A Variational Approach for Robust Online Fusion of Multiresolution Multispectral Images

Haoqing Li[1], Ricardo Borsoi[2], Tales Imbiriba[3,4], Pau Closas[3]

[1]Dept. of Geomatics Engineering, *University of Calgary* Calgary, Alberta (Canada)
[2]CNRS, CRAN, *University of Lorraine*, Nancy (France)
[3]Dept. of Electrical & Computer Engineering, *Northeastern University*, Boston, MA (USA)
[4]Institute for experimental AI, *Northeastern University*, Boston, MA (USA)
Email: haoqing.li@ucalgary.ca, ricardo-augusto.borsoi@cnrs.fr, talesim@northeastern.edu, pau.closas@northeastern.edu

***Abstract*—Multi-resolution image fusion is a key problem for real-time satellite imaging, which has a central role in detecting and monitoring the intensity of key natural phenomena such as floods. It aims to solve the trade-off between high temporal and high spatial resolution in remote sensing instruments. Although several algorithms have been proposed to solve this problem, the presence of outliers caused by, e.g., cloud cover downgrades their performance. In this paper, an online image fusion method based on a robust Kalman filter with a weakly supervised approach for temporal dynamics estimation is proposed. Outliers are modelled as a discrete variables, where the probability of contamination for each pixel and spectral band is modelled by a latent variable whose distribution is approximated under variational inference. Experiments fusing images from the MODIS and Landsat 8 sensors show that the proposed approach is significantly more robust against cloud cover, without losing its efficiency when no clouds are present.**

***Index Terms*—Multispectral imaging, Image fusion, Bayesian Filtering, Super-resolution, Variational Inference.**

I. Introduction

Satellite-based remote sensing of the environment is an essential tool for many applications such as monitoring land-cover [1], deforestation [2] and water quality [3]. A major concern when leveraging satellite-based imaging regards the trade-off among temporal, spatial, and spectral resolutions. Such trade-off is due to the large distances from space-borne instruments and target scenes, and limitations of multiband imaging systems. In practice, these limitations imply that higher spatial resolution leads to longer revisit times. For instance, the Moderate Resolution Imaging Spectroradiometer (MODIS) has pixels sizes of 250/500/1000 m (depending on the band) with daily revisiting period while Landsat 8 captures images with pixel sizes of 30/100 m with revisiting period of approximately 16 days [4]. This is a major issue when monitoring events that are rapidly changing and require high spatial resolution to be properly characterized.

To cope with these limitations image fusion approaches were proposed to generate high-spatial-temporal resolution images with emphasis on fusing image data from multiple space-borne instruments [5] generating daily high (e.g., 30 m pixels) resolution estimated images, impacting the study of drought-induced tree mortality [6], and daily construction of snow cover maps [7]. Spatial-temporal image fusion approaches can be roughly divided into four main categories, i.e., unmixing-based [8], weighted fusion [9], learning-based [10] and Bayesian approaches [11]. Recently, a recursive image fusion approach was proposed using a Bayesian filtering framework where the process noise covariance was estimated following a weakly supervised approach [12; 13].

Despite the efficiency of these methods, outliers caused by cloud contamination can severely impact their image fusion performance. Thus, the detection and removal of pixels contaminated by clouds constitute an essential step of existing image fusion pipelines. Different algorithms for cloud (and cloud shadow) detection and removal have been proposed, which can be divided into three categories [14]. The first kind is to restore the cloudy/shadow contaminated pixels by assuming they share the same statistical distribution or geometric structures as the surrounding cloudless ones. Typical methods include spatial interpolation [15], and deep learning algorithms [16]. The second kind is to use auxiliary information from different sensors, such as synthetic aperture radar (SAR) [17] or MODIS images [18]. The third kind is to utilize cloudless images from the same sensor on other dates as reference images [19]. However, since the quality of cloud cover information is limited, the images used in the fusion process might still contain outliers. Thus, the development of robust image fusion methods is paramount. Moreover, although noise-robust image fusion methods have been proposed [20], there is a lack of robust online image fusion approaches.

In this paper, the cloudy pixels are treated under a Bayesian filtering paradigm similar to [13], where the covariance of the transition noise is estimated by a weakly supervised method, and the fusion process is implemented by modeling the observed images of different modalities (with different spatial and temporal resolutions) as measurements of a Bayesian filter. For cloud removal, pixels under cloud/shadow influence is regarded as outliers, whose probability is modelled by latent variables defined as outlier indicators. The latent variables are estimated based on variational inference, following [21]. The novelty of this paper is a variational Kalman filter framework for recursive fusion of images from multiple modalities with robustness to outliers incorporating a weakly supervised estimator for the dynamical image evolution model. Experiments fusing images from the MODIS and Landsat 8 sensors illustrate the superior performance of the proposed approach.

