A hierarchical spatiotemporal adaptive fusion model using one image pair

Article in International Journal of Digital Earth · September 2016
DOI: 10.1080/17538947.2016.1235621

3 authors:

Bin Chen
University of California, Davis
28 PUBLICATIONS 146 CITATIONS
SEE PROFILE

Bo Huang
The Chinese University of Hong Kong
220 PUBLICATIONS 4,554 CITATIONS
SEE PROFILE

Bing Xu
Tsinghua University
215 PUBLICATIONS 4,018 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Virtual Environments for Geospatial Applications 2009 View project

Proceedings of the tenth international conference on Computers in Railways, Prague. WIT Transactions on Built Environment View project
A hierarchical spatiotemporal adaptive fusion model using one image pair

Bin Chen, Bo Huang & Bing Xu

To cite this article: Bin Chen, Bo Huang & Bing Xu (2016): A hierarchical spatiotemporal adaptive fusion model using one image pair, International Journal of Digital Earth, DOI: 10.1080/17538947.2016.1235621

To link to this article: http://dx.doi.org/10.1080/17538947.2016.1235621

Published online: 01 Nov 2016.
A hierarchical spatiotemporal adaptive fusion model using one image pair

Bin Chen, Bo Huang and Bing Xu

State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China; Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong; Ministry of Education Key Laboratory for Earth System Modeling, Center for Earth System Science, Tsinghua University, Beijing, China; Department of Geography, University of Utah, Salt Lake City, UT, USA

ABSTRACT

Image fusion techniques that blend multi-sensor characteristics to generate synthetic data with fine resolutions have generated great interest within the remote sensing community. Over the past decade, although many advances have been made in the spatiotemporal fusion models, there still remain several shortcomings in existing methods. In this article, a hierarchical spatiotemporal adaptive fusion model (HSTAFM) is proposed for producing daily synthetic fine-resolution fusions. The suggested model uses only one prior or posterior image pair, especially with the aim being to predict arbitrary temporal changes. The proposed model is implemented in two stages. First, the coarse-resolution image is enhanced through super-resolution based on sparse representation; second, a pre-selection of temporal change is performed. It then adopts a two-level strategy to select similar pixels, and blends multi-sensor features adaptively to generate the final synthetic data. The results of tests using both simulated and actual observed data show that the model can accurately capture both seasonal phenology change and land-cover-type change. Comparisons between HSTAFM and other developed models also demonstrate our proposed model produces consistently lower biases.

1. Introduction

Due to the technical limitations and the existing inevitable phenomenon of energy scattering in the sensors’ design, there is so far no such unified sensor that can produce remotely sensed data with fine spatial and temporal resolutions simultaneously. Facing an emerging need of remotely sensed data with fine spatial details and temporal frequency for global change detection (Zhang and Xu 2015), one possible cost-effective approach that blends multi-sensor spatial and temporal characteristics to generate synthetic data with fine resolutions has generated great interest within the remote sensing community (Gao et al. 2006; Camps-Valls et al. 2008; Huang and Song 2012; Chen and Xu 2014; Chen, Huang, and Xu 2015a, 2015b; Gao et al. 2015; Michishita et al. 2015). Several spatiotemporal fusion models have been proposed during the past decade. Based on the characteristics of the model framework and procedures of the model implementation, we further classify them into four major categories: (i) the transformation-based; (ii) reconstruction-based; (iii) unmixing-based; and (iv) learning-based models.
Transformation-based models include wavelet and tasseled cap transformations. Acerbi-Junior, Clevers, and Schaepman (2006) first proposed to increase the spatial resolution of MODIS data by integrating the Landsat imagery using a three-level coupled wavelet decomposition scheme. Hilker et al. (2009) used a tasseled cap transformation of both Landsat TM/ETM+ and MODIS reflectance data to capture temporal change information with a fine spatial resolution in a transformed space. However, the transformation-based models mainly focus on the integration of spatial and spectral information for image enhancement, instead of constructing a distinct fusion scheme between spatial and temporal information (Chen, Huang, and Xu 2015a). Thus, this kind of method is recommended to be combined with other fusion frameworks.

In the reconstruction-based models, the synthetic fusions are generated by a weighted sum of the spectrally similar neighboring information from fine spatial but coarse temporal resolution, and fine temporal but coarse spatial resolution data (Gao et al. 2015). Gao et al. (2006) proposed a spatial and temporal adaptive reflectance fusion model (STARFM) to blend Landsat and MODIS imagery to generate daily synthetic Landsat-like surface reflectance. Several improved STARFM-like models have since been developed. Hilker et al. (2009) proposed a spatial and temporal adaptive algorithm for mapping reflectance change (STAARCH) to identify highly detailed spatiotemporal patterns in land-cover change. STAARCH also produces synthetic Landsat-like images for each available date of MODIS data based on an extended STARFM. However, the prediction accuracy of the STARFM and STAARCH is sensitively correlated with landscape heterogeneity (Gao et al. 2006; Hilker et al. 2009). Addressing this issue, Zhu et al. (2010) developed an enhanced spatial and temporal adaptive reflectance fusion model (ESTARFM) considering conversion coefficients based on STARFM, so that homogeneous and heterogeneous pixels have different conversion coefficients in the prediction. A customized fusion model was developed by Michishita et al. (2015), based on the ESTARFM. Reflectance of the moderate-solution image on the target dates can be better predicted by the proposed C-ESTARFM than the original ESTARFM. Huang et al. (2013) combined the STARFM and bilateral filter to generate the land surface temperature (LST) with simultaneous spatial and temporal resolutions. Weng, Fu, and Gao (2014) modified the ESTARFM by considering the annual temporal cycle to derive accurate LST data with fine spatiotemporal resolutions in the heterogeneous urban landscape. Wu et al. (2015) proposed a spatiotemporal integrated temperature fusion model based on STARFM to retrieve fine-resolution LST data by combining multi-scale polar-orbiting and geostationary satellite observations. Given the common assumption under the STARFM-like framework that the land-cover types do not change during the prediction period, the reconstruction-based models cannot predict accurate objects when the shape or type changes and short-term disturbances are not recorded in any of the base fine-resolution images (Gao et al. 2006; Zhu et al. 2010; Gao et al. 2015).

