HOW TO EFFECTIVELY OBTAIN METADATA FROM REMOTE SENSING BIG DATA?

Körting, T. S.,* Namikawa, L. M., Fonseca, L. M. G., Felgueiras, C. A.

National Institute for Space Research – INPE Brazil
DPI – Image Processing Division at the OBT – General Earth Observation Coordination
Av. dos Astronautas,1758 São José dos Campos, Brazil
{thales.korting, laercio.namikawa, leila.fonseca, carlos.felgueiras}@inpe.br

KEY WORDS: Big data, Remote sensing, Metadata, Image processing, Vegetation indices, Water indices, Pattern recognition

ABSTRACT:

What can be considered *big data* when dealing with remote sensing imagery? In general terms, big data is defined as data requiring high management capabilities characterized by 3 V's: *Volume*, *Velocity* and *Variety*. In the past, (e.g. 1975), considering the computational and databases resources available, a series of Landsat-1 imagery from the same region could be considered big data. Nowadays, several satellites are available, and they produce massive amounts of data. Certainly, an image data set obtained by a single satellite, for a specific region and along time, fills the 3 V's requirements to be considered big data as well. In order to deal with remote sensing big data, we propose to explore the generation of metadata based on the detection of simple features. Besides the intrinsic geographic information on every remote sensing scene, no additional metadata is usually considered. We propose basic image processing algorithms to detect basic well-known patterns, and include them as tags, such as *cloud*, *shadow*, *stadium*, *vegetation*, and *water*, according to what is detectable at each spatial resolution. In this work we show preliminary results using imagery from RapidEye sensor, with 5 meter spatial resolution, composed by two full coverages of Brazil with RapidEye multispectral imagery (around 40k scenes).

1. INTRODUCTION

Nowadays, several references to the term *big data* are available, many of them without a proper understanding of what is the real meaning of it. Since 2001, Big data has been defined as data requiring high management capabilities characterized by 3 V's: *Volume*, *Velocity* and *Variety*, as proposed by (Laney, 2001). In (Plunkett et al., 2013), the authors provide examples of what has been considered big data so far. Examples include Web server and application logs, digital video and music, clickstream data, social networks, smartphone location-based services, real-time trading data, blogs and social media. It is clear that most of the given examples are Internet-related, however we can find other examples of big data generation, such as remote sensing.

Remote sensing satellites fill the requirements to be characterized as big data. Since at each day new images are obtained, and previously captured images are also combined as time series, it is possible to confirm the constant growing volume and also the increasing velocity of data gathering. New satellites are being launched and their design life (duration) are frequently overcome. One great example is the Landsat 5 satellite, projected with a 5-year design life, that returned scientifically viable data for 28 years (USGS, 2016). In terms of variety, remote sensing data is expanding the amount of spectral channels (i.e. Landsat 5 and 7 produces images in 8 spectral channels; Landsat 8, in 11 channels), which means different ways to capture spectral interaction between targets and eletromagnetic radiation. When we focus the analysis in the GEOBIA approach, the variety of data related to the same target increases more, since with the use of spectrally homogeneous regions, we combine the intra-region spectral information, such as average pixel values or texture, with spatial information, such as geometric features, and also relations to the neighborhood. Considering the aforementioned reasons, it is clear that remote sensing is a source of big data.

In this paper, we propose a method to work in a set of images composed by two full coverages of Brazil with RapidEye multispectral imagery (MMA, 2016), from 2012 and 2014. RapidEye images are generated from a constellation of 5 satellites located at the same orbital plane, and carrying the same sensors (BlackBridge, 2015). Available RapidEye imagery are processed into level 3A, which corresponds to geometric, radiometric and sensor correction, and mosaicked into 25 by 25 km tiles with a 5 meter pixel size, created from the acquisition sampled at 6.5 meters at the nadir. The multispectral bands are 5: blue ($0.44-0.51\mu$m), green ($0.52-0.59\mu$m), red ($0.63-0.685\mu$m), red edge ($0.69-0.73\mu$m), and near infra-red ($0.76-0.85\mu$m).

*Corresponding author

2. METHODOLOGY

In this section we describe our proposal to deal with remote sensing big data for metadata generation, which is depicted in the diagram of Figure 1. We also provide references for the used indices and algorithms to detect the presence of target patterns in images.