II. Dynamical Imaging Model

Definitions and notation: Let us denote the the ℓ-th band of the k-th acquired image reflectances from modality $m \in \Omega$ by $\boldsymbol{y}_{k,\ell}^m \in \mathbb{R}^{N_{m,\ell}}$, with $N_{m,\ell}$ pixels for each of the bands $\ell = 1, \dots, L_m$, and Ω denoting the set of image modalities (e.g., Landsat-8 and MODIS). We denote the corresponding high resolution latent reflectances by $\boldsymbol{S}_k \in \mathbb{R}^{N_H \times L_H}$, with N_H pixels and L_H bands, with $L_H \geqslant L_m$ and $N_H \geqslant N_{m,\ell}$, $\forall \ell, m$. Subindex $k \in \mathbb{N}_*$ denotes the acquisition time index. We also denote by $\mathrm{vec}(\cdot)$, $\mathrm{col}\{\cdot\}$, $\mathrm{diag}\{\cdot\}$ and by $\mathrm{blkdiag}\{\cdot\}$ the vectorization, vector stacking, diagonal and block diagonal matrix operators, respectively. The notation $\boldsymbol{x}_{a:b}$ for $a, b \in \mathbb{N}_*$ represents the set $\{\boldsymbol{x}_a, \boldsymbol{x}_{a+1}, \dots, \boldsymbol{x}_b\}$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Measurement model: The images acquired at each time k consist of spatially degraded, noisy versions of a high-resolution image $\boldsymbol{S}_k$. Following this assumption, the measurement model can be expressed according to:

$$\boldsymbol{y}_{k,\ell}^m = \mathcal{H}_\ell^m(\boldsymbol{S}_k)\boldsymbol{c}_\ell^m + \boldsymbol{r}_{k,\ell}^m\,, \qquad \ell = 1, \dots, L_m\,, \tag{1}$$

for each modality $m \in \Omega$, where $\boldsymbol{c}_\ell^m \in \mathbb{R}^{L_H}$ denotes a spectral transformation vector, mapping all bands in $\boldsymbol{S}_k$ to the ℓ-th measured band at modality m; $\mathcal{H}_\ell^m$ is a linear operator representing the band-wise spatial degradation, modeling blurring and downsampling effects of each high resolution band, and $\boldsymbol{r}_{k,\ell}^m$ represents the measurement noise. Note that, while we consider the spatial resolution of the high resolution bands in $\boldsymbol{S}_k$ to be the same, different bands from the same modality can have different resolutions. Most works assume the measurement noise to be Gaussian and uncorrelated among bands, that is, $\boldsymbol{r}_{k,\ell}^m \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_\ell^m)$ with time-invariant covariance matrix given by $\boldsymbol{R}_\ell^m \in \mathbb{R}^{N_{m,\ell} \times N_{m,\ell}}$. At each time index k the scene is measured through one of the imaging modalities $m \in \Omega$.

Using (1) and properties of the vectorization of matrix products, we can stack all bands of the m-th modality in the vector $\widetilde{\boldsymbol{y}}_k^m \in \mathbb{R}^{n_y^m}$, with $n_y^m = \widetilde{N}_m L_m$, leading to the equivalent reformulation of model (1) as

$$\widetilde{\boldsymbol{y}}_k^m = \widetilde{\boldsymbol{H}}_k^m \boldsymbol{s}_k + \widetilde{\boldsymbol{r}}_k^m\,, \tag{2}$$

where $\widetilde{\boldsymbol{H}}_k^m = \left[\left((\boldsymbol{c}_1^m)^\top \boldsymbol{H}_1^m\right)^\top, \cdots, \left((\boldsymbol{c}_{L_m}^m)^\top \boldsymbol{H}_{L_m}^m\right)^\top\right]^\top$ and

$\widetilde{\boldsymbol{y}}_k^m = \mathrm{col}\left\{\boldsymbol{y}_{k,1}^m, \dots, \boldsymbol{y}_{k,L_m}^m\right\}, \widetilde{\boldsymbol{r}}_k^m = \mathrm{col}\left\{\boldsymbol{r}_{k,1}^m, \dots, \boldsymbol{r}_{k,L_m}^m\right\},$
$\widetilde{\boldsymbol{o}}_k^m = \mathrm{col}\left\{\boldsymbol{o}_{k,1}^m, \dots, \boldsymbol{o}_{k,L_m}^m\right\}, \widetilde{\boldsymbol{R}}_k^m = \mathrm{blkdiag}\left\{\boldsymbol{R}_1^m, \dots, \boldsymbol{R}_{L_m}^m\right\},$

and $\boldsymbol{H}_\ell^m$ is a matrix form representation of the operator $\mathcal{H}_\ell^m$, such that $\mathrm{vec}(\mathcal{H}_\ell^m(\boldsymbol{S}_k)) = \boldsymbol{H}_\ell^m \boldsymbol{s}_k$. $\boldsymbol{s}_k \in \mathbb{R}^{L_H N_H}$ denotes a vector-ordering of the high-resolution image $\boldsymbol{S}_k$ obtained by grouping all pixels such that the bands of a single HR pixel, and positions corresponding to nearby pixels are adjacent to each other (see [13] for details about how the ordering is done).

