A Boosting-based Approach for Remote Sensing Multimodal Image Classification

Edemir Ferreira Jr, Arnaldo de A. Araújo, and Jefersson A. dos Santos
Department of Computer Science, Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 - Pampulha - Belo Horizonte - MG, CEP 31270-901, Brazil
edemirm, arnaldo, jefersson@dcc.ufmg.br

***Abstract*—Remote Sensing Images (RSIs) have been used as a major source of data, particularly with respect to the creation of thematic maps. This process is usually modeled as a supervised classification problem where the system needs to learn the patterns of interest provided by the user and assign a class to the rest of the image regions. Associated with the nature of RSIs, there are several challenges that can be highlighted: (1) they are georeferenced images, i.e., a geographic coordinate is associated with each pixel; (2) the data commonly captures specific frequencies across the electromagnetic spectrum instead of the visible spectrum, which requires the development of specific algorithms to describe patterns; (3) the detail level of each data may vary, resulting in images with different spatial and pixel resolution, but covering the same area; (4) due to the high pixel resolution images, efficient processing algorithms are desirable. Thus, it is very common to have images obtained from different sensors, which could improve the quality of thematic maps generated. However, this requires the creation of techniques to properly encode and combine the different properties of the images. Therefore, this paper proposes a boosting-based technique for classification of regions in RSIs that manages to encode features extracted from different sources of data, spectral and spatial domains. The new approach is evaluated in an urban classification scenario and a coffee crop recognition task, achieving statistically better results in comparison with the proposed baselines in urban classification and better results at some baselines for the coffee crop recognition.**

***Keywords*-Multimodal Classification; Remote Sensing; Data Fusion.**

I. Introduction

Over the years, there has been a growing demand for remotely-sensed data. Specific objects of interest are being monitored with earth observation data, for the most varied applications. Some examples include ecological science [1], hydrological science [2], agriculture [3], and many other applications.

RSIs have been used as a major source of data, particularly with respect to the creation of thematic maps. A thematic map is a type of map that displays the spatial distribution of an attribute that relates to a particular theme connected with a specific geographic area. This process is usually modeled as a supervised classification problem where the system needs to learn the patterns of interest provided by the user and assign a class to the rest of the image regions.

In the last few decades, the technological evolution of sensors has provided remote sensing analysis with multiple and heterogeneous image sources, which can be available for the same geographical region: high spatial, multispectral, hyperspectral, radar, multi-temporal, and multiangular images can today be acquired over a given scene.

Typically, these sensors are designed to be specialists in obtaining one or few properties from the earth surface. This occurs because each sensor, due to technical and cost limitations, has a specific observation purpose and operates at different wavelength ranges to achieve it. Since the sensors are specialists, they carry different and complementary information, which can be combined to improve classification of the materials on the surface and consequently increase the quality of the thematic map. In this scenario, it is essential to use a more suitable technique to combine the different features in an effective way.

The remote sensing community has been very active in the last decade in proposing methods that combine different modalities [4]. In addition to support the research on this important topic, every year since 2006, the IEEE Geoscience Remote Sensing Society (GRSS) has been developing a Data Fusion Contest (DFC), organized by the Image Analysis and Data Fusion Technical Committee (IADFTC), which aims at promoting progress on fusion and analysis methodologies for multisource remote sensing data. Also, other data fusion challenges have been proposed more recently by the International Society for Photogrammetry and Remote Sensing (ISPRS), devoted to the development of international cooperation for the advancement of photogrammetry and remote sensing and their applications. All the effort to reach advance in this research area shows the high interest and timely relevance of the posed problems.

Multimodal classification is a challenging task for several reasons. First, the data are generated by very complex systems, driven by numerous underlying processes that depend on the sensor used and a large number of variables which sometimes we have no access, e.g., the atmospheric constituents cause wavelength-dependent absorption and scattering of radiation, which degrade the quality of images. Second, combining heterogeneous datasets such that the respective advantages of each dataset are maximally exploited, and drawbacks suppressed, is not an evident task. Third, as pointed by [5], it is very difficult to conclude what is the best approach for multimodal data fusion, since it depends on the foundation of the problem, the nature of the data used and the source of information utilized.

