MACHINE LEARNING APPROACH TO FUSE MULTIPLE BAND FOR WATER BODIES DETECTION

G. Calvario,
Dptmt. of Comp. Sci. and A. I.
Faculty of Informatics
University of Basque Country
P. Manuel Lardizabal 1, 20018 Donostia
Spain

C. Hernandez,
Dptmt. of Comp. Sci. and A. I.
Faculty of Informatics
University of Basque Country
P. Manuel Lardizabal 1, 20018 Donostia
Spain

E. Lazkano,
Dptmt. of Comp. Sci. and A. I.
Faculty of Informatics
University of Basque Country
P. Manuel Lardizabal 1, 20018 Donostia
Spain

B. Sierra
Dptmt. of Comp. Sci. and A. I.
Faculty of Informatics
University of Basque Country
P. Manuel Lardizabal 1, 20018 Donostia
Spain

O. Dalmau
CIMAT Mexico

P. Manuel Lardizabal 1, 20018 Donostia
Spain

T. Alarcón
CUValles Jalisco Mexico

Abstract—Classifying surfaces and analyzing changes are among the most common applications of remote sensing. To detect water bodies changes, water features are extracted using satellite data—taken in different days—, and after analyzed and compared to detect changes.

The novelties presented in this paper are threefold: firstly, a new approach for water change detection; it is based on Machine Learning applied to a set of features, namely the Landsat 8 bands, and combinations of these bands, which are commonly used as Water Indexes.

Secondly, regarding the single band thresholding, it is necessary to identify an adequate threshold; this paper presents an approach to detect automatically an appropriate threshold value.

The third advance is related to the applicability of the proposed approach for surface water change detection during a time period.

The approach has been evaluated in comparison with common change detection methods, obtaining promising results.

1. Introduction

Water bodies detection is one of the most common applications of remote sensing. In its supervised form, field observations are used to train a classifier to detect water bodies in a determined area from its spectral radiance or reflectance, texture [18] and, in object-based classification, the shape, size and context of image segments [2]. These values can be derived from a single scene, or from a combination of scenes. For lakes nearby areas, the focus of this study, supervised classification has been used with
passive satellite images. Surface waters such as lakes, rivers, artificial reservoirs, and seas are essential for climate equilibrium, hydrological cycle as well as ecosystem balance, providing fundamental resources to terrestrial life [27], [31]. Comprehensive and precise knowledge in terms of size-distributions and dynamics of surface water bodies is of crucial importance for the conservation of related aquatic biodiversity.

Among various satellite sensors, the Landsat satellite series has been most frequently used to detect long-term spatiotemporal variations in land covers thanks to its remarkable data availability, global coverage and continuity over 40 years [1], [29]. Research devoted to this satellite resource is, therefore, meaningful to extend the water monitoring mission of Landsat series.

In this paper we focus on the combination of different aspects which need to be taken into account when water body detection is to be done:

- State-of-the-art water body detection paradigms are analyzed and tested, and a new approach on water detection from satellite images is presented, based on the combination of some of the existing ones.
- Application of different Machine Learning classifiers which combine different Landsat-8 bands – Blue, Green, Red, Near Infrared (NIR), SWIR 1 and SWIR 2 bands – aiming at improve the water bodies detection existing approaches. As a matter of facts, some well known water detection filters have been used as features for the classification algorithms.

2. Related work

Water bodies detection and tracking systems have been studied extensively due to the increase of demand [10]. In this paper we focus on two main aspects of water bodies detection from satellite images:

- Water Body Detection Approaches.
- Machine Learning to deal with Multi-Spectral Images

2.1. Water Detection Approaches

Several methods have been developed to detect water bodies by means of remotely sensed imagery. The most commonly used methods fall into the following categories:

1) Spectral bands: these methods identify water bodies by applying thresholds to one or more spectral bands; in general they are easy to implement, but often misclassify mountain shadows, urban areas or other background noise, classifying them as water bodies [23].