Note that satellite images may be corrupted by several effects, including dead pixels in the sensor, incorrect atmospheric compensation, and the presence of heavy cloud cover. Such pixels cannot be reliably used in the image fusion process as they may degrade the performance of the method. Most existing algorithms ignore the existence of such outlier pixels, which can lead to a considerable loss of performance when they are applied in real settings. Thus, we address this issue by considering two hypotheses for the measurements. Under the first hypothesis, denoted by $\mathcal{C}_0$, the pixels are only affected by Gaussian noise $\widetilde{\boldsymbol{r}}_k^m$, whereas under the second hypothesis, denoted by $\mathcal{C}_1$, the pixels are corrupted, being affected by a vector of outliers $\widetilde{\boldsymbol{o}}_k^m \in \mathbb{R}^{n_y^m}$. This leads to the following measurement model:

$$\widetilde{y}_k^{m,(i)} = \begin{cases} \widetilde{\boldsymbol{h}}_k^{m,(i)} \boldsymbol{s}_k + \widetilde{r}_k^{m,(i)}\,, & \text{under } \mathcal{C}_0 \\ \widetilde{\boldsymbol{h}}_k^{m,(i)} \boldsymbol{s}_k + \widetilde{r}_k^{m,(i)} + \widetilde{o}_k^{m,(i)}\,, & \text{under } \mathcal{C}_1 \end{cases} \tag{3}$$

for $i = 1, \dots, n_y^m$, where $\widetilde{y}_k^{m,(i)}$, $\widetilde{r}_k^{m,(i)}$ and $\widetilde{o}_k^{m,(i)}$ denote the i-th element of $\widetilde{\boldsymbol{y}}_k^m$, $\widetilde{\boldsymbol{r}}_k^m$ and $\widetilde{\boldsymbol{o}}_k^m$ respectively, and $\widetilde{\boldsymbol{h}}_k^{m,(i)}$ denotes the i-th row of $\widetilde{\boldsymbol{H}}_k^m$. Note that we have one hypothesis for each band and pixel in the measurement of modality m, which might be affected by an outlier. Moreover, the approach we will consider will not need a rigid statistical model for $\widetilde{\boldsymbol{o}}_k^m$, as will be shown in the following section.

Dynamical evolution model: To exploit the temporal information in the image sequence, an adequate dynamical model is necessary to describe the evolution of the HR images. We consider the following model proposed in [13]:

$$\boldsymbol{s}_{k+1} = \boldsymbol{F}_k \boldsymbol{s}_k + \boldsymbol{q}_k\,, \tag{4}$$

where $\boldsymbol{F}_k = \boldsymbol{I} \in \mathbb{R}^{n_s}$ is the state transition matrix ($n_s = L_H N_H \times L_H N_H$), and $\boldsymbol{q}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_k)$ with $\boldsymbol{Q}_k \in \mathbb{R}^{n_s}$ being the state process noise covariance matrix. Selecting an adequate $\boldsymbol{Q}_k$ is paramount to the performance of the method: at each instant k, the process noise $\boldsymbol{q}_k$ should have high variance on pixels that are expected to undergo larger changes, and low variance otherwise. In this work we follow the approach proposed in [13], in which a diagonal approximation of $\boldsymbol{Q}_k$ is computed based on historical high-resolution image data of the same geographical region.

III. A Robust Image Fusion Method

Given the probabilistic model described in the previous section for both the image generation and its temporal dynamics, the online image fusion consists of computing the posterior distribution of the high-resolution image conditioned on all past measurements, $p\big(\boldsymbol{s}_k | \{\widetilde{\boldsymbol{y}}_{1:k}^m\}_{m \in \Omega}\big)$. When the data follows a linear and Gaussian measurement model, this PDF can be computed efficiently using the Kalman filter [13; 22]. Nevertheless, such techniques do not account for the presence of disruptive outliers such as clouds or shadows.