Our proposal is to integrate a set of simple algorithms for pattern recognition (blocks called *Detector for pattern 1 . . . N* in Figure 1) without a strong compromise with accuracy, therefore we can consider these algorithms as *weak detectors*. Our expected level of metadata to be generated is as superficial as tags like *cloud*, *shadow*, *stadium*, *vegetation*, and *water*, according to what is detectable in each spatial resolution.

The basic idea in dealing with remote sensing big data plus a stream of incoming images, is to provide a continuous workflow, which means a system that keeps running (see the block *New images*), allowing the insertion of algorithms for detecting new patterns on-the-fly (block *New detectors*). With the provided structure, the algorithm can run more than once for the same image, which is useful when parameters are changed or new algorithms are inserted.

2.1 Pre-processing

When dealing with remote sensing big data, which should include heterogeneous sources of images, specific parameters for each

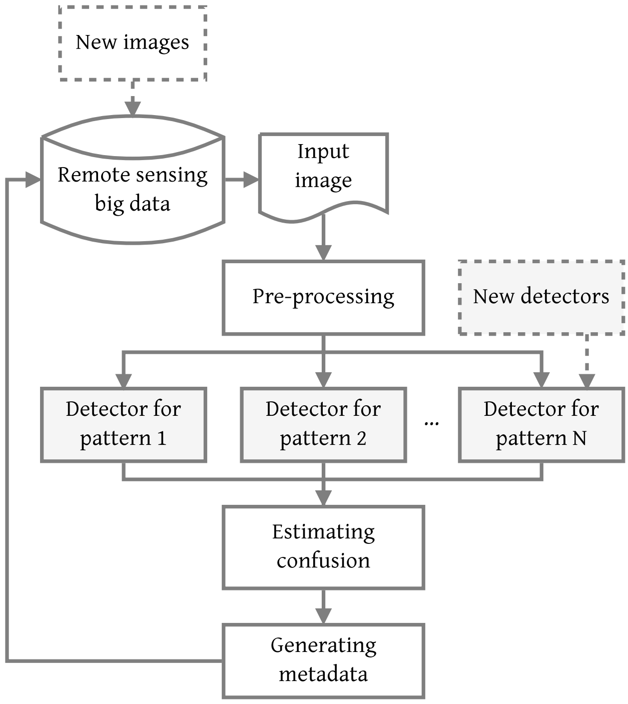

Figure 1. Diagram for metadata discovery in remote sensing big data.

sensor are often unavailable, and therefore algorithms should not rely on them. For this reason, we adopted an uniform normalization of the images, using the well-known *min-max* method (Equation 1), considering original pixel from image I at row r and column c ($r = 1 \ldots H$ and $c = 1 \ldots W$, and H and W are the Height and Width of I). This equation also uses $\min(I)$ and $\max(I)$ as the minimum pixel and maximum pixel values, respectively. As output we create the normalized image I_{norm} with values in the interval $[0, 1]$.

$$I_{\text{norm}}(r, c) = \frac{I(r, c) - \min(I)}{\max(I) - \min(I)} \tag{1}$$

This step also includes removal of null pixel values, or no data values, depending on how they are represented by the formats. Objects with these values will not be used by our algorithms. The detectors proposed in our approach are based on well-known remote sensing indices, such as vegetation or water indices. For other patterns without well-known indices, such as *stadium*, we are developing specific algorithms.

Although most of the classic indices are computed using reflectance information extracted by pixel values and often specific sensor parameters, we assume an approximation of the values by using the normalized images as inputs, since our proposal is not to map exact regions, but presence/absence of patterns.

2.2 Detectors

The main idea in our proposal is to create an image D_i, representing the *resemblance of an object* Γ *(pixel/region) to a certain pattern* i. This resemblance is based on thresholding processed images, represented as target indices, with values Φ. We defined a straightforward rule to be applied on each index, and it is up to the analyst to define the best expected threshold ζ to consider or not Γ in a specific pattern.

