A BAYESIAN REFLECTION ON SURFACES

The multiresolution inference of continuous-basis fields

DAVID R. WOLF
NASA Ames Research Center
MS 269-2 Moffet Field
CA 94035-1000†

Abstract. The topic of this paper is a novel continuous-basis field representation and inference framework applied to the inference of continuous surfaces from measurements (for example camera image data). Traditional approaches to surface representation and inference are briefly reviewed. The new field representation and inference paradigm is then introduced within a maximally informative (MI) (see [1]) inference framework. The knowledge representation is introduced and discussed in the context of MI inference. Then, using the MI inference approach, the here-named Generalized Kalman Filter (GKF) equations are derived. The GKF equations allow the update of field knowledge from previous knowledge *at any scale*, and new data, to new knowledge *at any other scale*. The GKF equations motivate a location-dependent scale or multigrid approach to the MI inference of continuous-basis fields.

Several problems are uniquely solved: The MI inference of fields, where the basis for the field is itself a continuous object and generally is not representable in a finite manner; the tradeoff between accuracy of representation in terms of information learned, and memory or storage capacity in bits; the approximation of probability distributions so that a maximal amount of information about the object being inferred is preserved by the approximation.

Key words: Field, inference, manifold, surface, Bayesian inference, Kalman Filter, Generalized Kalman Filter, multigrid, maximally informative inference, stochastic process, adaptive scale inference, knowledge representation, minimum description length

1. Introduction

1.1. TRADITIONAL SURFACE REPRESENTATION METHODS

Many methods for representing surfaces have been utilized previously, however all of these methods involve representing the surface by a *discrete* field, perhaps with

†Email: greywolf@ptolemy.arc.nasa.gov, Web: http://dino.ph.utexas.edu/ ~ wolf

a deterministic interpolation defined (bi-linear, tensor B-splines, etc.) to provide a definition for the surface at points intermediate to the discrete field. Probability distributions or densities of these discrete fields then often take the form of normalized exponentials of sums of clique energy functions, and produce a construct commonly known as a Markov Random Field. (See Geman [2], for the most often cited example.) There are several immediate observations on these approaches:

- The surface remains unspecified at points intermediate to the discrete field, except by the often undefined notion of interpolation.
- When interpolation is *not* defined, the discrete field probability distribution says nothing about the probability distribution of surface at points intermediate to the discrete field points.
- When interpolation is defined then, given a value of the discrete field, there is *no uncertainty* in the surface intermediate to the discrete field points. This is generally not a good representation of knowledge of physical fields.
- The surface distribution is not an intrinsic property of any physical object, rather a post-hoc imposition of the analyst attempting a useful regularization. The necessary scaling properties are ignored. See next section.

1.2. SCALING CONSISTENCY

The consistency condition mentioned in the last section, which must be imposed on probability distributions for continuous fields is:

Scaling of sample points consistency: For $S \subset A$, both sets being discrete sets of indices of regions of support of the continuous field,

$$P(X_S) = \int P(X_A)\, dX_{A\setminus S}. \tag{1}$$

There is one note on notation that needs to be made here: probability distributions of continuous objects are defined operationally throughout this paper, meaning that here the explicit distribution of a continuous object is accessible only via discrete marginalizations of the continuous object distribution, as in (1). That this can be done consistently within a rich domain is one of the achievements of continuous stochastic process theory.

1.3. ELEMENTS OF THE NEW PARADIGM

In the rest of this paper we discuss a new approach to continuous field inference which corrects the deficiencies, including the intermediate value and scaling problems, of traditional discrete-basis approaches to discrete height fields, for example. There are four objects of central importance within the inference approach described in this paper, one of which is a new object to Bayesian inference:

- The prior distribution for field. The prior holds all information about the field before any data is observed.
- The likelihood distribution. The likelihood is predictive for data, given the field. It incorporates all of the physics of the measurement process.

