difference time-domain (FDTD) numerical method. It discretizes Maxwell's equations in both space and time and obtains an approximate solution directly in the time domain through an iterative process. GprMax allows the user to specify different mediums, such as clay, soft sand, concrete, and different sizes of buried objects with varied diameters.

In Figure 6, we show one example to specify the buried assets and the obtained simulated GPR Data. Note that all the buried assets are assumed to be cylinders. In order to validate our algorithm, we have generated ten simulated GPR datasets like Figure 6 with varied pipe sizes in varied mediums. In each GPR data set, there are ten buried pipes. The radii of these pipes range from 4cm to 20cm and the depths range from 40cm to 120cm. Note that in order to obtain good results, we need to generate a relatively higher resolution GPR data to guarantee the number of data points in the primary hyperbola is significantly more than the number of noise points.

In this experiment, the previous algorithm Hyperbolic-algebraic, which uses the algebraic distance to fit hyperbola, one variant of the proposed algorithm, hyperbolic-k-means, and the classical algorithm, Hough transform, are included for comparison.

The hyperbolic-$k$means algorithm obtains these hyperbolae according to the principle of k-means and assigns the points to the hyperbola with the shortest algebraic distance. Since we could not define a likelihood function for the non-probabilistic model, hyperbolic-k-means, BIC will not be used in the hyperbolic-k-means algorithm and k will be chosen as the same value as the one in Hyperbolic-mixture, which is optimized by BIC. Since hyperbolic-k-means does not take the probabilistic model into consideration, it is very sensitive to noise.

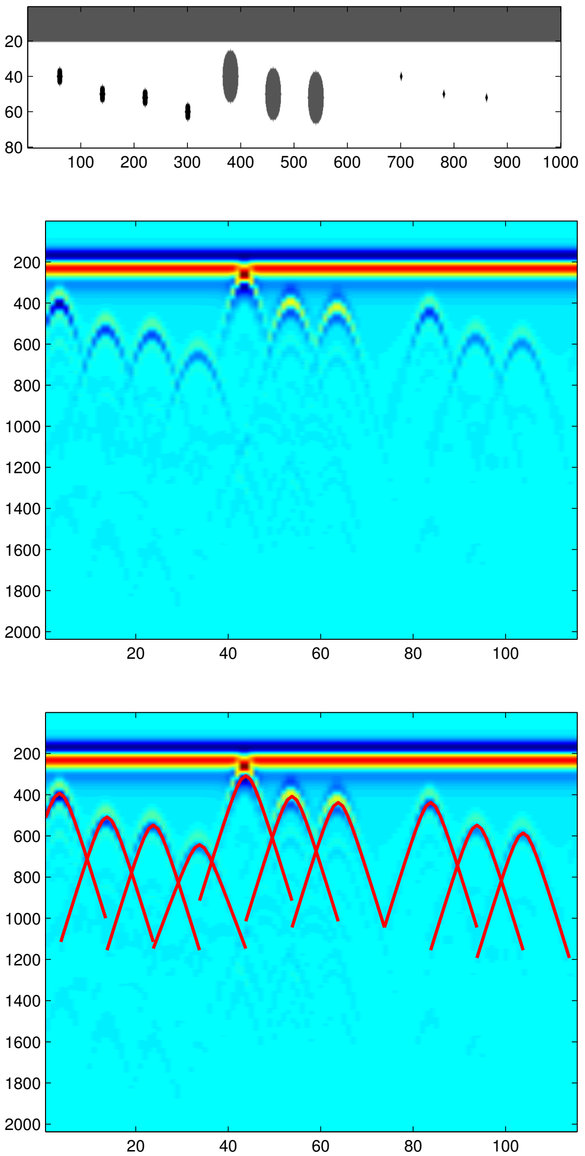
The Hough transform is a classical feature extraction technique used in image analysis to find imperfect instances of objects within a certain class of shapes by a voting procedure in the parameter space. In this application, we need to run thousands of Hough transforms with different combinations of hyperbola parameters (a, b) to search the best fit hyperbola shape. In the parameter space of the Hough transform, how to choose a suitable threshold for the number of votes to extract the number of hyperbolae in the image is usually a problem. In this experiment, we just use the same value k as the hyperbolic mixture model selected by BIC as the guide line to select the corresponding threshold for Hough transform.

