Bayesian Fusion for Multi-Modal Aerial Images

Alistair Reid
NICTA
Australian Technology Park
Eveleigh, NSW, 2015 Australia
Email: alistair.reid@nicta.com.au

Fabio Ramos
Australian Centre for Field Robotics
and School of Information Technologies
The University of Sydney, Australia
Email: fabio.ramos@sydney.edu.au

Salah Sukkarieh
Australian Centre for Field Robotics
The University of Sydney, Australia
Email: salah.sukkarieh@sydney.edu.au

***Abstract*—This paper presents a fusion method to combine aerial images from a low flying Unmanned Aerial Vehicle (UAV) with images of other spectral bands from sources such as satellites or commercial hyperspectral imagers. The proposed method propagates information from high-resolution images into other low-resolution modalities while allowing the images to have different spectral channels. This means the relationship between the high-resolution and low-resolution channels is expected to be non-deterministic, non-linear and non-stationary. A novel Gaussian Process (GP) framework was developed to define a stochastic prior over the estimated images. Its covariance function is computed to replicate the local structure of the high-resolution image, and allows the model to infer a high-resolution estimate from a low-resolution channel. Results are presented for natural images acquired by a UAV in a farmland mapping application.**

I. Introduction

Unmanned aerial vehicles (UAVs) are useful for geospatial exploration in applications such as ecological surveillance, agricultural management or mapping [6, 3]. UAVs are well suited to geospatial problems because they can obtain timely, high-resolution data. However, it is important to consider these advantages in the context of ubiquitous data modalities such as satellite imagery before investing in the development and deployment of a specialised UAV system. An effective image fusion algorithm allows the use of heterogeneous remote sensing strategies, combining UAVs with other sensor platforms, and would also benefit autonomous systems where the data acquisition strategy is planned [9, 11].

This paper presents an approach to fuse images from different modalities by combining information from a high-resolution image of one modality with a complementary low-resolution image with different spectral channels. We apply this technique to combine high-resolution images obtained from a UAV (using a generic colour camera) with multi-band reference imagery to predict new high-resolution modalities.

Image *super-resolution* (the estimation or fusion of images to increase their resolution) has received much attention in the computer vision community recently. However, imposing that the images measure different spectral information (for example, one image may be infrared, while another is RGB colour) leads to a very different multi-task learning problem where the degree of dependence between images is then induced by the latent spatial content of the scene. The observations of this dependence are mixed and indirect because of the change in resolution between modalities. This problem has been previously studied by the photogrammetry and geostatistics communities for fusing satellite data [2].

This paper presents a new Gaussian process (GP) fusion model that employs a novel covariance function as its structural prior. Our key contribution is the construction of a positive definite covariance function that provides three non-standard features. Firstly, it uses both spatial position and image intensity in its input space, allowing the model to derive contextual non-stationarity from the high-resolution reference image. Secondly, the new covariance function is defined over pixels of different sizes, as compared to a standard covariance function that is defined over point inputs only. This behaviour is achieved by convolving the underlying model over the spatial region of each observation, and allows the covariance function to unmix low-resolution pixels. Finally, the covariance model is sparse by design. This is critical for the model to scale to real images, as the high dimensionality of image data leads to an enormous number of potential dependencies that must be solved through matrix inversion. A high-resolution reconstruction of the low-resolution observations can be obtained by applying GP regression using the proposed covariance function.

Results are presented for both controlled simulation and practical application to the problem of combining high-resolution RGB images from our low flying Unmanned Aerial Vehicle (UAV) with co-registered infra-red reference images.

II. Related Work

Traditional image fusion approaches combine multiple images of a common scene using a sensor sensitivity model. This enables *super-resolution* images to be reconstructed that resolve detail beyond the Nyquist frequency of *any* of the observed images [8]. Super-resolution is an ill-posed inversion problem, so constraints on the solution space such as smoothness priors are introduced, allowing effective increases in resolution as supported by the data [14]. Unlike the work in this paper, typical super-resolution algorithms assume that the images have a common set of spectral channels.

Recent developments in the computer vision community have surpassed the traditional limitations of super-resolution reconstruction by exploiting the structural redundancy of natural images. This has led to solutions for many low level image processing problems such as de-noising and texture synthesis, increasing the resolution of a single input image [10], or creating plausible fill content using in-painting techniques [15]. Popular models of low level image structure include sparse image patch dictionaries [19], geometric partial differential equations [21], and probabilistic models such as Markov random fields that specify high order dependencies over pixel neighbourhoods [20]. GP models have also been used to model low level image structure [10]. While these approaches create convincing new visual detail, their purpose is image synthesis rather than sensor fusion. The reconstructed detail is one likely possibility from a large space of plausible images and unlike traditional super-resolution the synthesis is not necessarily supported by specific evidence in the observations, but rather from similar detail in the exemplar images. In this paper, a method is sought to infer the real scene as closely as possible by combining evidence from multiple modalities.

The fusion of images with different spectral channels is more frequently studied in the geostatistics community under the name *band-sharpening*. Some ad-hoc approaches such as luminance substitution or bilinear interpolation are still in common use, while state-of-the-art approaches transfer spatial pattern from the high-resolution image into the low-resolution image using a mapping in a feature space such as wavelet or curvelet coefficients [17]. These approaches rely on the feature representation to capture context, and the mappings themselves are usually stationary in the feature space. This makes it difficult, for example, for these approaches to handle objects of different colour that have the same greyscale appearance, and means that fusion errors in the latent features are transformed into texture artifacts in the image space [26].