There are also several research challenges in computational scope when working with RSI classification such as: (1) remote sensing data is inherently big, even at 250 m coarse spatial resolution, Moderate-Resolution Imaging Spectroradiometer (MODIS) product can contain more than 20 millions of pixels, jointly with a time series of five thousand observations. Most machine learning models described as a state of the art (e.g., Deep Neural Networks, non-linear Support Vector Machines), can not handle with the magnitude of this data; (2) segmentation scale, accompanied by the large amount of information at the level of object in very high spatial resolution images, segmentation algorithms have difficulty in defining the optimum scale to be used; (3) pixel mixture and dimensionality reduction, images with high spectral resolution must be preprocessed due to problems such as high dimensionality, treatment of noise and corrupted bands, mixture of pixels due to the low spatial resolution; (4) efficiency, even collecting information from various sensors, efficiency and capability to process that amount of data is desired or even crucial depending on the application. In applications such as tsunami or earthquakes, the data must be analyzed in near real time, and the difference of a few seconds can save hundreds or even thousands of lives in a seaquake.

In this work, we are interested in the use of RSI particularly with respect to the creation of thematic maps by exploiting multi-sensor data. So, in this work, we propose an approach for the classification task, projected to receive two images, over the same geographic region, with different domains as input: an image with very high spatial (VHS) resolution and another one with multi/hyperspectral (HS) resolution (Figure 1).

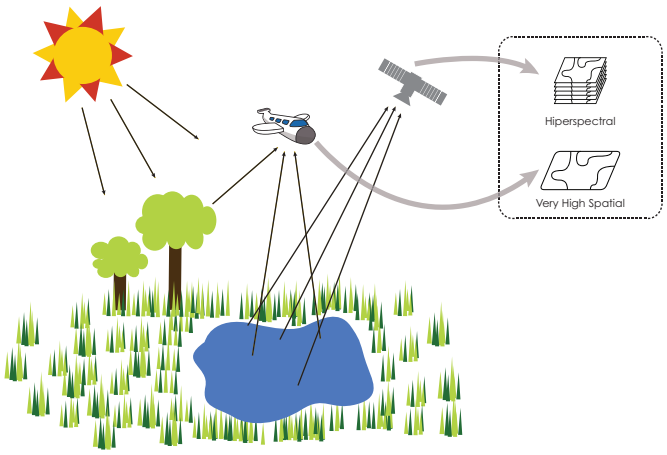


Fig. 1. An illustration of multimodal data acquisition. The figure shows two different platforms: a plane and a satellite; carrying sensors which extract different information (spectral and spatial) over the same region, creating a multimodal perspective.

Our approach is based on the SAMME Adaboost method [6], in which we created a framework based on a supervised learning scheme, divided into five steps: (1) data acquisition, the framework receives the VHS and HS images, acquired by different sensors but over the same geographic area as input; (2) object representation, the VHS image is segmented into regions using a segmentation algorithm while the HS image is analyzed by the spectral signature of each pixel; (3) feature extraction, feature vectors are extracted from the segmented regions of VHS using various descriptors and the spectral signatures are obtained by different dimensionality reduction methods; (4) training, using the features extracted from both domains and diverse learning methods a set of weak learners is created, which at every boosting iteration once is selected to compose the final strong classifier; (5) prediction, given the unseen samples and the set of selected weak learners, a predict for every new sample is made regarding to the linear combination of the weak learner predictions. In this approach, we exploit the inherent feature selection of the Adaboost for the combination of different modalities, as a natural process.

To summarize, the main contribution of this work is an approach capable of combining different modalities of sensor data by using the inherent feature selection of the boosting-based strategy.

This paper is organized as follows: Section II is an overview of the current multi-source remote sensing data fusion techniques. Section III presents details of the proposed approach. Section IV shows the corresponding evaluation protocol and experimental results of the proposed method. We conclude this work in Section V with some remarks and the future directions in the research.

II. Related work

In data fusion, each data source describing the same scene and objects of interest can be defined as a *modality*. In remote sensing image analysis, the different modalities often represent a particular data property carrying complementary information about the surface observed [7].