2) Classification: supervised or unsupervised machine learning algorithms are used to extract water bodies from multispectral imagery. In the supervised classification approach, maximum-likelihood classifiers, decision trees, artificial neural networks and support vector machines are the most notable paradigms. For unsupervised classification, Kmeans and iterative self-organizing data analysis are used [12], [15]. These approaches may achieve higher accuracy than spectral band methods under some circumstances; nevertheless, expert experience or existing reference data are needed to select appropriate training samples, which avoids the use of these methods from being applied over large areas [6].

3) Water indices (WIs): combine two or more spectral bands using various mathematical operations to enhance the discrepancy between water bodies and the rest. The principle used in most WIs is based on that of the normalized-difference vegetation index (NDVI) [26].
Apart from those aspects, Temporal evolution analysis can also be performed, to this end all above mentioned points need to be considered.

2.2. Classification Techniques

[34] present a systematic surface water extraction method by taking advantage of the complementarity between a water index (WI) and a modified FCM (WIMFCM) using the Landsat-8 Operational Land Imager (OLI) images. Most of the above mentioned approaches have mostly used predefined water bodies; few works consider the application of learning techniques.


[30] present an automated procedure that allows mapping of the actual number, size, and distribution of lakes using Landsat images. The same Research Team present, in a more recent work [31] GLOWABO, a system which allows for the global-scale evaluation of fundamental limnological problems, paving the line to an improved quantification of limnetic contributions to the biogeochemical processes at large scales.

The coastal zone of the Nile Delta witnessed several changes during the last century. [4] estimate the spatiotemporal changes occurred in the coastal zone between 1973 and 2007.

In the study of [21] a new approach for surface water change detection is introduced, integrating pixel-level image fusion and image classification techniques. Artificial neural network (ANN), support vector machine (SVM), and maximum likelihood (ML) classification techniques were applied to extract and map the highlighted changes.

The approach most related to our work found in the literature is the one by [5], in which classifying surface cover types and analyzing changes are among the most common applications of remote sensing. They present a new Automated Water Extraction Index (AWEI) to improve classification accuracy in areas with shadow and dark surfaces.

Other authors present new water indexes to carry out shoreline detection in vast lakes. For instance, [16] present W1 index, based on a logical combination of the Tasseled Cap Wetness (TCW) index and the Normalized Difference Water Index (NDWI).

3. Study areas and data sources: Lake Chapala and Landsat 8 Imagery

Lake Chapala (Jalisco, Mexico) within latitude parallels 20_070 and 20_210 North and longitude meridians 102_4004500 and 103_2503000 West [11]. Located east of the State of Jalisco and northwest of the State of Michoacán, is the largest lake in Mexico, has a maximum area of 114,659 ha, of which occupies 86 percent Jalisco and Michoacán 14 percent, and the main source of drinking water in Guadalajara, Jalisco because contributes 60 percent of the water coming into the city [3].
Figure 1. Lake Chapala, Jalisco, Mexico (Source: Google Maps and Google Earth)

We have used a time series of Landsat-8 OLI images acquired from June 2013 to May 2014 (one image per month) of Lake Chapala in Jalisco, Mexico. The images were acquired from [28], as Level 1 Terrain Corrected product (L1T), and pre-georeferenced to UTM zone 13 North projection using WGS84 datum. Table 1 presents the name, date, overall cloud cover and cloud cover around the Lake Chapala of time series of Landsat-8 OLI images.

