Data Fusion Approach for Employing Multiple Classifiers to Improve Lake Shoreline Analysis

Alejandra A. López-Caloca

Centro de Investigación en Geografía y Geomática, Mexico D.F., Mexico
alecaloca@yahoo.com

Abstract. Remote sensing images have been widely employed to analyze bodies of water and have become essential to studying their dynamics. While the use of indices based on the threshold segmentation technique is preferred, the search for methods that define water edge contour continues. The segmentation algorithm introduced in this study is based on Mean-Shift and Watershed methods. We propose a fusion classifier strategy which allows us to obtain results that are consistent with the segmentation process. The use of two or more segmentation processes has been shown to improve pattern recognition. It is important to implement a good data integration scheme. Preliminary results suggest that the approach reported herein can improve the definition of lake shorelines.

Keywords: multi aspect, data fusion, information fusion, fusion of sensors, methodological frameworks.

1 Introduction

Data integration involves understanding the complexity and heterogeneity of several kinds of data; analyzing possible technological solutions; evaluating the characteristics of the integration process; selecting data sources to be included; and proposing a resulting model that presents a consistent interpretation of the data. Recent literature reports on the fusion of data, information and sensors as well as data and information integration. Over the last few years, the research community in this area has achieved important advances since data fusion is not limited to one field but rather has spread into medicine, biology, geosciences, geomatics, robotics, air and space systems and security activities, among other areas. For some time now, information from multiple remote sensing sources has become of interest for data fusion knowledge. The key challenge is to acquire a comprehensive view in order to take advantage of the data provided by different sensors, analyze redundancies and complementarities in the available data and generate relevant information.

Although pattern recognition systems have traditionally employed only one classifier, currently a combination of different classifiers is being proposed in order to obtain a comprehensive classification. Thus, data fusion is a tool which considerably improves the recognition of objects under study and leads to a classification of data fusion algorithms. Most data fusion methods currently work with terms such as classifier fusion, multiple classifier systems (MCS) and classifier ensembles. MCS have performed better than a single classifier and improve the efficiency and robustness of

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 1022–1029, 2014.

the pattern recognition system. Several works have proposed simultaneously using two or more classifiers that complement each other, along with remote sensing data [1][2][3].

In order to demonstrate the use of MCS to identify shorelines, this work employs the strategy of using remote sensing data to identify objects such as bodies of water and determine the boundaries of lakes. The problem with segmenting inland waters is the land-water interphase [4]. It is difficult to determine the shoreline, or water boundary, because the transition between the land and the water fluctuates due to internal changes in a lake. Several methodologies have been implemented to improve the extraction of bodies of water, such as visual digitalization and classification and segmentation by thresholds using indexes such as Normalized Difference Water Index (NDWI).

In order to extract this pattern, different water indexes are calculated using multispectral data from the Landsat TM. The Mean-Shift Segmentation (MSS) and the Watersheds Segmentation (WS) methods are then used. The body of water is classified and separated from the soil and vegetation background using both methods and the results of these classifications are processed with the data fusion module. Based on the results obtained, the surface water mapping is compared to the reference water map and field data. The resulting fused map presents an improved definition of the shoreline.

The following is presented below: Section 2 contains the basic principles related to data fusion work; Section 3 describes the case study and data treatment; Section 4 describes the experimental set-up and the results; and lastly, Section 5 presents the conclusions and future work.

2 Data Fusion Framework

Data fusion techniques that use classifiers assume that all classifiers are equally expert and complementary. Each classifier makes a decision regarding each study pattern presented. The design stage is important to ensure the effectiveness of a multi-classifier system. Basic design elements are based, firstly, on the information within the system. Observations constitute the input data that characterize the pattern recognition. In our case, observations were made by means of satellite images and field data. Then, a topology is proposed with a parallel, serial, conditional, hybrid or hierarchical structure. In our case, we employed a parallel structure (Fig.1). The segmentation and classification algorithms must then be defined, taking into account the desired characterization, the established accuracy and the interoperability desired from the final results. For this last step, the use of water indices enhancements and their segmentation is proposed. Lastly, information aggregation algorithms are defined.

The most commonly used fusion methods can be divided into three groups [5] : first, probabilistic methods such as Bayesian inference, Bayesian networks and Dempster-Shafer inference; second, methods based on artificial intelligence techniques such as abductive reasoning and semantic fusion; and third, methods based on information theory, that is, majority voting. Simple voting (majority voting) is the most common decision fusion. To combine the results from input classifiers, the classification of each input is considered as an equally weighted vote. As described in Tsymbal et al, “the class value that receives the biggest number of votes is selected as the final classification” [6]. In this work, majority voting is applied to obtain the final decision map.