The joint complementarity exploitation of different remote sensing sources has proven to be very useful in many applications of land-cover classification, and the capability of improving the discrimination between the classes is a key aspect towards a detailed characterization of the earth [5]. Concerning multisource data, a diversity of fusion techniques has been proposed in the remote sensing literature, which can be divided into levels according to the modalities used in the fusion, as follows:

1) **Fusion at subpixel level**: Given k modalities datasets, which usually involve different spatial scales, the modalities are fused at subpixel level using appropriate transforms [8]. These fusions are commonly used in the cases where the main objective is to preserve the valuable spectral information from multispectral or hyperspectral sensors, with low spatial resolution, as an alternative to pan-sharpening methods which can produce a spectral distortion [9].
   In the subject of proposed works based on spectral unmixing for data fusion, the spatial and temporal adaptive reflectance fusion model proposed in [10], was used in [11] for combining information from Landsat (30-m resolution) and MODIS (250-m to 1-km resolution),

and a set of methods for increasing spatial resolution associated to [12] was used for classification task [13], [14]. An overview of the majority of nonlinear unmixing methods used in hyperspectral image processing and many recent developments in remote sensing are presented with details in [[15], [16]].

2) **Fusion at pixel level**: Given k modalities datasets, in the fusion at pixel level exists a direct pixel correlation between the modalities, which is used to produce data fusion. In general, that fusion level attempts to combine data from different sources in intent to produce a new modality, which, afterward, could be used for different applications. Some examples that rely on that case is pan-sharpening, super resolution, and 3D reconstruction from 2D views [5]. An evaluation of spatial and spectral effectiveness of more common pixel-level fusion methods was realized in [17]. Regarding [17], several pan sharpening methods have been proposed in the literature [[18], [19]] primarily based on algebraic operations, component substitution, high-pass filtering and multi resolution analysis.
   More recently, [20] made an analysis of the different fusion techniques in images, also applied to remote sensing at a pixel level, showing that all techniques have their own limitation when used individually and they also encouraged the utilization of hybrid systems.
3) **Fusion at feature level**: Given k modalities datasets, various features are extracted individually from each modality, e.g., edges, corners, lines, texture parameters, followed by a fusion, which involves extraction and selection of more discriminant attributes. Regarding [4], one of the new research directions on feature level multimodal fusion are the Kernel methods. At the domain of remote sensing, there is a considerable number of studies about kernel methods [21], once they provide an instinctive way to encode data from different modalities into classification and prediction models. One of the first attempts to combine data from different modalities, using a combination of kernel functions, was realized by [22], who created a compound kernel by using the weighted summation of spatial and spectral features from the co-registered region. Extending the proposition for more than two sources, a multiple kernel learning [23] was applied to [24] for combining spatial and spectral information, to combine optical and radar data [25], using the same sensor but in different places [26], also using different optical sensors to change detection [27].
4) **Fusion at decision level**: Given k modalities datasets, an individual process path is made for each modality, followed by a fusion of the outputs, assuming that the k outputs combined can improve the final accuracy [28]. In this way, the combination of complementary information from different modalities is done through the fusion of the results obtained considering each modality independently. There are several ways to combine the decisions, such as including voting methods, statistical methods, fuzzy logic-based methods, etc. When the results are explained as confidences instead of decisions, the methods are called soft fusion; otherwise, they are called hard fusion. An example of this type of fusion was presented in the 2008 [29] and 2009-10 [30] data fusion contests. [31] used a scheme of weighted decision fusion, which uses the SVM and the Random Forest for the probability estimation in the Landsat 8 and MODIS sensors; [32] made a combination of fusion by feature level using a graph-based feature fusion method together with a weight majority voting of outputs from differents SVM's for the classification of hyperspectral and Light Detection and Ranging (LiDAR) data.

The above-described levels do not cover all the possible fusion methods since input and output of data fusion can be different for each level of processing. In the most cases, the fusion procedure is a junction of the four fusion levels considered previously.

III. Boosting-Based Approach

In this work, we aim at exploiting multi-sensor data in a more general way, using the idea of boosting of classifiers, based on the SAMME Adaboost method [6].

The choice of an approach based on boosting is related to the inherent advantages of the strategy and its application in a multimodal classification of RSIs. Regarding the advantages, we can highlight: (1) algorithm flexibility, being possible to combine any method of learning as well extracted features obtained from different domains; (2) efficiency, when dealing with RSIs, the use of robust and efficient methods is desired, due to the complexity of the data (e.g., images with hundreds of spectral bands, very high spatial resolution) and the high computational cost for processing; (3) tuning parameters, unlike most of the robust methods in the literature (e.g., SVMs, Neural Networks) that use non-linear models thus requiring various parameter settings, the boosting approach uses a combination of weak linear models to create a more complex function, and requires only a single parameter, the number of rounds to be trained; (4) well-known algorithm, in addition to the solid mathematical foundation behind the method, the literature also indicates successful works using boosting in remote sensing [33] and for other applications to computer vision [34].