The unmixing-based models rely on the pixel unmixing techniques, which downscale the coarse-resolution images to generate fine-resolution synthetic images while preserving the spectral richness and fidelity. The basic implementation proposed by Zhukov et al. (1999) includes four steps: (i) cluster the fine-resolution map and determine the endmembers at coarse-resolution scale; (ii) calculate the endmember fraction for each coarse-resolution pixel and obtain the proportion matrix; (iii) unmix the coarse-resolution pixel at the predicted date within a moving window convolution; and (iv) restore a fine-resolution image by assigning unmixed spectra to each class according to the clustering map. Later, Zurita-Milla, Clevers, and Schaepman (2008) introduced constraints into the linear unmixing process to reconstruct synthetic images with the spectral and temporal resolution provided by a medium resolution imaging spectrometer (MERIS), but a Landsat-like spatial resolution. Zurita-Milla et al. (2009) then applied the constrained linear unmixing model to a time series of MERIS images to produce synthetic fused images. Gevaert and García-Haro (2015) directly unmixed the temporal change from the coarse-resolution pixels to estimate the corresponding change of endmembers by introducing the Bayesian theory to restrain the unmixing process. However, the unmixing-based models require a prior unsupervised classification for input fine-resolution
images, or a high spatial resolution land cover/use database as auxiliary materials for pixel unmixing. Similar to the reconstruction-based models, they also require that no land-cover-type change occurs during the prediction period. In addition, the unmixing-based models are incapable of predicting accurate spatial and spectral variability within the intra-class. Recently, Zhu et al. (2016) proposed a flexible spatiotemporal data fusion model (FSDAF) based on spectral unmixing analysis and a thin plate spline interpolator. It claims to be able to predict both gradual phenology change and type change, which requires to be a further test and validation.

In the learning-based models, sparse representation and dictionary learning have generated wide interest in various fields, especially in the image processing community. Huang and Song (2012) presented a sparse representation-based spatiotemporal reflectance fusion model (SPSTFM) to produce the synthetic prediction using both prior and posterior pairs of Landsat and MODIS images and one MODIS image on the predicted date. It may be regarded as an initial attempt to introduce sparse representation into the spatiotemporal fusion filed. Following SPSTFM, Song and Huang (2013) developed another dictionary learning-based spatiotemporal fusion model to produce synthetic predictions using only one base Landsat-MODIS image pair (SP-One for short hereafter). One of the greatest strengths of the learning-based models is that it can predict both phenology and type changes. However, the learning-based models only use the statistical relationship between the fine- and coarse-resolution images pair instead of taking any physical properties of remote sensing signals and combining the physical temporal change into the fusion procedure. Although their capability in predicting type changes has been validated, there still remain open topics in preserving better spectral fidelity and spatial details of the predictions.

Addressing the aforementioned limitations of existing spatiotemporal fusion models, we proposed a hierarchical spatiotemporal adaptive fusion model (HSTAFM) in this article. Compared with current spatiotemporal fusion models, it has the following highlights: (i) it combines sparse representation techniques into the physical fusion procedure; (ii) it can predict arbitrary temporal change including both seasonal phenology change and type change using only one image pair; (iii) it proposes a prior detection of temporal change and a two-level selection strategy of similar pixels, which ensures the accurate capturing of temporal change information. We tested the HSTAFM using both the simulated and observed dataset and compared it with other three state-of-the-art developed algorithms including STARFM, FSDAF, and SP-One. The remainder of this paper is organized as follows. Section 2 provides a detailed illustration of the proposed method. An algorithm test and comparison are described in Section 3. In Sections 4 and 5, discussion and conclusions are provided.

2. Methodology

Figure 1 presents a flowchart outlining the entire implementation of the proposed HSTAFM algorithm, including two major hierarchies: (i) super-resolution of the coarse-resolution image based on sparse representation and (ii) prediction of the synthetic data by combining the fine-resolution image and transitive-resolution image derived from the first hierarchy. All of the implementation steps are provided in detail below.

2.1. Data pre-processing

Both fine-resolution and coarse-resolution images are required to be processed geometrically and radiometrically, which is the prerequisite of implementing the HSTAFM algorithm. In this study, the Landsat-7 ETM+ images were radiometrically and atmospherically corrected using the 6S approach, while the MODIS 09GA surface reflectance products obtained from the Land Processes Distributed Active Archive Center (https://lpdaac.usgs.gov/lpdaac) were reprojected to Universal Transverse Mercator (UTM) by MODIS Reprojection Tools (MRTs). The coarse-resolution image
is then up-scaled to the same spatial resolution of the fine-resolution image, and clipped to the same extent as the fine-resolution image for each study site.