- The knowledge-representation (KR) distribution. Within the usual Bayesian point of view, the KR distribution is a new object. In the paradigm described in this paper the KR distribution is the object being updated when new data arrives. The KR distribution and the prior are the only objects remaining after the new data has been used to update.
- The posterior distribution. The posterior distribution summarizes everything known about the field given the prior knowledge and all data. The prior distribution and the KR distribution determine an approximation (possibly exact) to the surface posterior distribution.

In section 2 we make explicit the continuous field representation and inference paradigm via an example on surface inference.

2. Surface representation and inference

In this section the main ideas of the field representation and inference paradigm presented in this paper are given via the example of continuous-basis height field inference. The height field taken here may be thought of as a scalar function of a two-dimensional domain of support. However, the technique is general, and it will become obvious that it extends to arbitrary-topology, arbitrary-support, arbitrary-dimension continuous fields.

2.1. SURFACE DISTRIBUTIONS

The surface and height field distributions (the prior, likelihood, and posterior surface and height field distributions) are discussed in this section. As to notation, $\boldsymbol{s}$ represents a continuous-basis height field (surface), $\boldsymbol{h}$ represents a vector of heights (necessarily discrete), and $\boldsymbol{v}$ is reserved for the position basis-vector for those heights. The symbol $\boldsymbol{x}$ is reserved for measurements, θ represents parameters for the prior and ϕ represents parameters which determine the measurement operator, M. The symbol P is reserved for probability distributions or densities, as appropriate, while later $\hat{P}$ is associated with the KR distribution. The vector delta-function is defined as the product of the delta-functions for the components as in $\boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s},\boldsymbol{v})) = \Pi_{i=1}^{n} \delta(h_{v,i} - h_i(\boldsymbol{s},\boldsymbol{v}))$.

2.1.1. *Surface and height field prior distributions*

Consider a set S of surfaces where each element $\boldsymbol{s} \in S$ is a height field, i.e. such that $\boldsymbol{s} = s(x,y)$ is scalar function of two variables. Write the prior probability distribution for surfaces in S given the parameters θ which determine the prior distribution as

$$P(\boldsymbol{s} \mid \theta). \tag{2}$$

Note, as mentioned before, continuous probability distributions like that of (2) are used operationally by marginalizing to a discrete basis. Justification for this treatment of probability distributions over continuous-basis fields appears in [3]. Consider a vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ of discrete (x,y) points, $v_i = (x_i, y_i)$, and for any given surface $\boldsymbol{s}$ denote the associated vector of heights by $\boldsymbol{h}(\boldsymbol{s},\boldsymbol{v}) =$

$(h_1(\boldsymbol{s},\boldsymbol{v}),\ldots,h_n(\boldsymbol{s},\boldsymbol{v}))$. Write the prior distribution of the surface heights at the chosen points v as $P(\boldsymbol{h}_v \mid \theta)$. This discrete height distribution may be found as follows:

$$
\begin{aligned}
P(\boldsymbol{h}_v \mid \theta) &= \int P(\boldsymbol{h}_v \mid \boldsymbol{s},\theta)\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (3)\\
&= \int P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (4)\\
&= \int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s},\boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (5)
\end{aligned}
$$

Now, given that what is known is the surface heights $\boldsymbol{h}_v$ at a vector $\boldsymbol{v}$ of discrete (x,y) points, the posterior distribution of surfaces is found from Bayes' theorem as

$$
\begin{aligned}
P(\boldsymbol{s} \mid \boldsymbol{h}_v,\theta) &= \frac{P(\boldsymbol{h}_v \mid \boldsymbol{s},\theta)\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)} && (6)\\
&= \frac{P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)} && (7)\\
&= \frac{\boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s},\boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)}{\int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s},\boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s}} && (8)
\end{aligned}
$$

where the denominator distribution was found in (5). Note that in going from (4) to (5) (and from (7) to (8)) the fact that every surface determines the heights on itself everywhere leads to the delta-function representation for the conditional $P(\boldsymbol{h}_v \mid \boldsymbol{s})$, just another way of saying that the surface is certain everywhere, given itself

2.1.2. *Measurements: The Likelihood*

In general, a surface $\boldsymbol{s}$ and some other parameters ϕ not dependent upon $\boldsymbol{s}$ (i.e. camera point spread function, camera position and direction, lighting position and direction, etc.) specify the probability distribution for data (likelihood)

$$
P(\boldsymbol{x} \mid \boldsymbol{s},\phi,\theta) = P(\boldsymbol{x} \mid \boldsymbol{s},\phi) \tag{9}
$$

where the data distribution is independent of θ once $\boldsymbol{s}$ is known. In practice an approximate model for the surface is utilized to make the computation of the likelihood tractable, as in a rendering step. Much confusion will result if the reader fails to distinguish the surface model for approximating the likelihood with the actual surface on which the likelihood depends.