We create a framework based on a supervised learning scheme, dealing with different scenarios, regions, and objects, on the creation of thematic maps for the classification task. We propose a scheme, with a combination of a pixel, feature, and decision levels, to handle an amount of information from different modalities, and combine them for a final decision for each pixel in the thematic map. Contrary to approaches from the literature, our method uses the inherent feature selection of the boosting for the combination of different modalities, as a natural process.

The proposed method is projected to receive two images from the same place with different domains as input: an

image with very high spatial resolution and another one with hyperspectral resolution.

The boosting approach is divided into five main steps: data input, object representation, feature extraction, training, predicting. Figure 2 illustrates the proposed framework. We detail each step next.

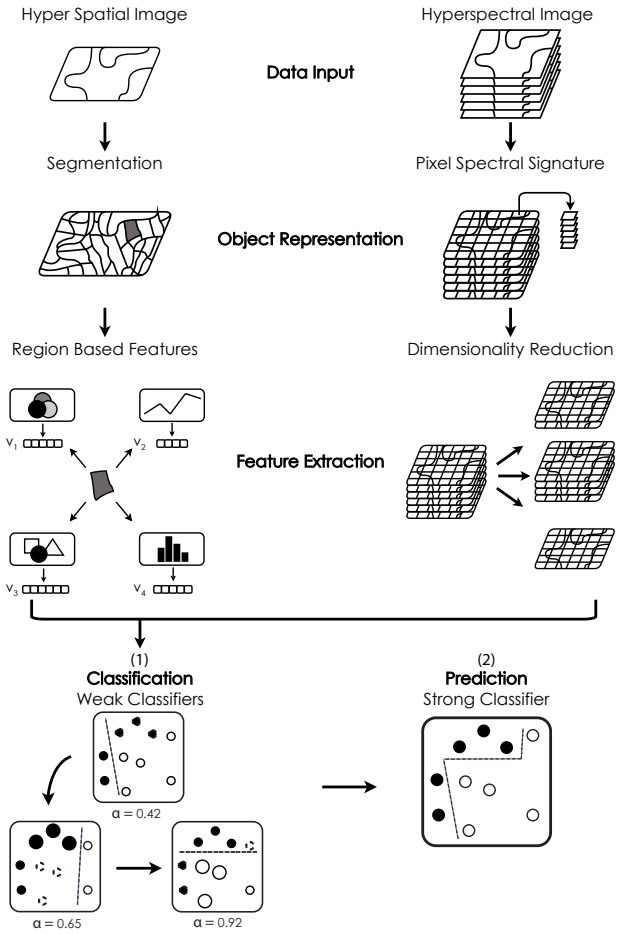


Fig. 2. The Proposed *Boosting-based approach* framework. (1) The proposed method is projected to receive two images from the same place with different domains as input: an image with very high spatial resolution and another one with hyperspectral resolution; (2) the VHS image is segmented into regions using a segmentation algorithm while the HS image is analyzed by the spectral signature of each pixel; (3) feature vectors are extracted from the segmented regions of VHS using various descriptors and the spectral signatures are projected by using different dimensionality reduction methods; Given the amount of feature extracted from both domains, the boosting training starts in (4), where for every round one weak classifier will be chosen to compose the final strong classifier. The samples which are incorrectly labeled in every round, have their weight increased and will be focused by the learners in the next round. The collection of selected weak classifiers are combined in (5) to build the strong final classifier, which is used to predict the samples of the test data regarding the confidence of each weak model.

A. Object Representation

The first step is to define the objects to be described by the feature extraction algorithms. Let I_{VHS} and I_{HS} be input images with VHS and HS resolutions, respectively. Let $Y_R^{t'}$ and Y_R^t be image labels from training and test data respectively. In an experimental scenario, $Y_R = Y_R^t \cup Y_R^{t'}$, where Y_R is the image labels of the entire dataset. For the I_{VHS}, we performed a segmentation process over the regions of Y_R^t in order to split the entire image into more spatially homogeneous objects. It allows the codification of suitable texture features for each part of the image.