2.2. Super-resolution of the coarse-resolution image based on sparse representation

2.2.1. Theoretical basis of the sparse representation

The past decade has witnessed a growing popularity in the field of sparse representation of signals (Yang et al. 2010), and sparse representation also achieved state-of-the-art results in image processing studies, such as image super-resolution, image deblurring, image inpainting (Shen et al. 2009; Yang et al. 2010; Dong et al. 2011). Suppose some \( \sqrt{n} \times \sqrt{n} \) image patches are represented by \( \{x_1, x_2, \ldots, x_N\} \) after stacking the pixel values lexicographically, using an overcomplete dictionary matrix \( D \in R^{n \times m} \), \( n < m \), which contains \( m \) prototype atoms for \( n \) columns, the given image patch can be represented as a sparse linear combination of these atoms with corresponding coefficient vector \( \alpha \), that is, \( x \approx D\alpha \), satisfying \( \|x - D\alpha\|_p \leq \epsilon \). To achieve this representation, the dictionary \( D \) needs to be obtained through training samples that have similar features with the targeted signal \( x \), and the coefficient vector \( \alpha \) is confined to be quite sparse, namely, it includes very few non-zero components, which require being solved efficiently with respect to the dictionary \( D \) using sparse coding algorithms (Yang et al. 2010; Song and Huang 2013).

The K-singular value decomposition (K-SVD) algorithm (Aharon, Elad, and Bruckstein 2006) is employed in this study to design the dictionary for its simplicity and high efficiency. To find the best
dictionary $D$ to represent the given training samples $X = [x_1, x_2, \ldots, x_N]$, the K-SVD algorithm minimizes the following objective function:

$$
\min_{D, \Lambda} \{\|X - DA\|_F^2\} \quad \text{s.t.} \forall i, \|\alpha_i\|_0 \leq \tau,
$$

(1)

where $\Lambda = [\alpha_1, \alpha_2, \ldots, \alpha_N]$, $\|\alpha_i\|_0$ denotes the number of nonzero elements in $\alpha_i$ and $\tau$ is a prefixed threshold of the number of nonzero elements of $\alpha_i$.

Moreover, our next goal is to find the optimal representation coefficient $\alpha$ with the fewest nonzero entities, to represent a targeted signal $x$ with respect to the established dictionary $D$. Mathematically, this sparsity problem can be expressed by the following objective function:

$$
\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|x - Da\|_2^2 \leq \varepsilon,
$$

(2)

where $\varepsilon$ denotes the bias threshold. This optimization problem is known as a nondeterministic polynomial-time hard (NP-hard) problem. It can be efficiently solved by using several developed approximating algorithms, which can be categorized into three groups, (i) greedy strategy; (ii) convex relaxation strategy; and (iii) non-convex relaxation strategy (Yin 2012). In this study, we adopted the orthogonal matching pursuit (OMP) algorithm from the group (i), because the OMP is a stepwise forward selection algorithm, and it solves the optimization problem using the least-square method for each iteration. It has been found to be efficient in accelerating the convergence speed in Equation (2), and it is also open-sourcing and easy to implement.

2.2.2. Super-resolution of the coarse-resolution images using a single image pair

Due to the large spatial resolution difference between the fine-resolution and coarse-resolution images, we cannot directly capture sufficient temporal changes in fine spatial details from the coarse-resolution image (e.g. MODIS data). Therefore, our aim is to enhance the spatial details of the coarse-resolution image through one fine- and coarse-resolution image pair learning, and then to fuse the transitive-resolution image with improved spatial details and the original fine-resolution image.

Based on the related description of sparse representation in Section 2.2.1, we performed the super-resolution process in two steps: (i) the dictionary-pair training on the known fine- and coarse-resolution images ($F_1$ and $C_1$) at time $t_1$ and (ii) the transitive-resolution image prediction. For training a dictionary pair, the fine- and coarse-resolution features are extracted from the difference image space of $F_1 - C_1$ and the gradient features of $C_1$ in corresponding patches, respectively. Here, we adopted the first- and second-order derivatives as the gradients feature of the low-resolution patches (Chang, Yeung, and Xiong 2004), because they can capture the relative reflectance changes within a patch, and require low computation costs as well. The four 1-D filters are:

$$
\begin{align*}
    f_1 &= [-1, 0, 1], \\
    f_2 &= f_1^T \\
    f_3 &= [1, 0, -2, 0, 1], \\
    f_4 &= f_3^T
\end{align*}
$$

(3)

The fine- and coarse-resolution training samples are obtained by stacking these two categories of feature patches into columns while their columns are in regular correspondence. The fine- and coarse-resolution samples are denoted by $Y$ and $X$, respectively. To obtain the coarse-resolution dictionary $D_c$, we employed the aforementioned K-SVD algorithm (Aharon, Elad, and Bruckstein 2006) in training samples $X$ via the following optimization:

$$
\{D^*_c, \Lambda^*_c\} = \arg \min_{D, \Lambda} \{\|X - Dc\|_F^2\} \quad \text{s.t.} \forall i, \|\alpha_i\|_0 \leq \tau.
$$

(4)

To establish correspondence between fine- and coarse-resolution training samples, the fine-resolution dictionary is constructed by minimizing the approximation error on $Y$ with the confined
same representation coefficient $\Lambda^*$, 

$$D_f^* = \arg \min_{D_f} \{ \| Y - D_f \Lambda^* \|^2_F \}. \quad (5)$$

Given that $\Lambda^*$ has full low rank, the solution of Equation (5) can be directly derived from the following pseudoinverse format:

$$D_h = Y \Lambda^* (\Lambda^* \Lambda^{*T})^{-1}. \quad (6)$$

Then, we aimed to predict the transitive-resolution image ($T_2$) at time $t_2$ using the established dictionary pair and sparse coefficients. Since the sparse coefficients are enforced to be the same in the dictionary training process, if only we obtain the representation coefficients of coarse-resolution image patch with respect to the coarse-resolution dictionary $D_c$, the corresponding fine-resolution image patch can be reconstructed using the same coefficients and the fine-resolution dictionary $D_f$. 