2.1.3. *Conditioning on data*

Conditioning on data is necessary when updates to the distribution functions of the last section are to be found. Given data, we may utilize the likelihood of the previous section and steps similar to those of (3)–(5) and (6)–(8) to find the data conditioned distributions $P(\boldsymbol{s} \mid \boldsymbol{x},\phi,\theta)$, $P(\boldsymbol{h}_v \mid \boldsymbol{x},\phi,\theta)$, and $P(\boldsymbol{s} \mid \boldsymbol{h}_v,\boldsymbol{x},\phi,\theta)$. We refer the reader to [3] for the details.

2.2. KNOWLEDGE REPRESENTATION

The solution to the surface representation problem presented here addresses the competition for representational resources (computer memory) issue in a unique manner. To notation, Θ represents parameters for the KR distribution, later we specialize to $\Theta = (\boldsymbol{\mu}, \Sigma)$. The full posterior may be written in the form

$$P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta) \quad = \quad \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \boldsymbol{x}, \phi, \theta)\, P(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)\, d\boldsymbol{h}_v \tag{10}$$

where the distributions inside the integral appear in section 2.1.3. The following approximation defines the KR distribution. The prior conditioned on a set of heights, along with a new distribution, the *knowledge representation* distribution $\hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)$, are substituted for the distributions inside the integral of 10 to approximate the posterior as

$$\hat{P}(\boldsymbol{s} \mid \ \hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)\,) \quad = \quad \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)\, \hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)\, d\boldsymbol{h}_v. \tag{11}$$

Commentary on conditioning on the KR distribution, and a proof of the ability of the KR distribution to achieve a good approximation to the posterior appears in [3]. In practice, it is useful to take a multinormal distribution over a discrete-point height field as the KR distribution; then the parameters for the KR distribution may taken as $\Theta_v(\boldsymbol{x}) = (\boldsymbol{\mu}_v(\boldsymbol{x}), \Sigma_v(\boldsymbol{x}))$ so that the posterior for surfaces is then approximated by

$$\hat{P}(\boldsymbol{s} \mid \Theta_v, \theta) = \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)\, \hat{P}(\boldsymbol{h}_v \mid \Theta_v)\, d\boldsymbol{h}_v. \tag{12}$$

2.3. UPDATING THE KNOWLEDGE REPRESENTATION

Now we discuss updating the KR parameters Θ_v when new data are acquired. Then the GKF equations, which are the appropriate parameter update when either or both the scale and position of the basis are allowed to vary, are written. Temporarily restrict attention to the fixed $\boldsymbol{v}$ case.

2.3.1. *Bayes' theorem*

Having acquired $\Theta_v^n = \Theta_v(\boldsymbol{x}^n)$, from previously seen data $\boldsymbol{x}^n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and upon seeing new data $\boldsymbol{x}_{n+1}$, the goal is to find Θ_v^{n+1} such that the surface distribution given Θ_v^{n+1} approximates the surface distribution given $\boldsymbol{x}_{n+1}$ and Θ_v^n. Given new data $\boldsymbol{x}_{n+1}$ in the context of the previously seen data $\boldsymbol{x}^n$ summarized by Θ_v^n, our updated surface distribution is found via Bayes' theorem

$$\begin{aligned} \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \quad &= \quad \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \Theta_v^n, \phi, \theta)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \phi, \theta)}{\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta)} \\ &= \quad \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)}{\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta)} \end{aligned}$$

$$= \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)}{\int P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)\, d\boldsymbol{s}} \tag{13}$$

where we defined

$$\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta) = \int P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)\, d\boldsymbol{s}. \tag{14}$$