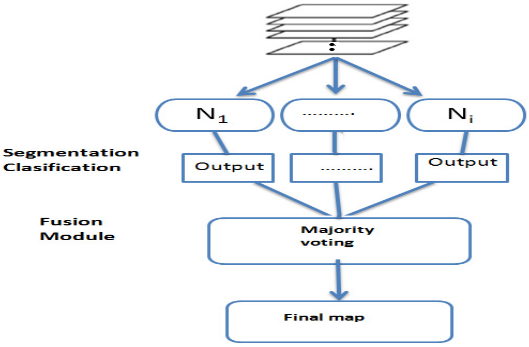

Fig. 1. Majority voting scheme. This method considers that the class with the highest number of votes is selected for each pixel.

Consider a basic scheme for classifier combinations, assuming that *n* classifiers produces a unique decision for each training input [7] [8]. In the decision fusion module, individual decisions are independent of each other and each voter has the same probability *p* of voting one way. The sample is assigned the class for which there is a consensus. The combined choice is the probability of consensus being correct. The majority voting rule is defined as follows:

$$\text{Let} \quad \Delta \mathrm{k_i} = \begin{cases} 1, if\ p(\mathrm{x_i}|w_k) \\ 0 \quad otherwise \end{cases} \tag{1}$$

where Δk_i represents binary-valued functions; w_k possible classes assigned for $k=1,2,3,\ldots\ldots, m$ possible classes; $p(x_i\ |w_k)$ is the probability density function of x_i given class k; and $i = i^{th}$ is the classifier group.

3 Case Study and Data Treatment

From the point of view of human well-being, inland waters play an important role for the community. Remote sensing has become a prime source of information to monitor surface waters. The basic problem is how to use satellite data to automatically create the separation between a body of water and the surrounding grounds. It is important to take into account the difficulty of defining the shoreline of a lake. Water levels change at the land-water interphase at the edge of a body of water, which frequently contains sediments in suspension (mud). The main problem with image processing segmentation is the intensity changes that occur along the edge of the shoreline between open water and dry land. Therefore, extraction methods need to be improved to enhance the contrast between open water and the surrounding land cover. Several techniques have been used for the water segmentation process, such as the water index method [9], [10]; iterative classification algorithms [11]; and morphological segmentation [12]. The work by Verpoorter et al. [13] presents a method called GWEM, which proposes the combination of several sources of information obtained through thresholding and classification techniques and establishes rules to generate supervised classes. The work by Lopez [14] proposes a data fusion module using the spatial classifiers Markov Random Field and support Vector Machine. This process is useful to determine several classes, such as water, water/sediments and vegetation.

The most successful method used to identify water features is the water index, since it can contrast the boundary between the body of water and other classes surrounding the lake. The idea behind water index techniques is to apply band ratios using the following spectral bands (μm): green (0.52-0.60); near infrared (NIR) (0.77-0.90); middle infrared (MIR) (1.55-1.75); and short wave infra-red (SWIR) (2.09-2.35). Ji et al. [15] emphasize the importance of understanding which is the best indicator since the indexes may result in different definitions of the border of a body of water. For the Landsat 7 TM sensor, the use of NDWI with the green and NIR bands is recommended. Bai et al. [16] recommend the use of green and SWIR bands based on the evaluation of different sensors, including Landsat, ASTER, SPOT and MODIS. For our purposes, we will describe the strengths of three different types of NDWI and evaluate two segmentation processes, considering two classes: water and no water. The objective is to improve the delineation of water with the use of a fusion module.

3.1 Data Sources and Reference Data Set

Landsat image. For the purposes of this work, we will focus on the use of Landsat 7 TM images. Landsat 7 TM images were acquired from the United States Geological Survey (USGS) portal, 2014. Even though the spatial resolution of the Landsat TM images is relatively high (30m) in the case of large lakes such as Chapala, the image used (Landsat 7 path 29 row 46 dated 03/10/1999) covers a scene and thereby satisfies a complete segmentation. The images were preprocessed (geometric and radiometric correction) and then transformed into reflectivity values.

Reference data. The reference data provided by Lopez et al [17] for the Landsat 7 TM 29/46 image were used. In this work, the authors compared the boundaries of a body of water obtained with satellite data to volume data obtained with field measurements from the lake.