The inter-modality fusion problem is also related to domain adaptation, where the training data is in a different modality to the testing data [13]. The key conceptual difference is that domain adaptation seeks to apply the training examples in another modality, while image fusion seeks to create a new complementary modality. GP models have also featured in domain adaptation problems, for example to recognise objects in images that have a different resolution to the labelled training data [7].

GP priors can also be used to directly encode inter-modality relationships between multiple output functions. The strategy is known in the machine learning community as multi-task GP regression [4, 24], and also in the geostatistics community as cokriging [2]. The method places a joint Gaussian prior over all the output data modalities, so inter-modality dependencies can be learnt if they can be incorporated into a cross covariance function.

While GP models usually operate on point data, recent work has also led to the development of strategies to accommodate a change of support where the inputs consist of different spatial geometries such as image pixels [2]. Our proposed approach also employs these strategies to handle the change of resolution. We model the low-resolution channel as a GP rather than placing all channels into the output space. This allows the high-resolution image channel to augment the *input* dimensionality of the covariance function, so local non-linear relationships may be inferred.

III. Preliminaries

We briefly review Gaussian processes (GP) regression as it is the main building block for our approach. A GP is a nonparametric Bayesian framework with appealing analytical properties that make it suitable to deal with images as a regression problem. It places a Gaussian prior distribution over the space of functions $f(\mathbf{x})$ mapping inputs $\mathbf{x}_i \in \mathbb{R}^D$ to outputs $y_i \in \mathbb{R}$, where $\mathbf{y} = f(\mathbf{x}) + \epsilon$ is a noisy observation, and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is a Gaussian distribution with zero mean and standard deviation σ. For N observations represented as $\{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$ and $\mathbf{y} = \{y_i \in \mathbb{R}\}_{i=1}^N$, and M query inputs $\{\mathbf{X}^*\} = \{\mathbf{x}_i^* \in \mathbb{R}^D\}_{i=1}^M$, the joint distribution over observations and query points is given by

$$\left[\begin{array}{c} \mathbf{y} \\ \mathbf{y}^* \end{array}\right] \sim \mathcal{N}\left(\mu\begin{pmatrix}\mathbf{X} \\ \mathbf{X}^*\end{pmatrix}, \left[\begin{array}{cc} k(\mathbf{X},\mathbf{X}) + \sigma_n^2 I & k(\mathbf{X},\mathbf{X}^*) \\ k(\mathbf{X}^*,\mathbf{X}) & k(\mathbf{X}^*,\mathbf{X}^*) \end{array}\right]\right) \quad (1)$$

where $\mu(\mathbf{x})$ is a mean function and $k(\mathbf{x}, \mathbf{x}')$ is a positive semi-definite kernel or covariance function. Conditioning the joint distribution on the observations, the mean and variance for query points $\mathbf{X}^*$ is given by

$$\mathbf{y}^* | \mathbf{X}, \mathbf{y}, \mathbf{X}^* \sim \mathcal{N}(\mu^*, \Sigma^*)\ , \quad (2)$$

where,

$$\begin{aligned} \mu^* &= k(X^*, X) K_X^{-1} \mathbf{y}, & (3) \\ \Sigma^* &= k(\mathbf{X}^*, \mathbf{X}^*) - k(\mathbf{X}^*, \mathbf{X}) K_X^{-1} k(X, X^*), & (4) \end{aligned}$$

with $K_X = k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I$. The hyper-parameters of the model θ include parameters for the mean and covariance functions as well as the noise term σ_n^2, and can be computed by maximising the log of the marginal likelihood (LML) defined as:

$$\log(p(\mathbf{y}|\mathbf{X}, \theta)) = -\frac{1}{2}\mathbf{y}^T K_X^{-1} \mathbf{y} - \frac{1}{2}\log|K_X| - \frac{N}{2}\log 2\pi. \quad (5)$$

LML is robust to over-fitting, as the first component seeks data fit, and is balanced against the second component that penalises model complexity [18].

IV. Non-stationary Image Fusion

We present a new GP approach to combine a high-resolution image with a low-resolution image of a different spectral band, to produce a new modality that has the spectrum of the low-resolution image but the resolution of the high-resolution image. The approach uses GP regression to infer the pixels of this new modality. This requires us to define a new covariance function model to address three key problems.

The first challenge is to define a covariance function to conduct inference over pixels of different resolutions (the change of support problem). This has been approached by extending a point-observation covariance function into a multi-task covariance function over areas using an integral kernel derivation [16].

The second challenge is defining a prior for the image structure. A GP will usually model smoothly varying spatial functions, but the image data are expected to be non-smooth, exhibiting discontinuities and spatial non-stationarity. While the high-resolution image is not an observation of the output we wish to model, it does contain cues as to this spatial structure. To use this information, the input space of the covariance function is augmented with the high-resolution pixel observations.

The final challenge is to ensure that the covariance function is tractable on image data (where there are many potential dependencies between pairs of pixel observations). This has been addressed by ensuring a high degree of sparsity in the covariance function. The design of this covariance function is outlined below.

A. Defining a Covariance over Areas

It is assumed that there is a continuous two-dimensional GP output f underlying the scene depicted in the low-resolution image. We may then model an image pixel as the result of observing the output function $f(x)$ over a discrete area A rather than at a point x, as depicted in Fig. 1.