To address this issue, we consider an approach based on the general VBKF (GVBKF) proposed in [21], which is summarized in this section. First, let us introduce the outlier indicator vector $\boldsymbol{z}_k^m = \left(z_k^{m,(1)}, \dots, z_k^{m,(n_y^m)}\right)^\top \in \mathcal{Z} = \{0,1\}^{n_y^m}$, such that $z_k^{m,(i)} = 0$ if there is an outlier on the i-th (corrupted) element of $\widetilde{\boldsymbol{y}}_k^m$, i.e., $\widetilde{y}_k^{m,(i)}$, and $z_k^{m,(i)} = 1$ if the i-th element is otherwise clean (not corrupted). The clean elements of $\widetilde{\boldsymbol{y}}_k^m$ can be used nominally in the image fusion process, whereas the

contribution of the corrupted ones should be down-weighted. This is performed by modifying the observation model such that the i-th position of indicator vector, $z_k^{m,(i)}$, adjusts the variance of a modified (referred to as *improper*) Gaussian noise distribution, leading to

$$p\left(\widetilde{\boldsymbol{y}}_k^m|\boldsymbol{s}_k,\boldsymbol{z}_k^m\right)=\frac{1}{c(\boldsymbol{z}_k^m)}\exp\Big(-\frac{1}{2}\|\widetilde{\boldsymbol{y}}_k^m-\widetilde{\boldsymbol{H}}_k^m\boldsymbol{s}_k\|^2_{\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{z}_k^m)}\Big), \tag{5}$$

where the covariance matrix $\boldsymbol{\Sigma}_k(\boldsymbol{z}_k^m)$ is given by

$$\boldsymbol{\Sigma}_k(\boldsymbol{z}_k^m)=\begin{bmatrix}\sigma^2_{1,1}/z_k^{m,(1)} & \sigma^2_{1,2} & \cdots & \sigma^2_{1,n_y^m}\\ \sigma^2_{2,1} & \sigma^2_{2,2}/z_k^{m,(2)} & \cdots & \sigma^2_{2,n_y^m}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma^2_{n_y^m,1} & \sigma^2_{n_y^m,2} & \cdots & \sigma^2_{n_y^m,n_y^m}/z_k^{m,(n_y^m)}\end{bmatrix},$$

where $\sigma^2_{i,j}\triangleq[\widetilde{\boldsymbol{R}}_k^m]_{i,j}$ (the dependence on k and m has been omitted from $\sigma^2_{i,j}$ for notational simplicity). This corresponds to the original matrix $\widetilde{\boldsymbol{R}}_k^m$ with indicators $\{z_k^{m,(i)}\}_{i=1}^{n_y^m}$ dividing its diagonal terms, such that $\widetilde{\boldsymbol{R}}_k^m=\boldsymbol{\Sigma}_k(\mathbf{1})$ (i.e., when there are no outliers). The normalization constant in (5) can be computed as $c(\boldsymbol{z}_k^m)=\sqrt{(2\pi)^{n_y^{m\prime}}|\mathbf{C}(\boldsymbol{z}_k^m)|}$. Note that it depends on the indicator vector, through matrix $\mathbf{C}(\boldsymbol{z}_k^m)$, which is a transformation of $\boldsymbol{\Sigma}_k(\boldsymbol{z}_k^m)$ where the rows/columns corresponding to those $z_k^{m,(i)}=0$ are removed. Special cases are $i)$ $\mathbf{C}(\mathbf{1})=\boldsymbol{\Sigma}_k(\mathbf{1})=\widetilde{\boldsymbol{R}}_k^m$ (resulting in the original model without indicators), and $ii)$ $\mathbf{C}(\mathbf{0})$, corresponding to the absence of measurements and defined as $\mathbf{C}(\mathbf{0})=\mathbf{I}$.

Therefore, the dimension $n_y^{m\prime}$ of the variables that are Gaussian distributed in $\widetilde{\boldsymbol{y}}_k^m$ is effectively reduced by the amount of zero indicators: $n_y^{m\prime}=\sum_{i=1}^{n_y^m}z_k^{m(i)}\leqslant n_y^m$, with equality when all indicators are one. To solve the image fusion problem, we need to approximate the posterior distribution $p\left(\boldsymbol{s}_k,\boldsymbol{z}_k^m|\widetilde{\boldsymbol{y}}_k^m\right)$ of both the HR image and of the outlier indicator vector. Following a Bayesian framework, we impose a beta-Bernoulli hierarchical prior to each indicator element,

$$p\left(z_k^{m,(i)}|\pi_k^{m,(i)}\right)=\left(\pi_k^{m,(i)}\right)^{z_k^{m,(i)}}\left(1-\pi_k^{m,(i)}\right)^{1-z_k^{m,(i)}}, \tag{6}$$

where $\pi_k^{m,(i)}$ is a beta distributed random variable parameterized by (unknown shape hyper-parameters) $e_0^{(i)}$ and $f_0^{(i)}$,