In Equation 2, we formalize the pattern detector. Given an object Γ, its detection index Φ and a threshold ζ, the outcome of the detector will be one of the following values:

$$\text{pattern}(\Gamma) = \begin{cases} 2, & \text{if } \Phi \geq \zeta \\ 1, & \text{if } \Phi < \zeta \text{ and } \Phi \geq 0.95.\zeta \\ 0, & \text{if } \Phi < 0.95.\zeta \end{cases} \tag{2}$$

The value 0.95 was defined empirically to distinguish a high from a medium resemblance outcome (values 2 and 1, respectively). When object Γ is not detected, the output is value 0. By applying each detector on the normalized image I_{norm}, the result is an image $D_i, i = 1 \ldots N$. All detections are then combined to estimate confusion and generate metadata.

Our preliminary results are composed by 5 weak detectors for the patterns *cloud*, *shadow*, *stadium*, *vegetation*, and *water*. For detecting the pattern *cloud*, we defined a threshold in the brightness band, with is basicaly computed by selecting, for each pixel, the maximum value among all bands. For pattern *shadow*, we applied a threshold in the near infra-red band to detect low values, adapting the proposal for Landsat TM band by (Abreu et al., 2013). The rules for pattern *stadium* are based on finding parts of images in which a compact block with high resemblance to vegetation contrasts with a surrounding area with low resemblance with vegetation. The average size of each stadium is defined as a relation between the spatial resolution of the image, the number of pixels expected to define the stadium, and it is computed using a basic highpass filtering algorithm. For pattern *vegetation*, we applied a threshold in the Normalized Difference Vegetation Index (NDVI), as described in (Rouse et al., 1974), that uses bands in the region of red and near infra-red. For pattern *water*, we applied a threshold in the Normalized Difference Water Index (NDWI), as described in (McFeeters, 1996), that uses bands in the region of green and near infra-red.

2.3 Estimating confusion

Independently of the scale of the objects (represented by Γ), in this step we process the detections for each pixel and create an output image E with estimated confusion from the previous step. By summing up individual detection values from $D_i, i = 1 \ldots N$, it is possible to infer if a pixel was detected as more than one pattern, or even if the object was not classified by any detector. It is important to highlight that in this step we account only for pixels marked with value high resemblance (value 2). Formally, the image E is computed as follows:

$$E(r, c) = \sum_{i=1}^{N} 1 \iff D_i(r, c) = 2 \tag{3}$$

The highest value we obtain in this sum, the more unconfident we are with the resultant detection. In Figure 2 we provide one example of some pattern detections and the estimated confusion.

2.4 Generating metadata

Our proposal is to output tags related to the detected patterns for each input image. It is known that certain targets appear in the images with higher probability than others. One example, when analyzing the full coverage of Brazil, is the pattern of *vegetation*. When compared to, for example, *water*, there is a strong difference in terms of area of occurrence.

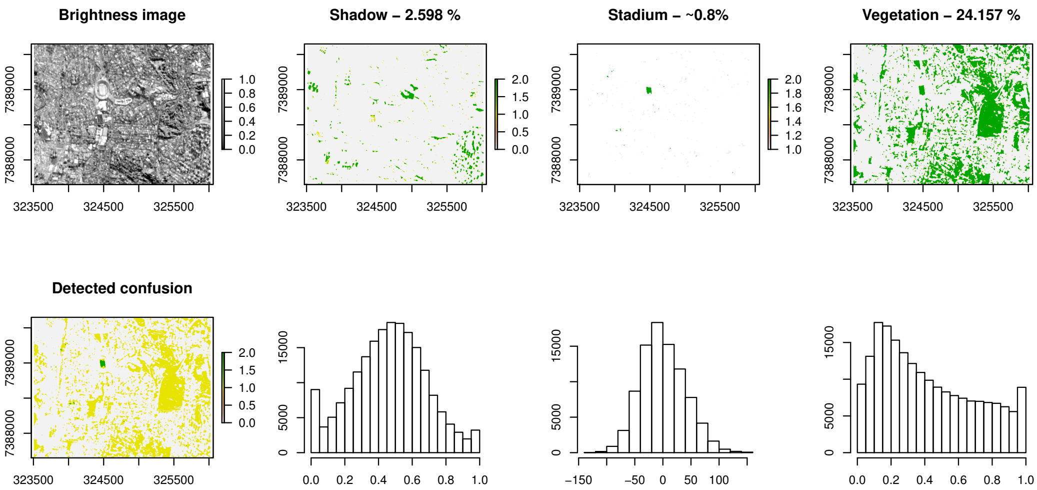

Figure 2. Estimating confusion between 3 pattern detections. Pixels marked as 2 were recognized by 2 pattern detectors. Pixels marked as 1 were detected by a single pattern, and pixels marked with 0 were not detected by any of the pattern detectors.