3.2 NDWI and Segmentation

We need to evaluate all the water indexes to determine the one with the best performance. We tested three different forms of the Normalized Difference Water Index (NDWI) using Landsat TM: $\text{NDWI}_{\text{L\_2,4}}$ = (Green – NIR) / (Green + NIR) [9], $\text{NDWI}_{\text{L\_2,5}}$ = (Green - MIR) / (Green + MIR) [10], and $\text{NDWI}_{\text{L\_2,7}}$ = (Green-SWIR)/(Green+SWIR) [14].

After the indexes were obtained, the image was normally segmented by using a threshold. For this method, the adjustment of the NDWI threshold is a key step in the extraction of water features. There are many methods to obtain an optimal threshold. One such method is parametric, where it is necessary to presume the probability distribution of the pixel values, i.e. a histogram form. The other method is non-parametric, for which the shape of the histogram does not need to be presumed. In the case of monitoring applications for bodies of water, this can be done through visual interpretation or by applying the Otsu method [19] [20] which is a non-parametric method, or by applying iterative processes [10]. The advantage of using NDWI is that it assigns values above 0 to the water body regions. Nevertheless, the segmentation by thresholds is not yet an optimal methodology because the intensity is not uniform throughout the contours of a body of water and soft local changes are presented along the boundary. Although the effect of the non-uniform intensity is limited to the manual interpretation of images, it is a significant problem for automated algorithms, which use the digital values (gray levels) of pixels.

We calculated the segmentation of the water index using Mean-Shift segmentation and Watershed-Based Segmentation. The Mean Shift technique is an algorithm that clusters pixels by searching for neighboring pixels within a given spatial radius. This is based on mathematical morphological principles for the growth of regions.

Mean Shift is a non-parametric segmentation technique in which a probability density function is estimated according to the pixels in an image. The dense regions with a high nearest-neighbor of pixels correspond to the local maximum of the density function. For a given pixel, the Mean Shift algorithm clusters or labels the pixels according to the center that most appears to have a bounded vicinity within a given spatial radius. The Mean Shift technique is a fixed-point iterative process that converges at a local maximum. The algorithm must proceed until a convergence is found. The Watershed technique enables extracting edges or boundaries of the regions where there is an image, since the pixels are assigned according to the spatial proximity, the gradient of its gray levels and the homogeneity of textures.

4 Experimental Results

A. Input Data. The MSS and WS segmentation methods were applied before applying the fusion module to each $NDWI_{L2,4}$, $NDWI_{L2,5}$, and $NDWI_{L2.7}$ index. Six results were generated: $NDWI_{L2,4\_MSS}$, $NDWI_{L2,5\_MSS}$, $NDWI_{L2.7\_MSS}$, $NDWI_{L2,4\_WS}$, $NDWI_{L2,5\_WS}$ and $NDWI_{L2.7\_WS}$.The first three were obtained with MSS and the rest with WS. The image processing used for all of these was Orfeo Toolbox 4.0 (OTB) (http://orfeo-toolbox.org/otb/) software. In order to perform the segmentation with the Mean Shift algorithm, the values of the algorithm's parameters were configured in the following way: spatial radius of the neighborhood, 3; range radius, 15; convergence threshold, 0.1; maximum number of iterations, 100; and minimum region size, 5 per pixel unit. For the traditional Watershed algorithm, the depth threshold units as a percentage of the maximum depth in the image was 0.01 and the flood level used to generate the merge tree from the initial segmentation was 0.1.

A mask was applied to these results to remove the bottom floor and the vegetation outside the lake to obtain only the region of the edge of the water, thereby generating a water/non-water map. The six NDWIs segmentations were evaluated based on two criteria: a correlation analysis with the reference image and a subjective detection criteria analysis.

The correlation analysis (ca) was performed for the MSS and WS segmentations versus the reference. The correlation values obtained were very high and close to each other. Therefore, a subjective detection criteria was used since it was necessary to establish criteria in order to compare the segmentation results from the NDWI against those noted in the reference image, making a visual comparison with several experts. The best results were obtained with $NDWI_{L2,5\_MSS}$ (ca=0.986), $NDWI_{L2.7\_MSS}$ (ca=0.987), $NDWI_{L2.7\_WS}$ (ca=0.985). An overestimation was observed with $NDWI_{L2,4\_WS}$ and $NDWI_{L2,5\_WS}$ compared to the reference, while for $NDWI_{L2.7\_WS}$ a lower over estimation was observed. A region was missing with the $NDWI_{L2,4\_MSS}$.