First, the same first- and second-order gradient features $X_2$ are extracted from the coarse-resolution image $C_2$ at the time $t_2$. Supposing that $x_k^2$ denotes the $k$th $\sqrt{n} \times \sqrt{n}$ image patch of $X_2$, we can estimate the representation coefficients by minimizing the following $l_1$-norm problem:

$$\alpha^* = \arg \min_{\alpha} \{ \| x_k^2 - D_c \alpha \|^2_2 + \lambda \| \alpha \|_1 \}. \quad (7)$$

This optimization problem can be solved by applying the sparse coding algorithm OMP. Thus, the corresponding fine-resolution patch $y_k^2$ can be calculated as 

$$y_k^2 = D_f \alpha^*. \quad (8)$$

After stacking the predicted patches $\{ y_1^2, y_2^2, \ldots, y_k^2 \}$ into the image format $Y_2$, which is the difference between the transitive- and coarse-resolution images ($T_2$ and $C_2$) at time $t_2$, we can obtain the transitive-resolution image $T_2$ by the sum of $Y_2$ and $C_2$. Similarly, the transitive-resolution image $T_1$ at time $t_1$ can be obtained using the same process.

### 2.3. Calculation of conversion coefficient

As the transitive-resolution images are much closer to the actual fine-resolution image in spatial details through the super-resolution procedure in the first hierarchy, we assume that the pixel purity between the transitive- and fine-resolution images is approximate. Since the fine-resolution image (e.g. Landsat) is regarded as relatively homogeneous compared with the coarse-resolution image (e.g. MODIS), the pixel-based unit of the fine-resolution image is the finest spatial details that we can reach in this study.

The pixel-based conversion coefficient can be computed from the transitive-resolution images at the prior/posterior and predicted time,

$$V_t(x, y, b) = \frac{T_2(x, y, b)}{T_1(x, y, b)}, \quad (9)$$

where $V_t$ denotes the conversion coefficient of the transitive-resolution image. $(x, y)$ is the location of a given pixel. $b$ denotes the $b$th band of the transitive-resolution image.

Supposing that the transitive- and fine-resolution images from $t_1$ to $t_2$ have the same conversion coefficients, due to their approximate pixel purity. Therefore, the conversion coefficients for the fine-resolution images from $t_1$ to $t_2$ can be expressed as

$$V_f(x, y, b) = V_t(x, y, b) = \frac{T_2(x, y, b)}{T_1(x, y, b)}, \quad (10)$$
where $V_f$ denotes the conversion coefficient of the transitive-resolution image. $(x, y)$ is the location of a given pixel. $b$ denotes the $b$th band of the transitive-resolution image.

### 2.4. Pre-selection of temporal change

After the conversion coefficients have been calculated according to Equation (10), the initially predicted fine-resolution image at $t_2$ can be obtained through

$$F_2(x, y, b) = F_1(x, y, b) \cdot V_f(x, y, b).$$ (11)

The temporal change between the prior/posterior and target dates commonly includes seasonal phenology change and land-cover-type change. Previous spatiotemporal fusion models often enforce the invalid assumption that the land-cover change does not occur during the prediction period, which limits them in producing accurate fusions in some practical applications. In this study, the initially predicted fine-resolution image at $t_2$ according to Equation (11) is utilized to perform a pre-selection of temporal change.

The reflectance difference is computed between the initially predicted fine-resolution image at $t_2$ and the actual fine-resolution image at $t_1$. The difference is employed to explain how much is the temporal change instead of identifying which specific change it is. Here, we categorized all possible temporal change into two classes: significant change (including land-cover change and phenology disturbance) and non-significant change (seasonal phenology change). Each class will be tackled with different strategies to select similar pixels in Section 2.5.

If all bands of the target pixel satisfy Equation (12), it can be recognized that 'significant change' occurs at this pixel. The significant change may be land-cover-type changes or phenology disturbance. Otherwise, we will label it with non-significant change. The temporal change of the target pixel is within reasonable seasonal phenology fluctuation.

$$\left| F_2(x, y, b) - F_1(x, y, b) \right| \leq \gamma(b),$$ (12)

where $\gamma(b)$ denotes the threshold determined by a semi-empirical approach in this study. Provided that a normal distribution with respect to the difference between two remotely sensed scenes acquired on different dates for the same site (e.g. Figure 2) exists, we can identify the threshold $\gamma(b)$ with the customized confidence intervals based on the three-sigma rule (i.e. 68-95-99.7 rule) (Devore 2011). For example, if the reflectance difference that lies within one standard deviation of the mean reflectance difference is assigned as non-significant change, it

---

**Figure 2.** Histogram showing the frequency of reflectance difference between two Landsat ETM+ scenes (Figure 5(c,d)) acquired on 1 November 2000 and 7 November 2002 in the study site Shenzhen. (a–c) denote the histograms of reflectance difference in the bands green, red, and near-infrared (NIR), which satisfy the Gaussian distribution, respectively.
will be at a 68.27% level.

\[ \gamma(b) = \rho \cdot \sigma(F_2(x, y, b) - F_1(x, y, b)), \]  

(13)

where \( \rho \) denotes the cutting off probability, \( \sigma(\cdot) \) denotes the operator of standard deviation. In this study, we assigned the cutting off probability as 68.27%.