Due to the low spatial resolution of the I_{HS}, we consider the pixel as the unique spatial unit. Anyway, we are more interested in exploiting the spectral signature of each pixel.

B. Feature Extraction

Concerning I_{VHS}, we have used image descriptors based on visible color and texture information to encode complementary features. For the I_{HS}, we exploit dimensionality reduction/projection properties from the spectral signature in order to obtain diversity. Notice that feature extraction process requires a region mapping between spatial and spectral resolutions, since I_{VHS} and I_{HS} are from different domains.

C. Training

Let $R = \{r_i \in R : r_1, ..., r_n\}$ be a set of regions r_i created by the segmentation process over Y_R^t. For each region r_i we have a set of features X_i extracted from the I_{VHS} and I_{HS} images. Let $X_{train} = \{(X_1, y_1), ..., (X_n, y_n)\}$ be a family of sets of features extracted from both domains, also refer here as the training data, where $X_i \subset X_{train}, \forall i$, and y_i the real label of the region r_i.

Algorithm 1 outlines the steps for the boosting approach.

Algorithm 1 Boosting-Based Approach.

Input: Number of rounds T, Number of classes K, and the training data X_{train}.
Initializing: For each region r_i, initialize the weights $w_i = \frac{1}{n}, i = 1, 2, ..., n$.
for t = 1 to T **do**
  for each learning algorithm l_i **do**
    Train weak classifiers $H_{l_i}^t(x)$ using X_{train} regarding to the weights w_i
    Evaluate each $H_{l_i}^t(x)$ on X_{train}, by computing $Err_{l_i}^t$ (Equation 1)
  end for
  Select the weak classifier H_t^* with minimum error, $Err^* = min(Err_{l_i}^t)$
  Compute $\alpha^t = ln\frac{1-Err^*}{Err^*} + ln(K-1)$
  Update the weights $w_i = w_i * exp(\alpha^t | (y_i \neq H^t(x)))$
  Normalize the weights $w_i = \frac{w_i}{\sum_{i=1}^n w_i}$
end for
Output:
F(x) = $\underset{k}{argmax} \sum_{t=1}^T \alpha^t | (H_t^* = k)$

$$Err_{l_i}^t(H_{l_i}^t(X_{train})) = \sum_{i=1}^{n} \frac{w_i | (y_i \neq H_{l_i}^t(X))}{\sum_{i=1}^n w_i} \quad (1)$$

In initializing phase, we assign for every region r_i a weight w_i with same value equal to $\frac{1}{n}$, where n is the number of

total regions (Line 2). The strong classifier $F(x)$ is built in a sequential boosting scheme. Therefore for every round t (Line 3), a set of weak classifiers $H^t_{l_i}(x)$ is trained using all features in X_{train} regarding to the weights w_i (Line 5). Afterwards, we evaluate every weak classifier in $H^t_{l_i}(x)$, by computing the accuracy weighted error ($Err^t_{l_i}$) on the same set of features X_{train} (Line 6). Given the amount of the weak classifiers $H^t_{l_i}(x)$, we select the one evaluated with the minimum error (Err^*), and compute the coefficient α^t at the round t in concern of the error (Lines 8-9). Thereafter, the regions r_i, which were misclassified at the round t, have their weight w_i updated by a factor of e^{α^t}, and finally all the regions have their weights normalized (Lines 10-11).

D. Predicting

In the same matter of the training phase, let $R' = \{r'_i \in R' : r'_1, ..., r'_n\}$ be a set of regions r'_i created by the segmentation process over $Y^{t'}_R$. For each region r_i, we have a set of features X'_i extracted from the I'_{VHS} and I'_{HS} images. So, let $X_{test} = \{(X'_1), ..., (X'_n)\}$ be a family of sets of features extracted from both domains, also referred here as the test data, where $X'_i \subset X'_{test}, \forall i$.