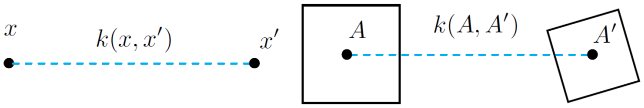

Fig. 1. While a standard covariance function is defined over points x, a covariance function over areas A may also be obtained.

A simple averaging relationship is then assumed to relate the pixel observation $P(a)$ to $f(x)$:

$$P(A) = \frac{1}{|A|} \iint_{x \in A} f(x) dx \tag{6}$$

where $|A|$ is the surface area of A. Now $f(x)$ is modelled as a Gaussian process over points ($f(x) \sim GP(\mu(x), k(x,x'))$). Here $k(x,x')$ defines the covariance between point x and point x'. By considering the observation model in Eq. 6, if x' is replaced with an area A', this is equivalent to integrating:

$$k(x,A) = \frac{1}{|A'|} \iint_{x' \in A'} k(x,x') dx. \tag{7}$$

This is a valid construction, as it has been established in the GP literature that covariance functions can be summed or convolved into more complex forms [18, 1]. Repeating the process yields a covariance between two areas:

$$k(A,A') = \frac{1}{|A||A'|} \iint_{x \in A} \iint_{x' \in A'} k(x,x') dx dx'. \tag{8}$$

For the image fusion problem, let A_H denote the geometry of the small *high-resolution* pixels, and A_L denote the geometry of the large *low-resolution* pixels. Discretisation of the two dimensional input space is used to simplify the computation of k across different areas as follows. Firstly, the covariance between two high-resolution pixels is given by:

$$k(A_H,A'_H) = \frac{1}{|A_H|^2} \iint_{x \in A_H} \iint_{x' \in A'_H} k(x,x') dx dx'. \tag{9}$$

Rather than defining $k(x,x')$ and integrating, the proposed approach is to directly define a base covariance $k(A_H,A'_H)$ between high-resolution pixels. Here $k(x,x')$ still exists, but is latent and unused. The design of $k(A_H,A'_H)$ is very important, as it basically defines a prior over high-resolution images, from which the output is predicted. This will be further explained in the next section. The large pixels A_L are then approximated by a set of high-resolution pixels A_h discretising the problem, as depicted in Fig. 2. Consequently, an integration over A_L with respect to x is equivalent to a sum of piecewise integrations over its constituent A_H (the integral has been avoided by defining the covariance $k(A_H,A'_H)$ directly),

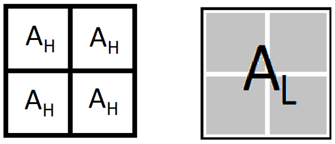

Fig. 2. The low-resolution pixels of area A_L are discretely integrated by assuming they correspond to a set of A_h areas.

$$k(A_H,A'_L) = \frac{1}{N_H} \sum_{A'_H \in A'_L} k(A_H,A'_H), \tag{10}$$

where N_H is the number of high-resolution A_H areas that compose A_L . The corresponding covariance between two low-resolution pixels is then given by:

$$k(A_L,A'_L) = \frac{1}{N_H N'_H} \sum_{A_H \in A_L} \sum_{A'_H \in A'_L} k(A_H,A'_H). \tag{11}$$

B. Defining the Image Prior $k(A_H,A'_H)$

In this section, the design of $k(A_H,A'_H)$ is described. While most GP covariance functions force the output to be smooth, the proposed function is well suited to images that contain discontinuities. Contextual non-stationarity is achieved by augmenting the input space, adding the intensity values of the supporting high-resolution image $P_H(A_H)$ and $P_H(A'_H)$ that occur inside of A_H and A'_H respectively. The augmented covariance function is constructed as the separable product of a positive definite spatial covariance function K_S and a positive definite pixel-intensity covariance function K_P:

$$k(A_H,A'_H) = \sigma_f^2 k_S(A_H,A'_H) k_P(P_H(A_H), P_H(A'_H)). \tag{12}$$

The role of k_S is to provide a smooth, sparse and local covariance envelope, while the role of k_P is to couple and decouple pixels within this envelope based on the contextual information present in the high-resolution image. σ_f is simply an amplitude hyperparameter. In this design, k_S is defined as a function of the midpoint coordinates $mid(A_H)$ and $mid(A'_H)$ using a sparse piecewise polynomial kernel that decreases to zero for displacements larger than the hyper-parameter λ_S [18]:

$$\begin{aligned} Z &= \tfrac{|mid(A_H) - mid(A'_H)|}{\lambda_S} \\ k_S(A_H,A'_H) &= \left\{ \begin{array}{cc} (1-Z)^3(3Z+1) & Z < 1 \\ 0 & Z \geq 1 \end{array} \right\}. \end{aligned} \tag{13}$$

Within this spatial envelope, the non-stationarity inducing K_P couples A_H to A'_H based on similarity in the high-resolution function, so that any smoothness or discontinuity can be transferred to the output image. K_P is defined as a squared exponential kernel with a hyperparameter λ_P to control sensitivity:

$$k_P(P_H(A_H), P_H(A'_H)) = \exp\left\{-\frac{(P_H(A_H) - P_H(A'_H))^2}{2\lambda_P^2}\right\}. \tag{14}$$

The hyperparameters of this model can be tuned by maximising the marginal likelihood Eq. 5. While the initial discussion has assumed that all images have a single channel (a single output function), this definition of k_P already supports a multi-dimensional P_H.