B) Evaluation of the Result. Two tests were performed with the fusion module. The first (F1) was based on the entries from the products of $NDWI_{L2,5\_MSS}$ and $NDWI_{L2.7\_MSS}$, and the second (F2) on those from $NDWI_{L2,5\_MSS}$, $NDWI_{L2.7\_MSS}$ and $NDWIL_{2.7\_WS}$. The fusion algorithm was compiled in a MATLAB environment, version R2007a. Table 1 shows the performance results from the segmentation, evaluating an overlap measurement between the automated segmentation and a reference. The results from the tests applied with the fusion module (FM1 and FM2) are included. The problem of detection and localization of abnormalities involved correct segmentation (CS), over-segmentation (OS) and missing regions.

The graphic result is shown in figure 2. Only the results from the water's edge are shown. Two colors are clearly shown in the figure, which indicate the visual distinction along the edges of the body of water

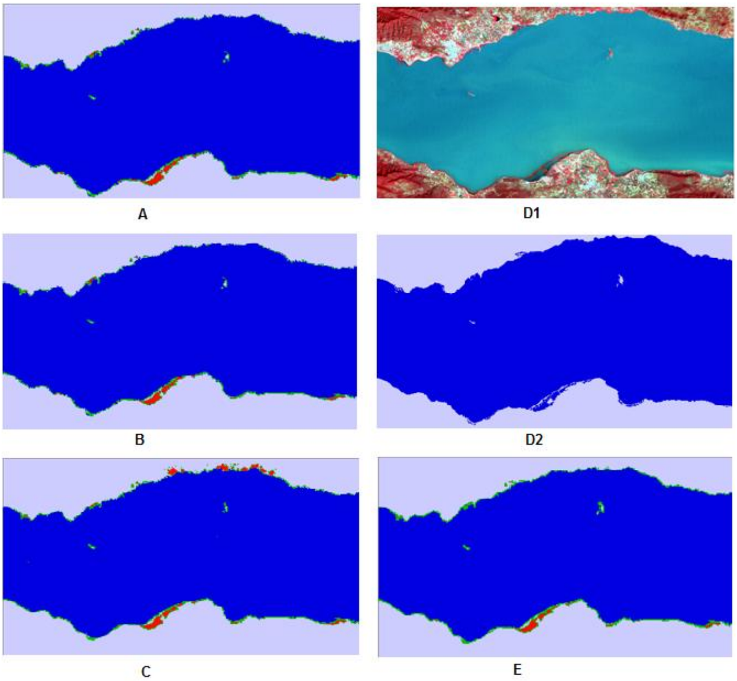

Fig. 2. Water/ Non-water map. A) $NDWI_{L2,5\_MSS}$, B)$NDWI_{L2.7\_MSS}$, C)$NDWIL_{2.7\_WS}$, D1) Landsat image, D2) reference mask, E) FM2. Correct segmentation (blue), over-segmentation (green) and missing region (red).

In the case of the overlap analysis [20], the following conditions were considered: A correct segmentation occurs when there is a high percentage of overlap between the automatically segmented region and the reference image. An over-segmentation results when the segmented image automatically presents more overlap coverage than the reference image. The third criteria, a missing region, occurs when a region noted in the reference image does not match with a valid region in the automatically segmented image.

Table 1. Overlap criteria related to the reference product. Correct segmentation % (CS), over-segmentation (OS) (unit pixels) and missing region (MR) (unit pixels). Input data NDWIL2,5_MSS, NDWIL2.7_MSS, NDWIL2.7_WS.

	$NDWI_{L2,5\_MSS}$	$NDWI_{L27\_MSS}$	$NDWI_{L2.7\_WS}$	FM1	FM2
CS	98.68	98.78	98.70	98.65	98.83
OS	>500	<500	>500	<500	<500
MR	>1000	<1100	<1000	<1000	<1100

After analyzing the 6 segmentations (Table 1), the segmentation of $NDWI_{L2,7\_ MSS}$ was determined to provide the best results, with an overlap of 98.78%, little over-segmentation and a low number of missing regions. Regarding the fused products, with the use of three classifier maps (FM2), an overlap of 98.83% was obtained with an equally low percentage of over-segmentation and a low number of missing regions.

5 Conclusions

In general, according to the experimental results, the performance of fusion classifiers positively impacts the resulting classification and the detection of water edges. The ideas proposed focused on using remote perception data to delineate an object such as a body of water. The segmentation of bodies of water using indexes is common and the effect of non-uniform intensity depends on the bands used. Visual interpretation continues to be important to this process. Automated methods of segmentation, with reduced computing time, will help to monitor lakes. Additionally, the application of methodologies for the characterization and extraction of bodies of water using satellite images makes it possible to determine morphological properties, which are valuable indicators of ecosystems.