Once the strong classifier $F(x)$ is built, and given a new region r'_i, the features X'_i are used to predict the label for each selected weak classifiers H^*_t at every round t, and to choose the final class k regarding to the maximum argument of a linear combination of the coefficients α^t at the class k (Eq. 2).

$$F(x) = \underset{k}{argmax} \sum_{t=1}^{T} \alpha^t | (H^*_t = k) \tag{2}$$

IV. Experiments

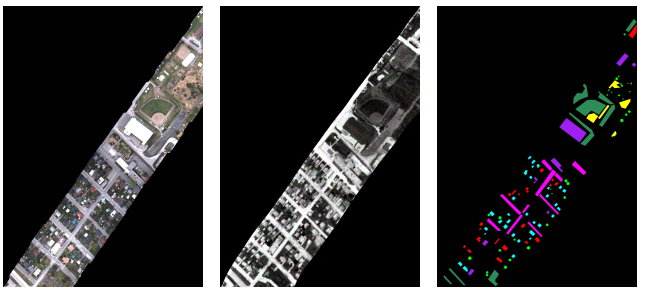
A. Datasets

1) Urban Land-Cover: This dataset was proposed in the IEEE GRSS Data Fusion Contest 2014[1], provided for the contest by Telops Inc., Québec, Canada, which involved two airborne imagery acquired at different spectral ranges and spatial resolutions: 1) a coarser-resolution Long Wave Infrared (LWIR) hyperspectral image and 2) fine-resolution visible (VIS) image. Both airbone data were acquired on May 21, 2013, using two different platforms with a short temporal gap, covering an urban area near Thetford Mines in Québec, Canada, which contain residential and commercial building, roads, vegetation, etc.

In order to perform a supervised learning, the dataset contains a training map which includes seven different classes: trees, vegetation, road, bare soil, red roof, gray roof, and concrete roof. Tables I, II and III show the main information about the dataset and the distribution of the classes in pixels. An illustration of the Urban Dataset is showed in Figure 3.

The dataset shows several challenges connected to the remote sensing multimodal classification: the multiresolution between the sources, the multisensor fusion, and also the complementarity between spectral and thermal data in terms of information extraction.

[1]2014 IEEE GRSS Data Fusion Contest. Online: http://www.grss-ieee.org/community/technical-committees/data-fusion



(a) VIS Image (b) LWIR hyperspectral (c) Ground Truth Train

(d) VIS Image (e) LWIR hyperspectral (f) Ground Truth Test

Fig. 3. Urban Dataset - IEEE GRSS Data Fusion Contest 2014.

TABLE I
General Information - Urban Land-Cover - Train.

	Visible Image	Hyperspectral Image
Width	2830 pixel	564 pixel
Height	3989 pixel	755 pixel
Spatial Resolution	0.2 m	1 m
Bands	3	84

TABLE II
General Information - Urban Land-Cover - Test.

	Visible Image	Hyperspectral Image
Width	3769 pixel	751 pixel
Height	4386 pixel	874 pixel
Spatial Resolution	0.2 m	1 m
Bands	3	84

TABLE III
Distribution of Pixels per Class.

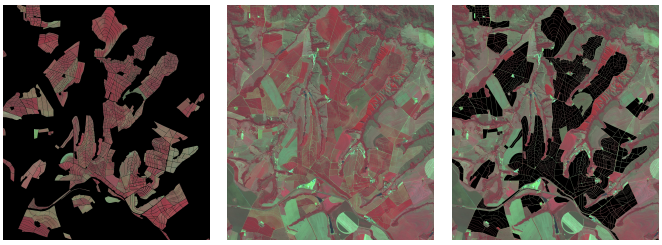
Classes	Train Pixels (%)	Test Pixels (%)
Bare soil	7.87	3.38
Road	19.79	55.73
Trees	4.87	6.93
Vegetation	32.61	7.13
Red roof	8.19	9.41
Gray roof	9.42	9.84
Concrete roof	17.21	7.54

2) Coffee Crop Recognition: This dataset is composed of two satellite-based imagery acquired at different spectral ranges and spatial resolutions: 1) moderate-resolution imaging spectroradiometer (MODIS) image and 2) high-resolution visible image. The satellite-based data were acquired on the following dates: May 21, 2010, and September 24, 2011, covering a region of coffee cultivation in Patrocínio, Minas

Gerais, Brazil. The MODIS utilized is a Surface-Reflectance Product (MOD 09), computed from the MODIS Level 1B land bands 1, 2, 3, 4, 5, 6, and 7 (centered at 648 nm, 858 nm, 470 nm, 555 nm, 1240 nm, 1640 nm, and 2130 nm, respectively).

The high-resolution visible image was obtained by using the satellite *Satellite Pour l'Observation de la Terre* (SPOT) 5, which offer a higher resolution of 2.5 to 5 meters in panchromatic mode.