**TABLE 1. LANDSAT-8 IMAGERY USED FROM JUNE 2013 TO MAY 2014.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Cloud Cover</th>
<th>Cloud Cover %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC080294620131641LGN01</td>
<td>2013-06-10</td>
<td>2.48</td>
<td>0.000</td>
</tr>
<tr>
<td>LC0802946201312092LGN00</td>
<td>2013-07-28</td>
<td>9.57</td>
<td>0.440</td>
</tr>
<tr>
<td>LC0802946201312421LGN00</td>
<td>2013-08-29</td>
<td>18.64</td>
<td>0.000</td>
</tr>
<tr>
<td>LC0802946201313273LGN00</td>
<td>2013-09-30</td>
<td>20.75</td>
<td>0.257</td>
</tr>
<tr>
<td>LC0802946201313289LGN00</td>
<td>2013-10-16</td>
<td>30.74</td>
<td>1.186</td>
</tr>
<tr>
<td>LC0802946201313213LGN00</td>
<td>2013-11-17</td>
<td>20.41</td>
<td>0.000</td>
</tr>
<tr>
<td>LC0802946201313237LGN00</td>
<td>2013-12-03</td>
<td>2.54</td>
<td>0.000</td>
</tr>
<tr>
<td>LC0802946201402091LGN00</td>
<td>2014-01-20</td>
<td>0.17</td>
<td>0.013</td>
</tr>
<tr>
<td>LC0802946201404161LGN00</td>
<td>2014-02-05</td>
<td>0.16</td>
<td>0.000</td>
</tr>
<tr>
<td>LC0802946201404864LGN00</td>
<td>2014-03-25</td>
<td>0.18</td>
<td>0.000</td>
</tr>
<tr>
<td>LC0802946201404160LGN00</td>
<td>2014-04-10</td>
<td>0.19</td>
<td>0.000</td>
</tr>
<tr>
<td>LC080294620141322LGN00</td>
<td>2014-05-12</td>
<td>7.67</td>
<td>0.012</td>
</tr>
</tbody>
</table>

To prepare the input satellite images for further processing, the usual pre-processing steps were performed: geometrically and orthometrically correction, radiometric calibration, co-registration of images, tessellation and mosaicking. We have omitted the clouds removal step because, on selected dates, there was no accumulation of clouds in the lake area (see Table 1).

4. Methods

Once acquired and pre-processed images in the time series, we have calculated several combinations of spectral bands and vegetation and water indices that have been used with the original bands in the classification process. This section presents three subsections, which are devoted, respectively, to the Water Index combination, Training and Testing selection and thresholding, and time series analysis.

4.1. Combination of bands (bands ratio) and Vegetation and Water indices

In order to finally classify the targets as water or not water, we have used six bands of each Landsat image (RGB bands, Near infrared band and two Shortwave infrared bands), four combinations of bands (bands ratio), one vegetation index (NDVI index) and four indices related with water detection (NDWI indices). These combinations of bands and vegetation and water indices have been widely used in the literature of detection of water bodies. We have used the following bands ratios: Mid-Infrared Stress Related Vegetation Index (MSVI2) [24], Ratio Vegetation Index (RVI) [17], Moisture Stress Index (MSI) [9], Midinfrared index (MidIR) [14]. We have also extracted, based on satellite imagery, the vegetation index known as Normalized Difference Vegetation Index (NDVI) [22] and four versions of Normalized Difference Water Index (NDWI): NDWI1 = NIR-SWIR/ NIR+SWIR known as Normalized Difference Moisture Index (NDMI) [7], NDWI2 = G-NIR/ G+NIR [13], NDWI3 = R-SWIR/ R+SWIR [20] and NDWI4 = G-SWIR/G+SWIR known as Normalized Difference Snow Index (NDSI) [33]. Train and test set for classification and thresholding

First, we have selected the training data (train and test sets) into known zones of water bodies and areas which do not contain pixels of water from specific software and panchromatic first image of time series (reference image). We have used the Semi-Automatic Classification Plugin [Congedo(2013)] of [19] to create the ROIs that were used as training data set for supervised classification of pixels (water and non-water). We have selected 25 ROIs with 39349 pixels
which 16614 samples were of class "water" and 22735 samples were of class "no-water". The training data set was composed of 3674 pixels which 712 pixels of "water" class and 2962 pixels of "no-water" class and the test data set was composed of 33675 pixels which 15902 pixels of "water" class and 19773 pixels of "no-water" class.