C. Image Reconstruction

The GP prior in Eq. 1 has been modified to suit the image fusion covariance function. Instead of X and y, the training data comes from the low-resolution image and consists of spatial areas A_L and the corresponding $P_L(A_L)$ pixels that were observed. Instead of point queries x^*, we query the model over the A_H of the high-resolution reference image (where P_H is defined). The image is obtained by querying the predictive mean of the GP model and performing a normalisation step as described below. A constant mean of $\mu = 0.5$ is assumed over the image (pixel outputs are continuous from 0 to 1). The predictive mean equivalent to Eq. 3 is now given by:

$$\begin{aligned} P_*(A_H) &= \\ &\mu + k(A_H, A_L)\left[k(A_L, A_L) + \sigma_n^2 \mathbf{I}\right]^{-1}(P_L(A_L) - \mu). \end{aligned} \tag{15}$$

In our results, an additional step has been taken to improve the image quality. It has been observed that the estimate of P_* has varying uncertainty due to the non-stationary covariance function: in non-smooth regions each pixel has less covariance with its neighbours than in the smooth parts. This causes the GP prediction to reduce its amplitude and return to its mean. While this is a reasonable behaviour in most regression problems, in an image context it has a negative impact on image contrast and quality. This effect can be removed by normalising the GP weights W to obtain corrected query pixels P_*^c:

$$\begin{aligned} W &= k(A_H, A_L)\left[k(A_L, A_L) + \sigma_n^2 \mathbf{I}\right]^{-1} \\ P_*^c(A_H) &= .5 + \tfrac{W(P_L(A_L) - 0.5)}{W\begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^\top}. \end{aligned} \tag{16}$$

Finally, P^c should lie in the range $(0, 1)$, and inspection of the fused output images over a range of test cases has shown that pixel intensities do (rarely) exceed this range. In a classification problem, this would be handled using a sigmoidal likelihood function [12]. However, there is no principled reason for an image function to lie between 0 or 1. It is the observations from a camera that have limited dynamic range. Consequently, the outputs are clipped, and allowed to occasionally saturate as seen in digital photographs.

D. Computational Aspects

It is important to consider whether the proposed method can scale to image data that contains a potentially large number of high-resolution pixels N_H and low-resolution pixels N_L. The covariance matrix over the low-resolution observations must be inverted at a nominal cost of $O(N_L^3)$. However, our covariance matrix is sparse because each row has at most N_K non-zero entries where N_K is the number of low-resolution pixels contained in the sparse envelope of k_S. Additionally, because of the grided structure of pixels in images, the resulting covariance matrix has its non-zero elements close to the diagonal, making it well suited to strategies such as incomplete Cholesky decomposition. For example, inference with a 102400×102400 covariance matrix with 27-connections per row has been solved in approximately 5 seconds on a core i7-3520M. In practice, the cost of computing the covariance function using Eq. 11 is often equal or larger to the cost of inversion. This requires $\frac{1}{2}N_K N_L$ elements to be computed, each involving $O(N_{KH}^2)$ operations where N_{KH} is the number of dependencies between high-resolution pixels. Thus the cost of populating the covariance matrix is linear with respect to the number of image pixels, and cubic with respect to the chosen degree of sparse dependencies.

V. Experiments

A. Example Problem

This example fuses a single-channel high-resolution image with a multi-channel colour image at $\frac{1}{16}$th the linear resolution. The multiple output channels are independently queried using a common covariance function parameterisation on different observations, so each channel can be queried at little additional cost. We have modelled the C_r and C_b channels (of a $YCbCr$ decomposition of the image) that have a complicated relationship to the shading channel. The fused full-resolution colour image is depicted in Fig. 3, and because the presence of luminance information can bias human perception, the estimated colour bands are also visualised for independent assessment.

The chrominance channels in Fig. 3 reveal that much of the detail has been successfully reconstructed, even if there is no simple mapping between shading and color. Object boundaries visible in the luminance channel have been unmixed in the chrominance image, although clearly this will only work where there is contrast between adjacent object surfaces.

B. Benchmarking

This section describes a controlled experiment constructed using aerial data to benchmark the proposed fusion algorithm against the state of the art. A set of 40 non-overlapping image tile pairs was draw from a mosaicked image database, across three different survey sites. These image pairs provide a visible band (of which the green channel was extracted) and a co-registered near-infrared (NIR) band. The relationship between the green and NIR spectral channels is difficult to model, as it exhibits local contrast inversions where an object that is dark in the visible spectrum reflects strongly in the infrared