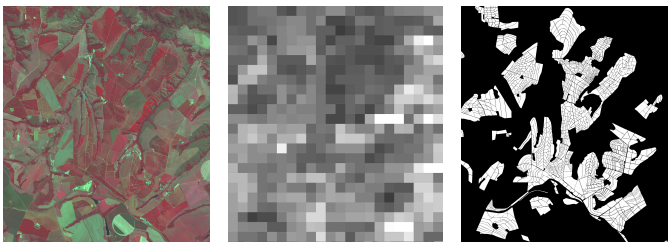
The main research challenge in this application is related to the large number of patterns that plantations can take. This effect on a coffee plantation is illustrated in Figure 4. The difference in age of the stands, the different types of management of plantations and the distortions caused by irregular relief are the main causes of these patterns.



(a) Coffee (b) Entire Image (c) Non-Coffee

Fig. 4. Intravariance class challenge in Dataset - Coffee Crop Recognition.

In order to perform a supervised learning, the dataset contains a training map which includes two different classes: coffee and non-coffee. Tables IV and V show the main information about the dataset and the distribution of the classes in pixels. An illustration of the Coffee Crop dataset is showed in Figure 5.



(a) SPOT VHS (b) MODIS Multispectral (c) Ground Truth

Fig. 5. Dataset - Coffee Crop Recognition.

TABLE IV
GENERAL INFORMATION - COFFEE CROP RECOGNITION.

	Visible Image	Multispectral Image
Width	3721 pixel	22 pixel
Height	4234 pixel	24 pixel
Spatial Resolution	5 m	500 m
Bands	3	7

3) Setup: We have used the two datasets described in subsection IV-A, evaluating one for an urban land-cover multiclass problem, and other for the coffee crop recognition binary problem.

TABLE V
DISTRIBUTION OF PIXEL PER CLASS - COFFEE CROP RECOGNITION.

Classes	Percentual (%)
Coffee	28,86
Non-Coffee	71,14

As **evaluation metric**, we used the Overall Accuracy and Cohen's Kappa metrics. For the statistical test of significance, we used paired Student t-test (confidence of 95%).

Segmentation. We used the IFT-Watershed [35], with parameters of spatial radius 10 and volume threshold equal to 100. Since we are dealing with large images, the IFT-Watergray is effective, efficient and capable of segmenting multiple objects in almost linear time.

Feature Extraction. In the urban dataset, we used four image descriptors to encode spatial information: BIC ([36]), CCV ([37]), GCH ([38]), and Unser ([39]), while in the coffee crop dataset, we used the four descriptors before mentioned jointly with ACC ([40]). In order to extract spectral information in urban dataset, we have used four different approaches: (1) the raw data of HS image (84 Bands), (2) the Fisher Linear Discriminant (FLD) components, (3) the first three principal components of PCA, and (4) the first four PCA components, while in the coffee crop dataset, we withdraw the FLD due to the low spectral resolution.

Training. In the urban dataset, we used the entire training set and fit a group of six weak learners: Gaussian Naive Bayes, k-Nearest Neighbors (3, 5 and 10-Nearest Neighbors), Decision Tree (DT), and a SVM with linear kernel, using the features extracted by each descriptors, resulting in the total of 48 classifiers (24 from each domain). We choose to use a training phase composed by 500 weak learners to construct the strong classifier since we are dealing with a multiclass dataset and the iterations necessary to the SAMME stabilization are commonly higher.

At the coffee crop dataset, the entire image was split in a Stratified ShuffleSplit 5-cross validation scheme, using a group of three weak learners: Multinomial Naive Bayes, one-level Decision Tree, and a SVM with linear kernel, using the features extracted by each descriptors, resulting in the total of 30 classifiers (15 from each domain). We chose to use a training phase composed by 200 weak learners to construct the strong classifier since we are dealing with a binary dataset and the iterations necessary to the SAMME stabilization are sufficient.

We have used the implementation of the learning methods available in the Scikit-Learn Python library. All learning methods were used with default parameters, which means we did not optimize them whatsoever. The management of HS data is made using the Spectral Python (SPy) Library, including the extraction of features from the spectral domain. In this work, we used the nearest neighbor assignment to map the regions extracted from a I_{VHS} to a set of pixels signatures in I_{HS}. We also used the up-sample nearest neighbor assignment method to analyze thematic maps created only using the spectral

features, in the same size of the spatial domain.