### 2.5. Selection of similar pixels

Pixels having similar spectral and land-cover information are called ‘similar pixels’, and the addition of similar pixels ensures that specific spatial and temporal information is used for the final fine-resolution image prediction. One of the highlighted strengths in the HSTAFM algorithm is the two-level selection strategy of similar pixels. It differs from the selection strategy in the STARFM and ESTARFM algorithms that only searches similar pixels from the prior/posterior fine-resolution image. In this article, based on the pre-selection of temporal change, our method will search similar pixels from both the prior/posterior fine-resolution image and the initially predicted fine-resolution image at the target time simultaneously. To be specific, if the given pixel is labeled with non-significant change, its corresponding set of similar pixels will be selected from the fine-resolution image at \( t_1 \). Otherwise, if the given pixel is labeled with significant change, it is reasonable to believe that the local land-cover type has changed obviously, or it experiences a phenology disturbance from \( t_1 \) to \( t_2 \). For this situation, the selected set of similar pixels from fine-resolution image at \( t_1 \) cannot stand for the ideal similar pixels at the predicted date \( t_2 \). It is also the major problem for previous algorithms that cannot accurately predict objects when shape or type changes and short-term disturbance are not recorded in any of the base fine-resolution images (Zhu et al. 2010). However, facing the ‘significant change’ condition, our algorithm will search other sets of similar pixels from the initially predicted fine-resolution image at \( t_2 \), and then extract the intersection of the two individual sets to obtain the final set of similar pixels.

We employed an adaptive threshold to identify similar pixels based on the reflectance difference between the candidate neighbor pixels and the central pixel. The threshold can be determined through local rules with a logistic function, and thus it is variable from window to window.

\[
Df_b = \sqrt{\frac{\sum_{i=1}^{\omega} \sum_{j=1}^{\omega} (F_i(x_i, y_j, b) - F_i(x_{\omega/2}, y_{\omega/2}, b))^2}{\omega^2}},
\]

(14)

\[
|F_i(x_i, y_j, b) - F_i(x_{\omega/2}, y_{\omega/2}, b)| \leq 0.01 \cdot 2Df_r.
\]

(15)

where \( F_i \) is the fine-resolution image (i.e. the actual fine-resolution image at \( t_1 \) or the initially predicted fine-resolution image at \( t_2 \) in this study). \((x_i, y_j)\) and \((x_{\omega/2}, y_{\omega/2})\) denote the locations of candidate similar pixel and central pixel, respectively. \( \omega \) is the moving window size.

For each candidate similar pixel, only if being equal to or greater than two bands satisfies the selection requirement will it be termed as a valid similar pixel. In this study, the lower limit of a similar pixel’s number is 15. When similar pixels are not sufficient to meet the algorithm’s requirement, the local window size will be enlarged for searching potential similar pixels. Otherwise, if the number of similar pixels is over 30, we will only select the first 30 pixels that have highest similarity.

### 2.6. Weight calculation of similar neighbor pixels

The weight accounts for how much similar pixels contribute to the final prediction of the central pixel’s reflectance. It can be determined by the spectral similarity and spatial distance difference. A greater weight should be assigned to those candidate pixels with higher similarity and closer distance to the central pixel.
The spectral similarity between candidate pixels and the central pixel can be calculated according to Equation (16),

\[ s_{ij} = \frac{P_{ij} \cdot \sum_{b=1}^{B} |F_i(x_i, y_j, b) - F_i(x_{w/2}, y_{w/2}, b)|}{\sum_{i=1}^{\omega} \sum_{j=1}^{\omega} P_{ij} \left( \sum_{b=1}^{B} |F_i(x_i, y_j, b) - F_i(x_{w/2}, y_{w/2}, b)| \right)} \tag{16} \]

where \( s_{ij} \) denotes the spectral similarity between the candidate pixel at location \((x_i, y_j)\) and the central pixel. \( P_{ij} \) is a binary matrix denoting whether the given pixel is a valid similar pixel. \( B \) is the total band number.

The spatial distance between the candidate pixels and central pixel can be computed according to Equation (17),

\[ d_{ij} = 1 + \frac{1}{\omega} \left( (x_i - x_{w/2})^2 + (y_j - y_{w/2})^2 \right) / (\omega / 2) \tag{17} \]

where \( \omega \) denotes the moving window size, which is used to normalize the spatial weight ensuring its range from 1 to 1 + \( \sqrt{2} \) (Zhu et al. 2010).

Therefore, the final weight of each similar pixel can be derived from combining these two factors using a normalization operator according to Equation (18). The range of \( W_{ij} \) is from 0 to 1, and the sum of total weight will be 1.

\[ W_{ij} = \frac{1}{s_{ij} \cdot d_{ij}} \sum_{i=1}^{\omega} \sum_{j=1}^{\omega} \frac{1}{s_{ij} \cdot d_{ij}} \tag{18} \]

### 2.7. Reflectance prediction of central pixels

After the two-level selection of similar pixels and weight calculation described above, we compute the final predicted fine-resolution image at \( t_2 \) according to Equation (19). Each prediction of the central pixel’s reflectance will be incorporated with the spectral and spatial information from its corresponding sets of similar pixels.

\[ F(x_{w/2}, y_{w/2}, b) = \sum_{i=1}^{\omega} \sum_{j=1}^{\omega} W_{ij} \cdot P_{ij} \cdot F_1(x_i, y_j, b) \cdot V_f(x_i, y_j, b) \tag{19} \]

where \( F \) denotes the final predicted fine-resolution image. \( W_{ij} \) denotes the weight. \( P_{ij} \) is a binary matrix denoting the set of similar pixels. \( V_f \) is the conversion coefficient matrix.