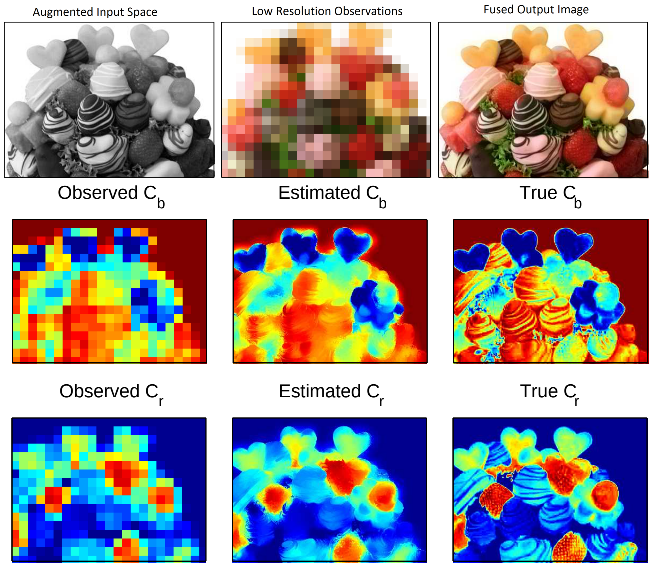

Fig. 3. Fusion of a coarse colour image with a fine greyscale image. The resulting reconstruction closely resembles the true output, but because human perception is easily misled by the presence of shading, the observed, true and reconstructed components of red and blue chrominance are shown separately.

spectrum. The NIR images were artificially downsampled by a range of magnification factors to construct a controlled testing scenario. The following key strategies were selected for comparison: *Bilinear Interpolation* has been included here as a lower bound on image quality; *Wavelet Fusion* is a fast popular technique from the geoscience literature where images are wavelet transformed and coefficients of the high-resolution image are merged with those from the low-resolution image using an inter-modality model [17]. *Downscaling Cokriging*, is a probabilistic data fusion framework that uses a multi-task, multi-resolution prior [2]; The *proposed* GP algorithm is also a probabilistic framework but we have designed a more flexible non stationary image prior.

Two metrics were used to compare the algorithm outputs to the true reference: Mean Squared Error (MSE) and *Universal Image Quality* (UIQ) [25]. MSE provides an overall quality indicator with a score closer to zero indicating higher algorithm performance. UIQ on the other hand is a statistical measure designed to penalise loss of correlation, luminance distortion, and contrast distortion, and an algorithm with a score closer to 1.0 has performed better. Statistics of the image quality scores as a function of magnification factor are presented in Fig. 4 and Fig. 5.

As expected, output image quality degrades for all the algorithms as a function of magnification because the problem becomes increasingly under-constrained. The UIQ scores in Fig. 4 show that our proposed method is out-performing the other approaches. We attribute this to the ability of the proposed GP framework to locally infer a non-stationarity relationship between the modalities. Other fusion approaches, including the wavelet benchmark, form an explicit relationship (in pixel or feature space) between the image modalities, and will therefore struggle with objects that have the same appearance in the visible channel but different infrared responses. Downscaling-cokriging handles this problem in a more principled way by treating the channels as dependent random processes, although the dependencies are usually stationary and linear.

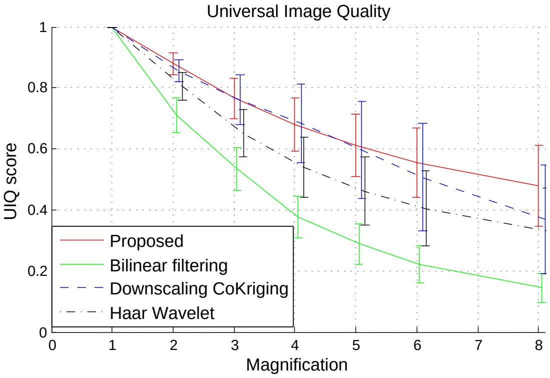

Fig. 4. The mean and error bounds of evaluating the Universal Image Quality (UIQ) index [25] of the reconstructed output compared to the true reference image over 40 test images as a function of algorithm and magnification factor.

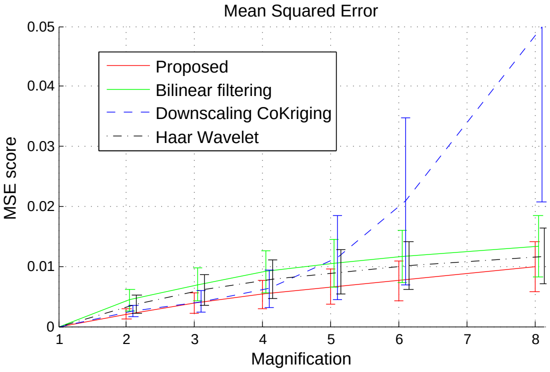

Fig. 5. The mean and error bounds of evaluating the mean squared error (MSE) index [25] of the reconstructed output compared to the true reference image over 40 test images as a function of algorithm and magnification factor.

Another trend is shown in the mean squared error (Fig. 5). Because our proposed method weights observations of the output band, it will remain faithful to the low-resolution image as the problem becomes ill-posed. This is also true of the bilinear filtering (which only uses the low-resolution image), and approximately true for the wavelet approach because it is only merging high frequency coefficients. The cokriging approach, on the other hand, increasingly relies on the *high-resolution* image for its predictions and may have low frequency errors that MSE is sensitive to.

C. UAV Experiments

This section applies the proposed fusion method to combine aerial images from different sensor platforms acquired as part of a project to survey remote farmland environments. High spatial resolution images were acquired by an Unmanned Aerial Vehicle (UAV) to resolve features such as individual trees, fences and waterways. Over the same region, low-resolution multispectral imagery was acquired from a manned aircraft, providing cues to vegetation and soil properties.