$$p\left(\pi_k^{m,(i)}\right)=\frac{\left(\pi_k^{m,(i)}\right)^{e_0^{(i)}-1}\left(1-\pi_k^{m,(i)}\right)^{f_0^{(i)}-1}}{\beta\left(e_0^{(i)},f_0^{(i)}\right)}, \tag{7}$$

and $\beta(\cdot,\cdot)$ is the beta function. Notice that we are assuming that the indicators are mutually independent

$$p(\boldsymbol{z}_k^m,\boldsymbol{\pi}_k^m)=\prod_{i=1}^{n_y^m}p\left(z_k^{m,(i)}\Big|\pi_k^{m,(i)}\right)p\left(\pi_k^{m,(i)}\right), \tag{8}$$

as well as independent from the observations since the underlying statistics modeling the outliers do not depend on the actual values of the data.

According to the Variational Inference (VI) principle [23], to estimate the posterior distribution of the latent variables $\boldsymbol{\theta}=\{\boldsymbol{s}_k,\boldsymbol{\pi}_k^m,\boldsymbol{z}_k^m\}$, that is $p(\boldsymbol{\theta}|\widetilde{\boldsymbol{y}}_{1:k}^m)$, we can use an auxiliary distribution $q(\boldsymbol{\theta})$ and independence assumptions such that:

$$q\left(\boldsymbol{\theta}\right)=q\left(\boldsymbol{s}_k\right)q\left(\boldsymbol{\pi}_k^m\right)q\left(\boldsymbol{z}_k^m\right)=q\left(\boldsymbol{s}_k\right)\prod_{i=1}^{n_y^m}q\left(\pi_k^{m,(i)}\right)q\left(z_k^{m,(i)}\right). \tag{9}$$

$\widetilde{\boldsymbol{y}}_k^m$ is conditionally independent on $\boldsymbol{\pi}_k^m$; $\boldsymbol{s}_k$ is conditionally independent on $\boldsymbol{z}_k^m$ and $\boldsymbol{\pi}_k^m$; $\widetilde{\boldsymbol{y}}_{1:k-1}^m$ is conditionally independent on $\boldsymbol{z}_k^m$, $\boldsymbol{\pi}_k^m$ and $\widetilde{\boldsymbol{y}}_k^m$. Thus, the various marginal distributions, $q(\cdot)$, are then obtained from the mean-field VI method, which attempts to compute $q\left(\boldsymbol{\theta}\right)$ which closely approximates the posterior under the true joint distribution:

$$p\left(\boldsymbol{s}_k,\boldsymbol{\pi}_k^m,\boldsymbol{z}_k^m,\widetilde{\boldsymbol{y}}_{1:k}^m\right)\propto p\left(\boldsymbol{s}_k|\widetilde{\boldsymbol{y}}_{1:k-1}^m\right)p\left(\widetilde{\boldsymbol{y}}_k^m|\boldsymbol{s}_k,\boldsymbol{z}_k^m\right)p(\boldsymbol{z}_k^m,\boldsymbol{\pi}_k^m), \tag{10}$$

Within the Gaussian filtering framework, the first term on the right-hand side of (10) is a predictive density, which can be approximated as $p\left(\boldsymbol{s}_k|\widetilde{\boldsymbol{y}}_{1:k-1}^m\right)\approx\mathcal{N}\left(\hat{\boldsymbol{s}}_{k|k-1},\boldsymbol{P}_{k|k-1}\right)$, where the corresponding mean and covariance are [22]

$$\hat{\boldsymbol{s}}_{k|k-1}=\int\left(\boldsymbol{F}\boldsymbol{s}_{k-1}\right)p\left(\boldsymbol{s}_{k-1}|\widetilde{\boldsymbol{y}}_{1:k-1}^m\right)d\boldsymbol{s}_{k-1}, \tag{11}$$

$$\boldsymbol{P}_{k|k-1}=\int\left(\boldsymbol{F}\boldsymbol{s}_{k-1}-\hat{\boldsymbol{s}}_{k|k-1}\right)\left(\boldsymbol{F}\boldsymbol{s}_{k-1}-\hat{\boldsymbol{s}}_{k|k-1}\right)^\top\times p\left(\boldsymbol{s}_{k-1}|\widetilde{\boldsymbol{y}}_{1:k-1}^m\right)d\boldsymbol{s}_{k-1}+\boldsymbol{Q}_k, \tag{12}$$

with $\hat{\boldsymbol{s}}_{k-1|k-1}$ and $\boldsymbol{P}_{k-1|k-1}$ the mean and covariance of the filtering posterior at $k-1$, that is $p\left(\boldsymbol{s}_{k-1}|\widetilde{\boldsymbol{y}}_{1:k-1}^m\right)\approx\mathcal{N}\left(\hat{\boldsymbol{s}}_{k-1|k-1},\boldsymbol{P}_{k-1|k-1}\right)$. The auxiliary distributions in (9), $q(\boldsymbol{s}_k)$, $q(\boldsymbol{\pi}_k^m)$ and $q(\boldsymbol{z}_k^m)$, are computed by updating them sequentially and iteratively under Bayesian filtering scheme, defined as GVBKF in [21]. For details of the algorithm, please refer to [21] for its full derivation.