The updated posterior $\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)$ will be approximated by the Θ_v^{n+1} parameterized KR distribution of (12) as

$$\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta) = \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)\, \hat{P}(\boldsymbol{h}_v \mid \Theta_v^{n+1})\, d\boldsymbol{h}_v. \tag{15}$$

The approximation condition for determining Θ_v^{n+1} is then written

$$\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta) \approx \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \tag{16}$$

Equation (16) suggests we minimize various measures of the closeness of the two distributions. For example, one measure is the average square difference of the two distributions,

$$\int |P_1(\boldsymbol{s}) - P_2(\boldsymbol{s})|^2\, d\boldsymbol{s} \tag{17}$$

but there is no good first-principles reason to use this form. In the next section we discuss the measure of distance which lead$\boldsymbol{s}$ to the *MI* choice of Θ_v^{n+1}.

2.3.2. *Maximally Informative inference*
The measure of distance which leads to the Θ^{n+1} providing the most information about the surface distribution is the *MI* choice for the statistic Θ^{n+1}. The condition for being MI, see [1], is

Find the Θ^{n+1} such that

$$\partial_{\Theta_v^{n+1}} \int \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)\, log \left(\frac{\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)}{\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta)} \right) d\boldsymbol{s} = \mathbf{0} \tag{18}$$

while at the Θ_v^{n+1} satisfying the derivative condition above the corresponding Hessian is negative definite and the extremum is a local maximum. If possible, choose the global maximum.

Note that the Kullback-Leibler distance above is asymmetric and that it is highly relevant which distribution contains the prior information and which distribution is being updated, a fact which many authors do not note. Maximum entropy techniques reverse the roles of the distributions which appear here. For a detailed explanation see [1]. In the following section the GKF update equations which follow from the MI approach are presented.

3. The Generalized Kalman Filter equations.

In this section the MI update equations for the KR, the GKF equations, are written. Details of the calculation and numerous appendices utilized in the derivation are found in [3]. Note the parameters of the updated KR (here $\overline{\boldsymbol{v}}$) need not have the same dimension nor position as those of the previous KR (i.e. $\boldsymbol{v}$), solving the problem of how to allow updates from one representation to the next, (finer, same, or coarser sampling), representation, i.e. it is not necessarily the case that $v_i \in \{\overline{v}_j\}$ or that $\overline{v}_i \in \{v_j\}$. Note that $\theta = (\boldsymbol{\mu}_s, \Sigma_s)$.

Denote the union of the components of $\boldsymbol{v}$ and $\overline{\boldsymbol{v}}$ by $\boldsymbol{v} \cup \overline{\boldsymbol{v}}$. Let A_{v_1,v_2} denote the projection from $\boldsymbol{v}_2$ to $\boldsymbol{v}_1$, for arbitrary position vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Skipping the mathematical details found in [3] the GKF equations are

$$\begin{aligned} \Theta_{\overline{v}}^{n+1} &= (\boldsymbol{\mu}_{\overline{v}}^{n+1}, \Sigma_{\overline{v}}^{n+1}) \\ \boldsymbol{\mu}_{\overline{v}}^{n+1} &= \boldsymbol{\mu}_{\overline{v}}^{R} \\ \Sigma_{\overline{v}}^{n+1} &= \Sigma_{\overline{v}}^{R} \end{aligned} \tag{19}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\overline{v}}^{R} &= \Sigma_R(\Sigma_Q^{-1}\boldsymbol{\mu}_{\overline{v}}{}^{Q} + (\Sigma_{\overline{v}}^{n})^{-1}\boldsymbol{\mu}_{\overline{v}}^{n} - \Sigma_{\overline{v}}^{-1}\boldsymbol{\mu}_{\overline{v}}) \\ \Sigma_R^{-1} &= \Sigma_Q^{-1} + (\Sigma_{\overline{v}}^{n})^{-1} - \Sigma_{\overline{v}}^{-1} \end{aligned} \tag{20}$$