To acquire the high-resolution modality, a UAV was developed to carry a payload consisting of an inertial measurement unit (IMU), global positioning system (GPS) unit, and a colour (RGB) camera. The aircraft and its payload box are depicted in Fig. 6. It was also necessary to implement fully autonomous flight capabilities on this platform, as the UAV flight paths were along narrow swaths, often taking the UAV beyond the visual range of its operators. The platform is ruggedised for operating from dirt roads rather than runways, and a ground station consisting of a differential GPS base station and telemetry computers were set up in the field for its operation.

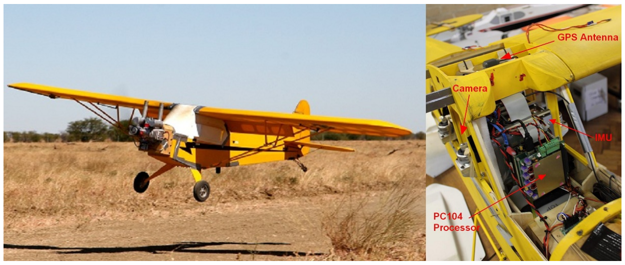

Fig. 6. Left: The UAV platform used to acquire geo-referenced aerial data. Right: The platform payload including a camera, GPS unit, Inertial Measurement Unit (IMU) and PC104 stack.

The UAV was operated 500m above ground to obtain a resolution of up to 20cm/pixel over the farmland environment. These images have been mosaicked to a ground resolution of 1m/pixel, a resolution in which key scene features can be identified, while georeferencing errors and motion blurring that might interfere with the fusion are minimal.The low-resolution image dataset covers the same region of interest, and was acquired from a manned aircraft carrying a multi-spectral imager. This data provides a wider choice of informative infrared spectral bands than the UAV data, but only at a relatively coarse resolution of 6 meters per pixel. For this experiment, three infrared spectral bands have been inferred, corresponding to 800, 1200 and 1700nm wavelengths. Visualisations of these bands are valuable for investigating vegetation stress and vegetation type.

While the UAV and multispectral datasets cover the same spatial region, they were collected from different platforms and have different geo-referencing errors. The spatial registration/alignment of the modalities is currently approached using normalised cross correlation [5] on the luminance outputs (because both platforms have sampled the visible spectrum) - as shown in Figure 7. This approach has provided an acceptable registration accuracy to conduct the experiment, but is simply a starting point for further development. In addition to registration error, the fusion algorithm faces distortion induced by a time difference between the commercial image acquisition and the UAV deployment. This delay could lead to occulation in the data - where objects or shadows have entered or left the scene [22]. As this is real data, the UAV mosaic also has gaps in coverage that were not predicted over.

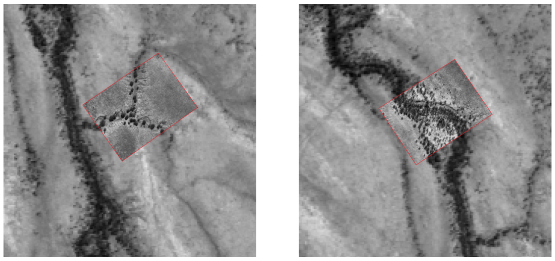

Fig. 7. Normalised cross correlation of luminance is used to register a UAV image tile (sharp foreground) to the lower resolution multispectral dataset (background).

The proposed fusion framework has been applied to this registered data, using the three UAV bands as supporting channels inside k_P (because P_H is three dimensional), and predicting three output infrared bands independently. This has resulted in infrared scene images with a 6 fold magnification factor. The full region is shown in Fig. 8, and high-detail zooms are given in Figures 9 to 11. The results presented here are open loop - this is a real investigation for which no ground truth exists. Inspecting the fused images in each scenario, we observe that the GP has noticeably sharpened the resolution of the estimated infrared channels, performing especially well in unmixing the tree crowns from the background soil due to their contrast in the supporting visible bands. Particularly noticeable in Fig. 11, different types of trees in the image data exhibit different crown colours in the false colour reconstruction due to the local spectral relationship model. Therefore the six fold improvement in resolution has made a critical difference to the scientific value of this data, as it is possible to distinguish tree crowns and other features in the data that were not resolved in the observed infrared bands.

As was discussed in the previous section, our algorithm uses the observed pixel values of the low-resolution image to construct its high-resolution reconstruction. In most situations this is a strong advantage of the model, because it means the output is not sensitive to distortions or contrast inversions in the high-resolution reference image. However, this also means that the method does not fabricate texture without evidence in the low-resolution observations. The large magnification ratio, combined with small registration errors and lighting differences between the two modalities ensure that although the model has performed excellently at recovering the infrared-colour of objects in the scene, it has had little opportunity to recover the original shading. Fortunately, it is very easy to combine a correctly coloured image from our GP fusion with a hallucinated shading model from the UAV images using an IHS decomposition (intensity-hue-saturation)[23]. We therefore present a second set of results that transforms the GP prediction into (Y, Cb, Cr) space and replaces the Y channel with the corresponding Y channel from the RGB image. This hallucination of plausible texture was found to improve the scene understanding for human perception and produce visually appealing images.