IV. Experiments and Results

In this section, we use the proposed methodology to fuse Landsat and MODIS images over time. The proposed GVBKF is compared with the Kalman filter with the weakly supervised process noise covariance estimation proposed in [13] with a block diagonal state covariance matrix, which we refer to simply as KF. This method is selected because it achieved the best performance among the online methods compared in [13].

Remotely Sensed data: We collected data from Oroville Dam, which is introduced in [13] and shown in Figure 1. Specifically, we collected the dataset separately for cloudless and cloudy cases. In terms of the cloudless case, we collected MODIS and Landsat data acquired on an interval ranging from 2018/07/03 to 2018/09/21. In terms of the cloudy case, we collected MODIS and Landsat data on a interval ranging from 2018/03/29 to 2018/07/19. This interval was selected because a typical cloud cover occurs at 2018/05/16 in the Landsat data. In this experiment, we will focus on the red and near-infrared (NIR) bands since they are often used to distinguish water from other landcover elements in the image. We also collected 50 Landsat data from 2013/04/09 to 2017/12/07 to

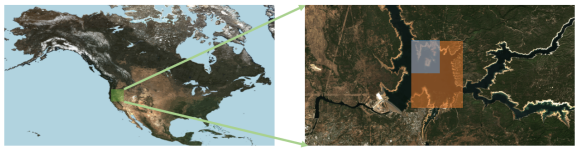

Fig. 1: Oroville dam. Blue and orange boxes delimit the study area used in the cloudy and cloudless datasets, respectively.

serve as a past historical dataset $\mathcal{D}_k$ and for the purpose of weakly supervised covariance estimation approach of [13].

The study region corresponds to Landsat and MODIS images with 81×81 and 9×9 pixels for cloudy case, and 162×162 and 18×18 pixels for cloudless case respectively After filtering for heavy cloud cover during the designated time periods, a set of 6 Landsat and 15 MODIS images were obtained for cloudless case. We used the first Landsat images for initialization leading to 5 and 15 images used in the remaining fusion process. In terms of the cloudy case, 6 Landsat and 10 MODIS images were obtained, among which the first Landsat is used for initialization leading to 5 Landsat and 10 MODIS as the observation.

In the cloudless case, from the set of 5 Landsat images of the Oroville Dam site that were available for testing, three of them were set aside and not processed by any of the algorithms. These images were acquired at dates 07/19, 08/20 and 09/05, when MODIS observations were also available, and will be used in the form of a reference for the evaluation of the algorithms' capability of estimating the high resolution images at these dates solely from the low resolution MODIS measurements. As for the cloudy case, 2 out of 5 Landsat images were set aside and not processed by all algorithms. These images were acquired at dates 06/01, 07/03, when MODIS observations were also available.

Algorithm setup: We initialized the proposed Kalman filter using a high resolution Landsat observation as the state, i.e., $\boldsymbol{s}_{0|0} = \widetilde{\boldsymbol{y}}_0^{\mathsf{L}}$, and set all the parameters according to the statistic of observations as well as the experience. $\boldsymbol{P}_{0|0} = 10^{-10}\boldsymbol{P}_0$. where $\boldsymbol{P}_0 = \frac{1}{10}\mathbb{1}_B + \frac{9}{10}\boldsymbol{I}$, with $\mathbb{1}_B$ being a block diagonal matrix of ones. The noise covariance matrices were set as $\boldsymbol{R}_\ell^{\mathsf{L}} = 3 \times 10^{-2}\boldsymbol{R}_0$ and $\boldsymbol{R}_\ell^{\mathsf{M}} = 10^{-4}\boldsymbol{R}_0$ for all ℓ, where $\boldsymbol{R}_0$ is a block diagonal matrix with the block as $\begin{bmatrix} 1 & 0.1 \\ 0.1 & 2 \end{bmatrix}$, which is selected based on the variance of the noise in each band. The blurring and downsampling matrices were set as $\boldsymbol{H}_\ell^{\mathsf{L}} = \boldsymbol{I}$ for Landsat, while for MODIS $\boldsymbol{H}_\ell^{\mathsf{M}}$ consisted of a convolution by an uniform 9×9 filter, defined by $\boldsymbol{h} = \frac{1}{81}\mathbb{1}_{9\times 9}$ (where $\mathbb{1}_{9\times 9}$ is a 9×9 matrix of ones), followed by decimation by a factor of 9, which represents the degradation occurring at the sensor. Vectors $\boldsymbol{c}_\ell^m$ contained a positive gain in the ℓ-th position to compensate for scaling differences between Landsat and MODIS sensors, and zeros elsewhere.