### 3. Algorithm test and comparison

To test the performance of our proposed algorithm, we applied it to both the simulated and observed data. Spatiotemporal fusion algorithms including the STARFM (Gao et al. 2006), the FSDAF (Zhu et al. 2016), and the SP-One (Song and Huang 2013) were chosen as three benchmarks for the performance comparison, for the following reasons: (i) minimum input requirement (i.e. only one image pair is required without any auxiliary data); (ii) simplicity of implementation; and (iii) wide acceptance within the remote sensing community. It should be clarified the SP-one algorithm provides a two-layer fusion scheme to enhance the spatial details of the coarse-resolution. Similarly, it can be applied to our algorithm as well. However, due to (i) the high computation cost of dictionary training and sparse coding, especially for larger study areas and (ii) the second layer fusion does not yield an obvious improvement than the first layer, both the SP-One and HSTAFM algorithms are enforced under the one-layer fusion strategy in the comparison section of algorithms’ performance.
3.1. Assessment of the algorithms’ performance

To quantify the comparison of fusion results derived from the selected spatiotemporal fusion algorithms, we employed some representative metrics along with a visual inspection. The average absolute difference (AAD), root-mean-square error (RMSE), and $R$-square ($R^2$) between the predicted and observed reflectance were calculated to verify the deviation.

3.2. Experiments with the simulated data

The selected algorithms were first applied to a set of simulated images, which allowed us to examine the performance without the interference of external factors such as radiometric and geometric inconsistencies between satellite sensors (Gevaert and García-Haro 2015). Unlike the simulated cases performed in Gao et al. (2006), Zhu et al. (2010) and Song and Huang (2013), we aimed to create a simulated dataset that is physically closer to the actual remotely sensed data. Thus, we digitalized a number of polygons representing six land-cover classes (i.e. forest, water, arable land, bare land, built-up area, and grassland). To be consistent with the observed dataset, we further transformed the digitalized polygons to a three-band raster image, and assigned them with real surface reflectance derived from Landsat data. Then, this image was degraded to produce the simulated coarse-resolution image using a degradation model in Equation (20),

$$C = SBF + n,$$

where $C$ and $F$ denote the simulated coarse- and fine-resolution images. The matrix $S$ denotes the down-sampling process (using the bicubic resampling at a ratio of 15:1 in this study). $B$ denotes the Gaussian blur filter. $n$ is the noise, which is assigned to be a null matrix due to the large difference between the simulated fine- and coarse-resolution images.

Similarly, we created another pair of fine- and coarse-resolution images. Thus, we have obtained two fine-resolution images (i.e. Landsat-like images) and two coarse-resolution images (i.e. MODIS-like images) for two different dates (i.e. $t_1$ and $t_2$). For the simulated dataset in Figure 2, we imposed both phenology change and land-cover-type changes. It was assumed that the temporal changes from $t_1$ to $t_2$ occurred in the following way: the elliptical grassland experienced phenomenology changes; the water body at the bottom experienced a man-made division with a built-up area; another abrupt type change was introduced in the bottom-right corner that a part of the arable land was converted to the built-up area.

We used the fine- and coarse-resolution image pair (Figure 2(c,a)) on the base date and the coarse-resolution image on the targeted date (Figure 2(b)) to predict the fine-resolution image that resembles the actual image on the targeted date (Figure 2(d)). Figure 3(a–d) shows the fused results of all four selected algorithms. It can be identified that all algorithms capture the phenomenology change of the elliptical grassland. With respect to another two sets of abrupt type changes, the STARFM seems to fail in dealing with land-cover-type changes, since both conversions from the water body to the built-up area, and from the arable land to the built-up area are not captured in Figure 3(e). The FSDAF predicts the general profile of the built-up area that is converted from the original arable land, but the conversion from water body to the built-up area is quite blurred in Figure 3(f) for its quite limited spectral information from accurate similar pixels. Compared with the STARFM and FSDAF, the SP-One (Figure 3(g)) and H-STAFM (Figure 3(h)) produce much better predictions that can capture abrupt type changes. Meanwhile, there still exist slight blurring effects at the edges of the change areas because of the large spatial resolution gap between the fine- and coarse-resolution images (a ratio of 15:1 in the simulated test). However, it can be clearly witnessed that our HSTAFM produces less noise than the SP-One in those change areas. The quantitative metrics including the AAD and RMSE in Table 1 also shows that the HSTAFM yields lower prediction biases than the other fusion models. It reveals that our proposed method is more robust to tackle with the type change prediction while capturing accurate phenomenology changes.
3.3. Experiments with the observed data

3.3.1. Phenology change prediction

A selection of the Coleambally Irrigation Area datasets (CIA) provided by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, was used for validating phenology change predictions. The CIA is a rice-based irrigation system located in southern New South Wales, Australia (145°04´ E, 34°00´ S) that has been extensively utilized for time-series remote sensing researches (Niel, McVicar, and R 2003; Van Van, Niel, and McVicar 2004; Emelyanova et al. 2013). In this study, we employed two cloud-free Landsat-MODIS pairs over the CIA during the austral growing season, acquired on 2 November 2001 and 25 November 2001. We further reselected the main irrigation area of the CIA, an area of 625 km² (1000 rows by 1000 columns at 25 m resolution). The major land-cover type in this region is crop within the irrigation area, which is surrounded by agricultural drylands and some woodlands (Emelyanova et al. 2013). From the base to predicted dates (Figure 4(c,d)), the study site witnessed significant phenology dynamics especially for the agricultural drylands and the greens. Moreover, the relatively small patches within the irrigation area lead it to be a spatially heterogeneous region for testing the fusion algorithms’ performance.