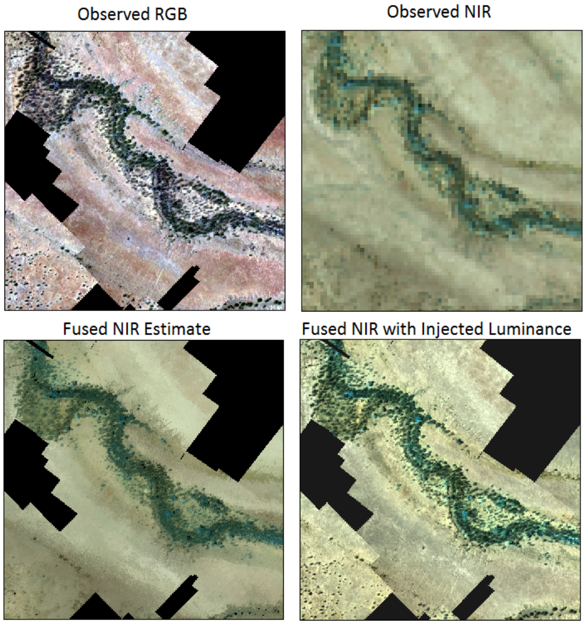

Fig. 8. Enhanced reconstruction of infrared bands. Top left: Supporting RGB. Top right: observed coarse infrared bands. Bottom left: Fusion output, Bottom right: fusion output with shading heuristic. The GP fusion provides a faithful estimate of the enhanced NIR image. Because of noise, time elapsed, and registration error, the model cannot extract colour. A shading model can improve the perceived realism by directly transferring luminance (if known).

VI. Conclusion

A new approach has been presented for fusing images that have different resolutions and different sets of spectral channels. The novelty of this approach lies in the construction of a new positive definite covariance function that can model large, non-stationary image data. Firstly, it uses spatial input dimensions augmented with image-intensity values to transfer contextual non-stationarity from the high-resolution reference image. Secondly, the new covariance function is defined over pixels of different areas, as compared to a standard covariance function that is defined over point inputs only. Finally, the covariance model is sparse by design, ensuring tractability on real images. The covariance function is used in a GP regression framework to estimate a high-resolution image of the low-resolution observations.

The proposed framework has been validated through controlled image degradation and restoration, exhibiting marked increases in image quality. In comparison with a selection of alternative techniques, the new framework was able to outperform the benchmarks in terms of quantitative image quality metrics. The GP fusion model has also been applied open loop to a multi-platform dataset: a low flying UAV was used to acquire high-resolution colour imagery, while a high altitude manned aircraft provided a comparatively coarse resolution multi-band dataset. While no quantitative ground truth exists for this scenario, the new algorithm has produced highly detailed false colour infrared images at 6 fold magnification, in this case adding significant scientific value to the channels by allowing tree crowns to be distinguished in the infrared spectrum. We have identified that future development of the model should focus on increasing the amount of texture inferred from the high-resolution image in the predictive output as the

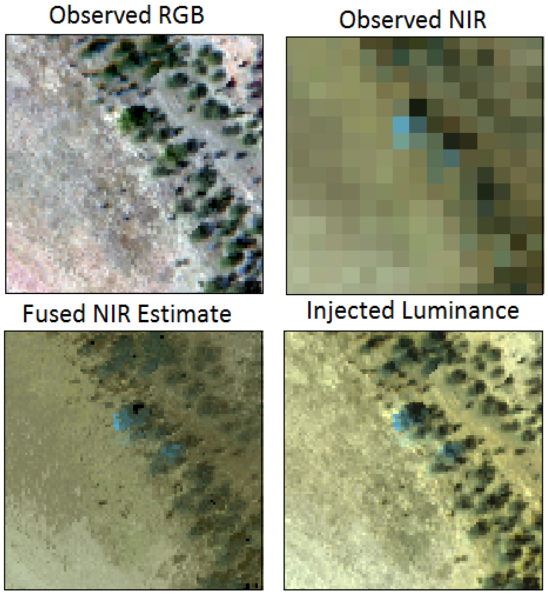

Fig. 9. High-detail zoom of the predicted infrared bands, using the layout of Fig. 8. In this case the bright blue pixels in the false colour image have been reconstructed into blue tree crowns corresponding to the corresponding tree crowns in the RGB image.

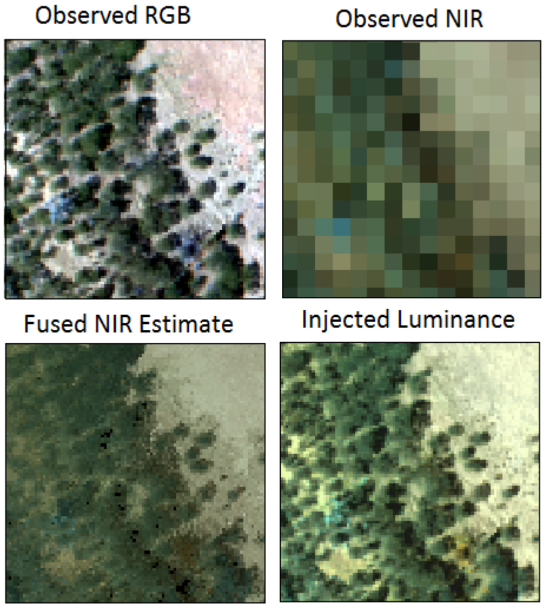

Fig. 10. High-detail zoom of the predicted infrared bands, using the layout of Fig. 8. This region presents a difficult problem as the many small tree crowns lead to very mixed pixels. The fusion estimate forces a consistent colour for tree crowns and for background, preventing colour bleeding in the fusion estimate.