and where

$$\begin{aligned} \boldsymbol{\mu}_{\overline{v}}^{Q} &= A_{\overline{v},v\cup\overline{v}}A_{v\cup\overline{v},s}\boldsymbol{\mu}_s^{P} \\ \Sigma_Q^{-1} &= A_{\overline{v},v\cup\overline{v}}A_{v\cup\overline{v},s}\Sigma_P^{-1}A_{v\cup\overline{v},s}^{T}A_{\overline{v},v\cup\overline{v}}^{T} \end{aligned} \tag{21}$$

$$\begin{aligned} \boldsymbol{\mu}_s^{P} &= \Sigma_P(\Sigma_s^{-1}\boldsymbol{\mu}_s + M^T\Sigma_\epsilon^{-1}\boldsymbol{x}_{n+1}) \\ \Sigma_P^{-1} &= \Sigma_s^{-1} + M^T\Sigma_\epsilon^{-1}M \end{aligned} \tag{22}$$

$$\begin{aligned} \boldsymbol{\mu}_{\overline{v}}^{n} &= A_{\overline{v},v}\boldsymbol{\mu}_v^{n} \\ (\Sigma_{\overline{v}}^{n})^{-1} &= A_{\overline{v},v}(\Sigma_v^{n})^{-1}A_{\overline{v},v}^{T} \end{aligned} \tag{23}$$

$$\begin{aligned} \boldsymbol{\mu}_{\overline{v}} &= A_{\overline{v},v}\boldsymbol{\mu}_v \\ \Sigma_{\overline{v}}^{-1} &= A_{\overline{v},v}\Sigma_v^{-1}A_{\overline{v},v}^{T} \end{aligned} \tag{24}$$

$$\begin{aligned} \boldsymbol{\mu}_v &= A_{v,s}\boldsymbol{\mu}_s \\ \Sigma_v^{-1} &= A_{v,s}\Sigma_s^{-1}A_{v,s}^{T} \end{aligned} \tag{25}$$

Equations (19)–(25) are the GKF update equations. Standard KF equations are discussed in many places, see for example [4]. Another helpful paper is [5]. Having these update equations allows one to consider updating a representation of any dimension and position relative to the original representation. Finally, the

GKF equations would not have been discovered without the MI inference approach, which led directly to the the correct interpretation of the Kullback Leibler distance, with application here. Evaluating the GKF update across new representations, a multigrid-like search algorithm for representation optimization, and the characterization of the measurement operator M for nonlinear systems are discussed in information-theoretic terms in [3]. This appears to provide the first information-theoretic justification for multi-grid approaches to surface or image inference.

4. Conclusion

Field inference has been generalized from the typical discrete fixed-basis setting to a continuous-basis setting. The problem of surface inference was solved in the context of continuous field inference. Using the approach of acquiring the MI KR distribution, the GKF equations were found. The GKF allows the updated KR parameters to be found at any scale and any position. The approach allows the learning of information at the relevant scales desired. It provides an information-theoretic justification for location-dependent adaptive multi-grid inference. To the knowledge of the author, this is the first time that the inference of continuous surface and field objects has been rigorously justified.

5. Acknowledgements

This work was supported by the NASA Ames Center for Excellence in Information Technology. I thank the members of the Ames Data Understanding group for their interest and comments, especially the invaluable contributions of Dr. Robin D. Morris, who painstakingly checked the results, and Dr. Peter Cheeseman, whose comments were thoughtful.

References

1. D. R. Wolf and E. I. George, "Maximally informative statistical inference," in *Bayesian Statistics, 6th Valencia Meeting*, vol. 6, 1998.
2. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intell.*, **PAMI-6**, (6), pp. 721–741, 1984.
3. D. R. Wolf, "A Bayesian reflection on surfaces," *http://dino.ph.utexas.edu/˜ wolf*, 1998.
4. R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering*, Wiley, NY, 1983.
5. H. Zhu, C. K. Williams, R. Rohwer, and M. Morciniec, "Gaussian regressian and optimal finite dimensional linear models," *http://www.ncrg.aston.ac.uk/*, 1997.