Similarly, we used Landsat-MODIS images (Figure 4(c,a)) on 2 November 2001 and the MODIS image (Figure 4(b)) on 25 November 2001 to predict the Landsat-like image on 25 November 2001. The bottom row in Figure 4(e–h) displays the predictions derived from the four selected fusion algorithms. A visual comparison between the actual observation (Figure 4(d)) and predictions (Figure 4(e–h)) shows that all predictions capture rapid phenology changes from the MODIS data.

Table 1. Quantitative assessment of predictions for the simulated dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>AAD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>STARFM</td>
<td>0.0019</td>
<td>0.0037</td>
</tr>
<tr>
<td>FSDAF</td>
<td>0.0016</td>
<td>0.0032</td>
</tr>
<tr>
<td>SP-One</td>
<td>0.0014</td>
<td>0.0027</td>
</tr>
<tr>
<td>HSTAFM</td>
<td>0.0012</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Note: The bold values represent the best accuracy performance.
while preserving the Landsat-like spatial details. In terms of the quantitative metrics in Table 2, our proposed HSTAFM achieves slightly better prediction accuracy compared with other algorithms. Meanwhile, the FSDAF also yields considerable prediction accuracy when tackling the heterogeneous study site for its incorporation of the pixel unmixing theory.

3.3.2. Type change prediction

Land-cover change stands for an important signal of the earth’s surface dynamics, which accounts for both natural disturbance and human activities (Song and Huang 2013). Meanwhile, type change prediction is also regarded as a difficulty within the spatiotemporal fusion field. To assess the proposed algorithm’s performance, we chose a selection of Landsat and MODIS data in the study site Shenzhen, China, which covers an extent of 225 km² (500 rows by 500 columns at 30-m resolution). The dataset comprises two cloud-free Landsat-MODIS image pairs acquired on 1 November 2000 and 7 November 2002, respectively. It witnesses significant land-cover type changes for some areas of the study site during this period. We employed Landsat-MODIS images (Figure 5(a,c)) on 1 November 2000 and the MODIS image (Figure 5(b)) on 7 November 2002 to predict a Landsat-like image on 7 November 2002. Though the acquisition interval of the two selected Landsat scenes is around two years, the landscape pattern remains spatially and temporally stable, except for several man-made land-cover-type conversions (e.g. the marked rectangle in Figure 5(c,d): from vegetation to built-up areas). As the two acquisition dates are within the same phenology period, most of the vegetation remains unchanged.

<table>
<thead>
<tr>
<th>Models</th>
<th>$R^2$</th>
<th>AAD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Red</td>
<td>NIR</td>
<td>SWIR</td>
</tr>
<tr>
<td>STARFM</td>
<td>0.3683</td>
<td>0.6654</td>
<td>0.6033</td>
</tr>
<tr>
<td>FSDAF</td>
<td>0.4258</td>
<td>0.7227</td>
<td><strong>0.7900</strong></td>
</tr>
<tr>
<td>SP-One</td>
<td>0.3726</td>
<td>0.6917</td>
<td>0.7293</td>
</tr>
<tr>
<td>HSTAFM</td>
<td><strong>0.4378</strong></td>
<td><strong>0.7336</strong></td>
<td>0.7676</td>
</tr>
</tbody>
</table>

Note: The bold values represent the best accuracy performance.
The bottom row in Figure 5 shows the predictions derived from the four selected fusion algorithms. Compared with the actual Landsat observation (Figure 5(d)), we can clearly identify that the STARFM (Figure 5(e)) does not succeed in capturing land-cover-type changes, and FSDAF (Figure 5(f)) only captures a blurred profile of land-cover-type changes with severe spectral distortions in the change areas. Both the SP-One (Figure 5(g)) and HSTAFM (Figure 5(h)) perform much better than STARFM and FSDAF in predicting land-cover-type changes, and their predictions are also much closer to the actual observation. However, compared with the SP-One algorithm, our proposed HSTAFM produces more spatial details with less blurring effects while preserving better spectral consistency. The quantitative metrics listed in Table 3 also demonstrate that the proposed HSTAFM yields lower biases than the other three models in capturing land-cover-type changes.

### 4. Discussion

The HSTAFM algorithm established pixel-based temporal conversion coefficients between the base date and the predicted date. Thus, there is no further constraint that the land-cover type should be stable during the prediction period. Being different from the previous STARFM-like methods that search for similar pixels only from the prior or posterior images, the HSTAFM algorithm conducts a two-level selection strategy to adaptively search for similar pixels from both the fine-resolution image at the base date and the initially predicted fine-resolution at the predicted date. It enables us to obtain a more accurate set of similar pixels. The tests with both simulated and observed data show that the proposed HSTAFM can accurately capture both seasonal phenology change

### Table 3. Quantitative assessment of predictions for the study site Shenzhen.

<table>
<thead>
<tr>
<th>Models</th>
<th>Green</th>
<th>Red</th>
<th>NIR</th>
<th>AAD</th>
<th>Green</th>
<th>Red</th>
<th>NIR</th>
<th>RMSE</th>
<th>Green</th>
<th>Red</th>
<th>NIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>STARFM</td>
<td>0.7022</td>
<td>0.6690</td>
<td>0.7925</td>
<td>0.0102</td>
<td>0.0153</td>
<td>0.0164</td>
<td>0.0163</td>
<td>0.0249</td>
<td>0.0244</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSDAF</td>
<td>0.6909</td>
<td>0.6834</td>
<td>0.7733</td>
<td>0.0103</td>
<td>0.0144</td>
<td>0.0170</td>
<td>0.0163</td>
<td>0.0235</td>
<td>0.0248</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP-One</td>
<td>0.7475</td>
<td>0.7914</td>
<td>0.8495</td>
<td>0.0090</td>
<td>0.0116</td>
<td>0.0134</td>
<td>0.0145</td>
<td>0.0191</td>
<td>0.0200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSTAFM</td>
<td>0.7576</td>
<td>0.7994</td>
<td>0.8510</td>
<td>0.0087</td>
<td>0.0114</td>
<td>0.0128</td>
<td>0.0139</td>
<td>0.0185</td>
<td>0.0200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The bold values represent the best accuracy performance.
and land-cover-type change. Meanwhile, some concerns about the implementation of the HSTAFM algorithm should be addressed.