To reduce the complexity of the proposed method, we consider the assumption of a block-diagonal state covariance matrix with one block per Landsat multispectral pixel, similarly to [13]. We also set $e_0^{(i)} = 0.5$ and $f_0^{(i)} = 0.5$. In addition, note that the VI process in GVBKF algorithm is an iterative

TABLE I: Misclassification Percentage without cloud cover.

	07/19	08/20	09/05	09/21	Average
KF	6.1119	8.6496	10.1242	10.4100	8.8239
GVBKF	3.7189	8.0590	9.5946	9.8384	7.8027

TABLE II: Misclassification Percentage with cloud cover.

	06/01	07/03	07/19	Average
KF	7.0873	7.9713	11.1264	8.7283
GVBKF	7.1635	7.4531	9.5565	8.0577

TABLE III: Mean Square Error without cloud cover.

	07/19	08/20	09/05	09/21	Average
KF	0.0028	0.0053	0.0049	0.0056	0.0047
GVBKF	0.0019	0.0046	0.0045	0.0050	0.0040

TABLE IV: Mean Square Error with cloud cover.

	06/01	07/03	07/19	Average
KF	0.0074	0.0049	0.0073	0.0065
GVBKF	0.0072	0.0070	0.0092	0.0078

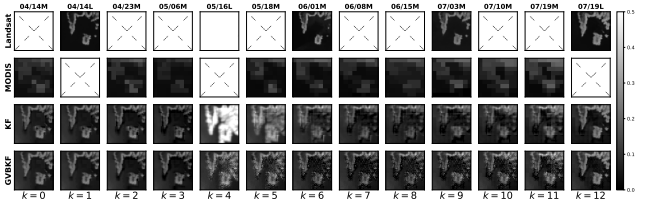

Fig. 2: Reconstruction Landsat image in red band in cloudy case (a cloud was present in 05/16). Acquisition dates are displayed in the top labels at each column with a character, M for MODIS and L for Landsat, indicating the image used in the fusion algorithms.

procedure, which was run until the relative difference between state estimation in different iterations was less then 10%, or up to a maximum of 20 iterations.

Results and discussion: In Figure 2, we show the fused reflectances as well as the acquired reflectance values from MODIS and Landsat for the red band. Quantitative results in the form of the misclassification percentage of water mapping results and mean squared error (MSE) of reconstruction results, with/without cloud cover, are shown in Tables I, II, III and IV. We recall that only the first and last Landsat images as well as the one at 05/16 were used in the fusion process, keeping the remaining two images as ground-truth for evaluation purposes. Note that the Landsat image at 05/16 is covered by clouds, leading the image in Figure 2 appearing as totally white. Analyzing the results we can see that the reconstruction results of KF, especially the ones around cloudy Landsat observation are heavily affected by the presence of the cloud, while the GVBKF holds a stable performance which is more robust against cloud cover. Figure 3 shows a classification water mapping results, where the performance of the KF and GVBKF are more similar. As shown in the quantitative results, the GVBKF holds a better performance than the KF both in the cloudless and cloudy cases in terms of lower misclassification percentages. This illustrates that the proposed method grants robustness to large outliers without considerable loss of performance. On the other hand, the GVBKF performs a bit worse in terms of MSE in the cloudy case. This is because the GBVKF has a lower amount of pixels with large errors but a higher error-per-pixel

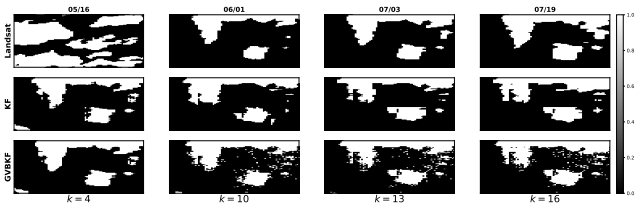

Fig. 3: Classification water mapping in cloudy case.

on average, which explains why it has a lower misclassification percentage but a higher MSE.

V. Conclusion

In this paper, we proposed an online image fusion method that is robust to outliers such as clouds and shadows. To achieve this goal we proposed a linear Gaussian imaging model, contaminated by discrete outliers, and whose time evolution is learned from a set of historical images. To estimate the high-resolution states, i.e. the high-resolution image estimates, we resorted to variational inference strategies to implement a variational Kalman filter under the proposed model. Experimental results show that the proposed algorithm is robust against cloud cover, without losing performance when no clouds are present.