Though the MCC application has potential, it is necessary to continue to identify and design its applications. A more in-depth investigation using Dempster-Shafer, fuzzy data fusion and weight majority voting algorithms is desired for future works.

Acknowledgements. Author is indebted to Rodolfo Sanchez and Ellen Weiss for their contribution in this work.

References

1. Schistad-Solberg, A.: Contextual Data Fusion Applied to Forest Map Revision. IEEE Transaction on Geoscience and Remote Sensing 37(3), 1234–1243 (1999)

2. Jeon, B., Landgrebe, D.A.: Decision Fusion Approach for Multitemporal Classification. IEEE Transaction on Geoscience and Remote Sensing 37(3), 1227–1233 (1999)
3. Zhang, J.: Multi-source remote sensing data fusion: Status and trends. International Journal of Image and Data Fusion 1(1), 5–24 (2010)
4. Song, C., Huang, B., Ke, L., Richards, K.S.: Remote sensing of alpine lake water environment changes on the Tibetan Plateau and surroundings: A review. ISPRS Journal of Photogrammetry and Remote Sensing 92, 26–37 (2014)
5. Castanedo, F.: Fusión de Datos Distribuida en Redes de Sensores Visuales Utilizando Sistemas Multi-Agente. Tesis Doctoral. Departamento de Informatica, Universidad de Carlos III de Madrid, Escuela Politécnica Superior 41 (2010)
6. Tsymbal, A., Pechenizkiy, M., Cunninghama, P.: Diversity in search strategies for ensemble feature selection. Information Fusion 6, 83–98 (2005)
7. Lam, L., Suen, C.Y.: Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 27(5), 553 (1997)
8. Jimenez, L.O., Moales-Morell, A., Creus, A.: Classification of Hyperdimensional Data Based on Feature and Decision Fusion Approaches Using Projection Pursuit, Majority Voting, and Neural Networks. IEEE Transaction on Geoscience and Remote Sensing 37(3), 1360–1366 (1999)
9. McFeeters, S.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing (17), 1425–1432 (1996)
10. Xu, H.: Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International Journal of Remote Sensin 27, 3025–3033 (2006)
11. Luo, J., Sheng, Y., Shen, Z., Li, J., Gao, L.: Automatic and high-precise extraction for water information from multispectral images with the step-by-step iterative transformation mechanism. J. Remote Sens 13, 604–615 (2009)
12. Bagli, S., Soille, P., Fermi, E.: Automatic delineation of shoreline and lake boundaries from Landsat satellite images. In: Proceedings of Initial ECOIMAGINE GI and GIS for Integrated Coastal Management, Seville, pp. 13–15 (2004)
13. Verpoorter, C., Kutser, T., Tranvik, L.: Automated mapping of water bodies using Landsat multispectral data. Limnol. Oceanogr. Methods 10, 1037–1050 (2012)
14. Lopez-Caloca, A.: Aplicaciones de fusión de Datos en datos geoespaciales: Caso de estudio fusión de clasificadores múltiples en el Lago de Chapala. GEOcibernética: i+g+s, Open Access, `http://www.geocibernetica.org/journal/` (to be published)
15. Ji, L., Zhang, L., Wylie, B.: Analysis of Dynamic Thresholds for the Normalized Difference Water Index. Photogrammetric Engineering & Remote Sensing 75(11), 1307–1317 (2009)
16. Bai, J., Chen, X., Li, J., Yang, L., Fang, H.: Changes in the area of inland lakes in arid regions of central Asia during the past 30 years. Environ Monitorig Assess 178, 247–256 (2011)
17. López-Caloca, A.A., Tapia-Silva, F.O., Escalante-Ramírez, B.: Lake Chapala change detection using time series. Remote Sensing for Agriculture, Ecosystems, and Hydrology X. In: Neale, C.M.U., Owe, M., D'Urso, G. (eds.) Remote Sensing for Agriculture, Ecosystems, and Hydrology X. Proceedings of the SPIE, vol. 7104, article id. 710405, p. 11 (2008)
18. Lua, S., Ouyangab, N., Wua, B., Weic, Y., Tesemma, Z.: Lake water volume calculation with time series remote-sensing images. International Journal of Remote Sensing 34(22), 7962–7973 (2013)
19. Otsu, N.: A threshold selection method from grey-level histograms. IEEE Trans. Syst. Man Cybern. 9(1), 62–66 (1979)
20. Sonka, M., Fitzpatrick, J.M.: Handbook of Med. Ima., vol. 2. SPIE Press (2000)