Dictionary training from the input image pair plays a critical role in the super-resolution of the coarse-resolution images. Generally, provided that a certain spatial extent of the input image pair is used for dictionary training using the sparse representation techniques, it will contain the majority of land-cover types sufficient for constructing an overcomplete dictionary. However, the accurate predictions will be limited when a new type of land cover presents on the targeted date. Here we provided two alternative solutions addressing this issue: (i) reconstructing or transforming the input image pair to provide richer spatial and spectral features for dictionary learning and (ii) constructing a prior database containing different kinds of land-cover types for dictionary training with a high quality. Although the HSTAFM works with only one fine- and coarse-resolution image pair, the dictionary training and prediction results can be further improved if more than one image pair is available. Supposing that an overcomplete dictionary with rich features is constructed in advance, the presented method can efficiently produce dense long-term time-series synthetic predictions with limited number of input image pairs. Such a capability is of great applicability in monitoring terrestrial surface change and ecological dynamics at both seasonal and annual scales.

For the pre-selection of temporal changes, the selection criterion is highly correlated to the preset cutting off probability. As we have assigned the cutting off probability as 68.27% (i.e. \(1\sigma\)) in this study, the selection criterion is then established that ‘only if all bands of the targeted pixel satisfy Equation (12), it would be termed with a significant change’. Here we took the study site Shenzhen for a sensitivity experiment in Figure 6. Given the cutting off probability as 68.27%, the significant change areas will be largely overestimated if only \(\geq 1\) or \(\geq 2\) bands meet the selection criterion (upper row in Figure 6(b)). The strict criterion adopted in this study ensures that it yields the plausible results of capturing obvious land-cover change and other phenology disturbance. However, it we raise the cutting off probability up to 95.45% (i.e. \(2\sigma\)), the selection criterion can be adjusted that ‘only \(\geq 1\) or \(\geq 2\) bands of the targeted pixel should be matched with the requirement of Equation (12)’. It will also achieve considerable results of capturing obvious changes (bottom row in Figure 6(b)).

In the comparison section of the algorithms’ performance, four spatiotemporal fusion algorithms including the STARFM, FSDAF, SP-One, and HSTAFM were simultaneously tested. It could be clearly found that both the STARFM and FSDAF algorithms did not succeed in capturing type changes. Another widely used algorithm named ESTARFM was not added in the preliminary comparison, because it requires two pairs of fine- and coarse-resolution images on the prior and posterior dates, which may lead to its incapability for practical predictions when the base images are not
sufficient. Inspired by the implementation of the HSTAFM, we proposed a similar strategy to implement the ESTARFM using only one image pair. To be specific, it includes two major steps: (i) super-resolution of the coarse-resolution image through sparse representation techniques, which is similar to the HSTAFM in Section 2.2 and (ii) spatiotemporal fusion through the pixel unmixing. The ESTARFM algorithm introduces conversion coefficients in the original model based on the pixel unmixing theory so that homogeneous/heterogeneous pixels have different conversion coefficients in the prediction. The conversion coefficients are computed from the image pairs on the prior and posterior dates by using a linear regression. When there is only one available image pair on the prior or posterior date, we have to seek for alternative approaches to calculate the conversion coefficients. The coarse-resolution image is improved greatly in spatial details after the super-resolution procedure, and it serves as a transitive-resolution image between the fine- and coarse-resolution images. Thus, followed by the assumption that conversion coefficients between the fine- and transitive-resolution images are equal to those between the transitive- and coarse-resolution images, we can retrieve the conversion coefficients by a linear regression analysis for each similar pixel in the moving window. Further steps are the same as the implementations in Zhu et al. (2010). In this way, we can solve the hard problem that existed in the original version of ESTARFM. Experimental tests in Figure 7 also show its applicability and stability.

5. Conclusions

This article proposes an HSTAFM which integrates both the sparse representation technique and the physical temporal change process of surface objects into the fusion strategy. Its main objective is to provide a unified fusion framework for predicting arbitrary temporal changes including both phenology and type changes. The tests with both simulated and observed data show that the proposed HSTAFM can accurately capture both seasonal phenology change and land-cover-type change. Comparisons between the HSTAFM and other developed algorithms also demonstrate the HSTAFM algorithm produces consistently lower prediction biases and achieves a more robust performance in tackling arbitrary temporal changes. The presented method advances the capability for capturing accurate temporal changes in the process of fusing fine spatial details and frequent temporal coverage from multi-source remotely sensed data. It will be of great utility in practical applications, especially for the ecosystems and urban systems with rapid changes.

Acknowledgements

We would like to thank the three anonymous reviewers and external editor for providing valuable suggestions and comments, which have greatly improved this manuscript.
Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the Ministry of Science and Technology, China, National Research Program [grant number 2012CB955501], [grant number 2013AA122003], [grant number 2012AA12A407]; the National Natural Science Foundation of China [grant number 41271099].

References