Acknowledgment

This work has been partially supported by the French National Research Agency under grants ANR-23-CE23-0024, ANR-23-CE94-0001, and by the National Science Foundation, under grants NSF 2316420, ECCS-1845833 and CCF-2326559.

References

[1] M. Lu, J. Chen, H. Tang, Y. Rao, P. Yang, and W. Wu, "Land cover change detection by integrating object-based data blending model of landsat and modis," *Remote Sensing of Environment*, vol. 184, pp. 374–386, 2016.

[2] M. Schultz, J. G. Clevers, S. Carter, J. Verbesselt, V. Avitabile, H. V. Quang, and M. Herold, "Performance of vegetation indices from landsat time series in deforestation monitoring," *International journal of applied earth observation and geoinformation*, vol. 52, pp. 318–327, 2016.

[3] M. H. Gholizadeh, A. M. Melesse, and L. Reddi, "A comprehensive review on water quality parameters estimation using remote sensing techniques," *Sensors*, vol. 16, no. 8, p. 1298, 2016.

[4] D. P. Roy, M. A. Wulder, T. R. Loveland, C. E. Woodcock, R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote sensing of Environment*, vol. 145, pp. 154–172, 2014.

[5] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images," *Remote Sensing of Environment*, vol. 204, pp. 31–42, 2018.

[6] Y. Yang, M. C. Anderson, F. Gao, J. D. Wood, L. Gu, and C. Hain, "Studying drought-induced forest mortality using high spatiotemporal resolution evapotranspiration data from thermal satellite imaging," *Remote Sensing of Environment*, vol. 265, p. 112640, 2021.

[7] K. Rittger, M. Krock, W. Kleiber, E. H. Bair, M. J. Brodzik, T. R. Stephenson, B. Rajagopalan, K. J. Bormann, and T. H. Painter, "Multi-sensor fusion using random forests for daily fractional snow cover at 30 m," *Remote Sensing of Environment*, vol. 264, p. 112608, 2021.

[8] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 116–127, 2020.

[9] Y. Zhang, G. M. Foody, F. Ling, X. Li, Y. Ge, Y. Du, and P. M. Atkinson, "Spatial-temporal fraction map fusion with multi-scale remotely sensed images," *Remote Sensing of Environment*, vol. 213, pp. 162–181, 2018.

[10] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, 2018.

[11] S. Xu and J. Cheng, "A new land surface temperature fusion strategy based on cumulative distribution function matching and multiresolution Kalman filtering," *Remote Sensing of Environment*, vol. 254, p. 112256, 2021.

[12] H. Li, B. Duvvuri, R. Borsoi, T. Imbiriba, E. Beighley, D. Erdoğmuş, and P. Closas, "Online multi-resolution fusion of space-borne multispectral images," in *2022 IEEE Aerospace Conference (AERO)*. IEEE, 2022, pp. 1–12.

[13] ——, "Online fusion of multi-resolution multispectral images with weakly supervised temporal dynamics," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 471–489, 2023.

[14] Z. Wang, D. Zhou, X. Li, L. Zhu, H. Gong, and Y. Ke, "Virtual image-based cloud removal for landsat images," *GIScience & Remote Sensing*, vol. 60, no. 1, p. 2160411, 2023.

[15] A. C. Siravenha, D. Sousa, A. Bispo, and E. Pelaes, "Evaluating inpainting methods to the satellite images clouds and shadows removing," in *Proc. International Conference on Signal Processing, Image Processing and Pattern Recognition*, 2011, pp. 56–65.

[16] M. Xu, F. Deng, S. Jia, X. Jia, and A. J. Plaza, "Attention mechanism-based generative adversarial networks for cloud removal in landsat images," *Remote sensing of environment*, vol. 271, p. 112902, 2022.

[17] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[18] H. Shen, J. Wu, Q. Cheng, M. Aihemaiti, C. Zhang, and Z. Li, "A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 3, pp. 862–874, 2019.

[19] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 148–160, 2020.

[20] Z. Tan, M. Gao, J. Yuan, L. Jiang, and H. Duan, "A robust model for MODIS and Landsat image fusion considering input noise," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[21] H. Li, D. Medina, J. Vila-Valls, and P. Closas, "Robust Variational-based Kalman Filter for Outlier Rejection with Correlated Measurements," *IEEE Transactions on Signal Processing*, vol. 69, pp. 357–369, 2020.

[22] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, no. 3.

[23] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. New York: Springer-Verlag, 2005.