Table I reports the statistical results of the experiment.

TABLE I

THE COMPARISON BETWEEN FOUR ALGORITHMS ON HYPERBOLA IDENTIFICATION AND HYPOTHESES EXTRACTION FROM SIMULATED GPR DATA. THE ACTUAL NUMBER OF HYPERBOLAE IN THE DATA SET IS 100.

Algorithm	Hyperbola Identified (#)	k selected by BIC	Running Time	Depth Error (%)	R Error (%)
Hyperbolic-Orthogonal	94	103	1.7s	2.8	2.3
Hyperbolic-algebraic	93	117	0.8s	5.7	4.7
Hyperbolic-k-means	51	103(fixed)	0.3s	16.3	14.9
Hough Transform	89	103(fixed)	226.1s	4.9	4.2



(c) Simulated GPR Data

Fig. 6. The Buried Assets, the simulated GPR Data and the hyperbolae identified by our algorithm

The proposed algorithm manages to identify 94 out of 100 hyperbolae and for the identified hyperbolae, the obtained hypotheses on the depth and size are quite accurate. The number of hyperbolae k, selected by BIC, is a little greater (103) than 100. This is because the secondary hyperbolae of some pipes with large diameters, buried in shallow subsurface, have large intensity even after a series of pre-processing steps, such as the background removal, noise reduction by wavelet and clutter reduction.

The algebraic distance fitting algorithm with probabilistic model identifies a similar number (93 out of 100) of hyperbolae as Hyperbolic-Orthogonal algorithm. However, the k selected by BIC is not as accurate as Hyperbolic-Orthogonal algorithm. This might be caused by the inaccurate estimation of the likelihood value because of the inaccurate estimation of hyperbola parameters. We also notice that the depth and the R error of Hyperbolic-algebraic is larger than the error of Hyperbolic-Orthogonal algorithm.

Compared with our algorithm, hyperbolic-$k$means, which does not incorporate the probabilistic model, only identifies 51 hyperbolae out of 100 and the calculated hypotheses are significantly worse than our algorithm in terms of the accuracy. It is also worth mentioning that both algorithms operate almost in real-time. Based on this experiment, the probabilistic conic mixture model achieves satisfactory performance in terms of the accuracy and time.

The performance of the Hough transform is fair in terms of the Hyperbola Identify and the depth/R error. However, the running time is significantly longer than other algorithms. This is because we need to conduct the grid search for a good combination of hyperbola parameters (a, b). As a robust algorithm, the Hough transform also suffers from another problem, which is how to specify the size of the suppression neighborhood. This is the neighborhood around each vote peak in the parameter space that is set to zero after the peak is identified. In the experiment, we used an empirical value of 25.

Although the probabilistic model incorporating the orthogonal distance fitting runs a little more slowly compared to the algebraic distance, the other performance indicators compare favourably.

This algorithm can achieve such a good performance since the robustness has been enhanced in two ways: (a) the robust orthogonal distance fitting can deal with the feature noise (the noise points around the hyperbola) and (b) the probabilistic model handles the background noise and the partitions nicely.

V. Conclusions

Previous algorithms for mining GPR data are unsuitable for on-site applications due to their computational complexity or the difficulty of obtaining sufficient appropriate training data for neural network based methods. We have addressed both of these problems in this paper.

In order to develop a novel GPR data mining algorithm, we extend an existing robust *single* hyperbola fitting algorithm by incorporating a probabilistic hyperbola mixture model and employing the classification expectation maximization algorithm for the final solution.

The proposed algorithm significantly contributes to both theoretical research and practical application areas. From a theoretical point of view, this research extends the existing single hyperbola fitting algorithm to a multiple hyperbola fitting algorithm and provides a robust solution compared to the previous hyperbola fitting algorithms. In this paper, we mainly focus on mixtures of hyperbolae although the proposed search can be trivially extended for other conics, such as ellipses and parabola.

For practical applications, the proposed techniques provide an effective and accurate GPR data interpretation tool, which is adequate for on-site applications. This is extremely useful for the advanced multi-channel GPR system that often generates volumes of data in each task. Therefore, this research will potentially play an important role in the GPR and related industries, such as utility detection, infrastructure and transportation industries.