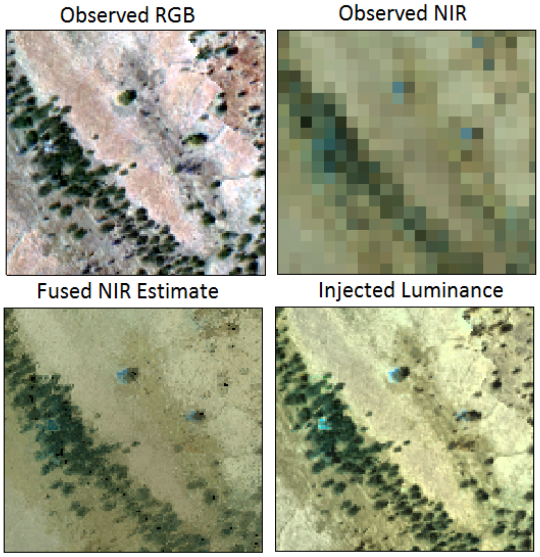

Fig. 11. High-detail zoom of the predicted infrared bands showing tree crown detail that is clearly not visible in the observed infrared bands. Layout is the same as Fig. 8. Interestingly there are two tree types present that look the same in the RGB imagery but have different appearances in the NIR bands.

magnification factor becomes large. The promising value of our results suggests that this fusion strategy has great potential in other multi-platform surveys.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work was originally produced at the Australian Centre for Field Robotics at the University of Sydney, where the UAV experiments were part of an ecological mapping project supported by Meat and Livestock Australia.

References

[1] M. Alvarez and N. Lawrence. Sparse convolved Gaussian processes for multi-output regression. *Advances in Neural Information Processing Systems*, 21:57–64, 2009.

[2] P. M. Atkinson, E. Pardo-Iguzquiza, and M. Chica-Olmo. Downscaling cokriging for super-resolution mapping of continua in remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(2):573–580, 2008.

[3] A. Barrientos, J. Colorado, J. Cerro, A. Martinez, C. Rossi, D. Sanz, and J. Valente. Aerial remote sensing in agriculture: A practical approach to area coverage and path planning for fleets of mini aerial robots. *Journal of Field Robotics*, 28(5): 667–689, 2011.

[4] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, 2008.

[5] K. Briechle and U. Hanebeck. Template matching using fast normalized cross correlation. In *SPIE*, volume 4387, pages 95–102, 2001.

[6] M. Bryson, A. Reid, C. Hung, F. Ramos, and S. Sukkarieh. Cost-effective mapping using unmanned aerial vehicles in ecology monitoring applications. In *International Symposium on Experimental Robotics (ISER)*, 2010.

[7] C. M. Christoudias, R. Urtasun, M. Salzmann, and T. Darrell. Learning to recognize objects from unseen modalities. In *European Conference on Computer Vision*, 2010.

[8] S. Farsiu, M. Elad, and P. Milanfar. Multiframe demosaicing and super-resolution of color images. *IEEE Transactions on Image Processing*, 15(1):141–159, 2006.

[9] F. D. Fave, A. Rogers, Z. Xu, S. Sukkarieh, and N. R. Jennings. Deploying the max-sum algorithm for decentralised coordination and task allocation of unmanned aerial vehicles for live aerial imagery collection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 469–476, 2012.

[10] H. He and W.-C. Siu. Single image super-resolution using Gaussian process regression. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 449–456, 2011.

[11] S. Huang and J. Tan. Adaptive sampling using mobile sensor networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 634–641, 2012.

[12] H.-C. Kim and Z. Ghahramani. Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intell.*, 28(12):1948–1959, 2006.

[13] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792, 2011.

[14] Z. Lin and H.-Y. Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (1):83–97, 2004.

[15] Y. Liu and V. Caselles. Exemplar-based image inpainting using multiscale graph cuts. *IEEE Transactions on Image Processing*, 22(5):1699–1711, 2013.

[16] A. Melkumyan and F. Ramos. Multi-kernel Gaussian processes. In *Neural Information Processing Systems Workshops "Understanding Multiple Kernel Learning Methods"*, 2009.

[17] X. Otazu and M. Gonzalez-Ausicana. Introduction of sensor spectral response into image fusion methods: Application to wavelet-based methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (10), 2005.

[18] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Thomas Dietterich editor.

[19] S. Ravishankar and Y. Bresler. Learning doubly sparse transforms for image representation. In *IEEE International Conference on Image Processing (ICIP)*, pages 685–688, 2012.

[20] T. Ruzic, A. Pizurica, and W. Philips. Markov random field based image inpainting with context-aware label selection. In *IEEE International Conference on Image Processing (ICIP)*, pages 1733–1736, 2012.

[21] J. Sun, J. Zhu, and M. F. Tappen. Context-constrained hallucination for image super-resolution. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 231–238, 2010.

[22] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot. Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5): 1301–1312, 2008.

[23] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang. A new look at IHS-like image fusion methods. *Information Fusion*, 2:177–186, 2001.

[24] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte. Non-stationary dependent Gaussian processes for data fusion in large-scale terrain modeling. In *IEEE Int Robotics and Automation (ICRA) Conf*, pages 1875–1882, 2011.

[25] Z. Wang and A. Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81 – 84, 2002.

[26] S. Zheng, W. Shi, J. Liu, and J. Tian. Remote sensing image fusion using multiscale mapped LS-SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1313–1322, 2008.