Once the training set is selected, we have used three known classification algorithms: KNN, J48 and Naïve Bayes [32]. To evaluate the accuracy of the classifiers, we have run 5-times repeated 2-fold cross-validation and we have collected statistics from this step of process. Finally, we have generated the final classifiers which have been applied to the test set with the aim of validate the performance of the classifiers over each image in the time series. Classification algorithms have been used from "RWeka" R-package [8].

A second contribution here presented is a proposal to determine the threshold of the vegetation and water indices; based on the training data set used in the classification, obtained from the reference image, we select a Machine Learning paradigm to decide the exact value of the threshold to be used for each water index. Afterwards, we have used these thresholds to binarize the vegetation and water indices throughout the year and to collect statistics of thresholding. For thresholding of vegetation and water indices, we have used a CART classification tree [25].

5. Experimental results

Performance of the water bodies detection system was evaluated in terms of detection rates and false positives or negatives. The same accuracy validation process has been applied to each image on each water detection standard approach. We have evaluated the classifiers (KNN, J48, Naïve Bayes) and obtained rules of thresholdings for vegetation index (NDVI) and water indices (NDWI1, NDWI2, NDWI3, NDWI4) from test set of reference image. Figure 2 shows the evaluation of classifiers and thresholdings from test data all year long. With only two exceptions all of the algorithms have provided high-quality results. It is worth mentioning that the first thresholding of water index (NDWI1) has a kappa of 0.60 and J48 classification algorithm has a kappa value less than 0.9. Therefore, we can say that practically all classifiers and thresholdings have given us a good rating for the first image in the time series (or reference image). We have decided not to show in the experimental results figures the results of NDWI1, as they distort the rest of the index and classifiers results. As we can see, for most of the algorithms the kappa value throughout the year is above 0.7 (except thresholding of water index NDWI1 and J48 algorithm). We can see small variations throughout the year probably due to seasonal variations but almost all algorithms follow the same trend.

As it can be seen, all approaches made good water discrimination with the exception of thresholded NDWI1, as has been already noticed in the previous subsection. Therefore, from now on the results obtained with the threshold of this index are omitted, as they disrupt the figures and tables.

The number of pixels identified as water is almost constant throughout the year (see Figure 3). It can be seen that most algorithms (except in this case the algorithm J48) maintain the same number of pixels of the year especially towards the end of the time series. For image differencing, we can see that changes (0 value) and no-changes (-1,+1 values) remain semi-constant throughout the year (Figure 3). In the experimental performed, the no-change is the most
stable feature, which can also be observed for image rationing (Figure 3).

6. Conclusions and Future Works
Water body monitoring is essential for the effective management and conservation of water resources, which is enormously benefited from the use of remotely sensed images; an efficient as well as robust method to perform water detection from satellite images remains challenging due to noise sources and heterogeneous backgrounds. A robust methodology was designed in this study to deal with water bodies from multi-temporal and multi-spectral Landsat-8 images. The proposed approach took advantages of the complementarity between water detection Indexes and Machine Learning paradigms. Moreover, a temporal series analysis method as well as a thresholding strategy for known Water Indexes has been incorporated into the proposed method to reduce the requirement of manual actions.

It is worth mentioning that, with the use of a supervised classification approach to find the appropriate threshold for each water index, this work has been highly simplified. We do not need to find the threshold for each index and image, it is obtained automatically.

Although the current work makes progress on the improvements of water extraction accuracy in heterogeneous backgrounds, a full automation of the methodology in different sites remains lacking due to the parameter tuning issue in extreme situations. Further experiments in different regions and with different classifiers are still necessary to enhance the adequateness of the method given the high diversity of aquatic environments globally. Nevertheless, the findings obtained in this study can provide a beneficial idea to enhance the accuracy and generalizability of surface water detection using satellite images.

Figure 2. Accuracy, Sensitivity, Specificity, Kappa of all methods and for test set corresponding to each image throughout the year.
References