However, it is up to the analyst to define a minimum estimated proportion of detected targets in one image, as a threshold $\Upsilon_i, i = 1 \ldots N$, to consider the existence of the pattern i in the image, and therefore creating the tag. As we are dealing with big data, it is mandatory that Υ_i is defined as a flexible value, since each remote sensing scene is supposed to include different combinations and proportions of targets. For this purpose, we compute in the following equations the value P_i, which stands for the estimated amount of pattern i in the input image (#pattern$_i$), disregard the detected confusion (#confusion).

$$\#\text{pattern}_i = \sum_{r=1,c=1}^{H,W} 1 \iff D_i(r,c) \in \{1,2\}$$

$$\#\text{confusion} = \sum_{r=1,c=1}^{H,W} 1 \iff E(r,c) > 3 \text{ and } D_i(r,c) = 2$$

$$\#\text{total}_i = \#\text{pattern}_i + \sum_{r=1,c=1}^{H,W} 1 \iff D_i(r,c) = 0$$

$$P_i = \frac{\#\text{pattern}_i - \#\text{confusion}}{\#\text{total}_i}$$

The value 3 for considering a confusion was empirically defined (in Equation $E(r,c) > 3$), based on the assumption that if more than 3 detectors output a high resemblance for the same pattern, this value should be disconsidered. At the end, the decision for creating or not a tag for pattern i is based on the following rule:

$$\text{tag for pattern } i\text{,if } P_i \geq \Upsilon_i$$
$$\text{no-tag,if } P_i < \Upsilon_i$$

3. CONCLUSIONS

This short paper presented preliminary insights in using remote sensing big data for metadata discovery. To work with big data, the flexibility of the thresholds must be considered, and also a week compromise with positional accuracy should be assumed. The analyst plays an important role on defining thresholds for detecting patterns, and also for creating or not metadata tags for presence of targets.

We are currently running our methodology in a big data set composed by remote sensing images from RapidEye sensor, composed by two full coverages of Brazil with RapidEye multispectral imagery (around 40k scenes). New detectors must be created to extend the metadata generation for more patterns, since we are currently working with 5 patterns, namely *cloud*, *shadow*, *stadium*, *vegetation*, and *water*. Although 5 is a small number of patterns, it has allowed us to refine the methodology for estimating confusion and deciding for tag generation. The algorithms were developed in R language, using packages *rgdal, raster, ggplot2* and *sp*.

ACKNOWLEDGEMENTS

The authors would like to thank FAPESP, under grant 2016/14545-7, São Paulo Research Foundation (FAPESP).

REFERENCES

Abreu, E., Fonseca, L., Santos, C. and Ribeiro, V., 2013. Cloud Detection Tool Uma ferramenta para a detecção de nuvens e sombras em imagens de satélite. *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR* (2002), pp. 4234–4241.

BlackBridge, 2015. Satellite Imagery Product Specifications. Technical Report April.

Laney, D., 2001. 3D Data management: Controlling data volume, velocity, and variety.

McFeeters, S., 1996. The Use of the Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features. *International Journal of Remote Sensing* 17(7), pp. 1425–1432.

MMA, 2016. Geo Catálogo Do Ministério Do Meio Ambiente.

Plunkett, T., Mohiuddin, K., Macdonald, B., Harding, D., Nelson, B., Segleau, D., Sun, H., Mishra, G., Hornick, M., Stackowiak, R. and Laker, K., 2013. *Oracle Big Data Handbook*. Oracle Press, New York.

Rouse, J., Haas, R., Schell, J. and Deering, D., 1974. Monitoring Vegetation Systems in the Great Plains with ERTS. *NASA Special Publication* 351, pp. 309.

USGS, 2016. Landsat 8 (L8) Data Users Handbook.