In this paper, a robust hyperbola fitting algorithm based on orthogonal distance fitting is employed for robust and real-time detection of buried infrastructure. This algorithm is more robust than our previous algorithm that incorporates the algebraic distance. As remarked earlier, one way to further improve the results will be to incorporate evidence from multiple scans along the length of a suspected linear object.

Acknowledgement

This research is supported by an EPSRC grant (EP/F06585X/1): Mapping The Underworld: Multi-sensor Device Creation, Assessment, Protocols.

References

[1] 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*, 43(2-4):157–165, 2000.

[2] W. Alnuaimy, Y. Huang, M. Nakhkash, M. Fang, V. Nguyen, and A. Eriksen. Automatic detection of buried utilities and solid objects with GPR using neural networks and pattern recognition. *Journal of Applied Geophysics*, 43:157–165, 2000.

[3] F. L. Bookstein. Fitting conic sections to scattered data. *Computer Graphics and Image Processing*, 9(1):56–71, 1979.

[4] C. Bruschini, B. Gros, F. Guerne, P. Y. Pièce, and O. Carmona. Ground penetrating radar and imaging metal detector for antipersonnel mine detection. *Journal of Applied Geophysics*, 40(1-3):59–71, 1998.

[5] L. Capineri, P. Grande, and JAG Temple. Advanced image-processing technique for real-time interpretation of ground-penetrating radar images. *International Journal of Imaging Systems and Technology*, 9(1):51–59, 1998.

[6] G. Celeux and G. Govaert. A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

[7] B. B. Chaudhuri and G. P. Samanta. Elliptic fit of objects in two and three dimensions by moment of inertia optimization. *Pattern Recognition Letters*, 12(1):1–7, 1991.

[8] H. Chen and A. G. Cohn. Probabilistic conic mixture model and its applications to mining spatial ground penetrating radar data. In *Workshops of SIAM Conference on Data Mining (WSDM10)*, 2010.

[9] S. Delbo, P. Gamba, and D. Roccato. A fuzzy shell clustering approach to recognize hyperbolic signatures in subsurface radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3):1447–1451, 2000.

[10] A. Dell'Acqua, A. Sarti, S. Tubaro, and L. Zanzi. Detection of linear objects in GPR data. *Signal Processing*, 84(4):785–799, 2004.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Of The Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[12] A. Dolgiy, A. Dolgiy, and V. Zolotarev. Optimal radius estimation for subsurface pipes detected by ground penetrating radar. In *Proceedings 11th International Conference on Ground Penetrating Radar, Columbus, Ohio, USA*, volume 4, 2006.

[13] A. Fitzgibbon, M. Pilu, and R. B. Fisher. Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):476–480, 1999.

[14] W. Gander, G. H. Golub, and R. Strebel. Least-squares fitting of circles and ellipses. *BIT Numerical Mathematics*, 34(4):558–578, 1994.

[15] A. Gasgupta and A.E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441), 1998.

[16] A. Giannopoulos. Modelling ground penetrating radar by GprMax. *Construction and Building Materials*, 19(10):755–762, 2005.

[17] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer vision, graphics, and image processing*, 44(1):87–116, 1988.

[18] E. Pasolli, F. Melgani, and M. Donelli. Automatic analysis of GPR Images: a pattern-recognition approach. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2206–2217, 2009.

[19] M. Pilu, A. Fitzgibbon, and R. Fisher. Ellipse-specific direct least-square fitting. In *Proceedings of International Conference on Image Processing (ICIP'06)*, volume 3, 1996.

[20] J. Porrill. Fitting ellipses and predicting confidence envelopes using a bias corrected Kalman filter. *Image and Vision Computing*, 8(1):37–41, 1990.

[21] K. Roeder and L. Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1995.

[22] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[23] S. Shihab and W. Al-Nuaimy. Radius estimation for cylindrical objects detected by ground penetrating radar. *Sensing and Imaging: An International Journal*, 6(2):151–166, 2005.

[24] C. G. Windsor, L. Capineri, and P. Falorni. The estimation of buried pipe diameters by generalized hough transform of radar data. In *Proceedings Progress In Electromagnetics Research Symposium (PIERS)*, pages 22–26, 2005.