Baselines. We have implemented the boosting approach method using different settings as baselines, regarding the dataset. In the Urban Dataset, we used as baseline a diversity-based fusion framework [41] adaptation to encode spatial and spectral feature, and a late fusion framework for RSIs multi-modal classification [42] ; In the Crop Coffee Dataset, we used the following baselines: (1) the boosting scheme using only the BIC descriptor from the spatial domain and the decision tree (DT) as weak classifier, resulting in a traditional Adaboost algorithm; (2) the boosting scheme with all descriptors from the spatial domain and the decision tree as weak classifier; (3) The same settings of (2), including the spectral features; (4) the boosting scheme using only the BIC descriptor from the spatial domain but using all the weak classifiers (CLFs); (5) The same settings of (4), but using all the features from spatial and spectral domains.

B. Results and Discussion

The results obtained by the proposed method (Boosting Approach) against the baselines, with the confidence intervals (95%), are presented in Tables VI (Urban Dataset) and VII (Crop Coffee Dataset).

TABLE VI
ACCURACY AND KAPPA INDEX IN URBAN DATASET

Techniques used	Accuracy (%)	Kappa
Spatial+Spectral (Faria's Paralel) [41]	83,34 +- 0,26	74,73 +- 0,39
Dynamic Majority Vote [42]	84,96 +- 0,48	77,19 +- 0,65
Boosting-based Approach	**88,02 +- 0,27**	**81,67 +- 0,42**

The comparison shows a statistically significant difference among the boosting approach and the baseline proposed, regarding with the t-student test, in both datasets.

Our boosting approach is based on the combination of different features and learner algorithms, letting the feature selection of the boosting decide how to incorporate the weak learners to create a strong model.

In the Urban Dataset, as showed in Table VI, the boosting-based approach outperformed the baselines proposed in the two metrics, showing the effectiveness of the method to handle with data from different sensors.

In the Crop Coffee Dataset, as showed in Table VII, the boosting-based approach outperformed the baselines proposed only in the Kappa index metric, showing the partial effectiveness of the method to handle with data from different sensors.

Even the boosting-based method handled well with the urban dataset, in the coffee crop dataset, the injection of an additional spectral information did not help the model to improve the final results in respect to the accuracy metric. In this case, we can see a huge difference between the pixel resolution of the images (5 m to 500 m), and a poor spectral information (7 bands), preventing the fusion model to get better results. In addition, even the spectral information does not help the model, the use of diverse learning methods and descriptors at spatial domain showed the complementarity of information, improving the final results.

TABLE VII
ACCURACY AND KAPPA INDEX IN CROP COFFEE

Techniques used	Accuracy (%)	Kappa (%)
Boost(DT) + BIC	79,73 +- 1,86	47,34 +- 2,31
Boost(DT) + spatial	82,66 +- 3,14	**56,44 +- 6,97**
Boost(DT) +spatial+spectral	82,83 +- 3,07	**57,14 +- 6,68**
Boost(CLFs) + BIC	81,83 +- 1,91	**53,78 +- 2,53**
Boost(CLFs) +spatial+spectral	82,57 +- 5,15	56,84 +-12,24

As shown in the literature, the use of multiple modalities does not always lead to improvements with respect to the use of a single mode [5]. This usually happens when the considered data that is not relevant for the application could pollute the analysis. In the proposed dataset the low size of spectral data (22 x 24 pixels), and a poor spectral resolution (7 bands) made unfeasible the extraction and combination of spectral information effectively.

V. CONCLUSION

This work addressed the use of RSI particularly with respect to the creation of thematic maps exploiting multi-sensor data. We dealt with two main challenges: the combination of referenced images from different domains (spatial and spectral) and how to exploit different types of features, extracted from these sensors. To this purpose, we proposed a boosting-based approach to the classification task, projected to receive two images, over the same geographic region.

We evaluated the Boosting-based approach in an urban scenario and coffee crop recognition conducting a series of experiments in the datasets that demonstrated a significant improvement using the Kappa index and Overall Accuracy metrics in comparison with the proposed baselines at the urban scenario, but not statistically relevant in concern to the coffee crop recognition dataset using the Overall Accuracy.

The joint complementarity exploitation of different remote sensing sources has proven to be very fruitful in the urban scenario dataset proposed, however the efforts to combine the information from a multispectral data with great difference of spatial resolutions and a poor spectral resolution prevent the Boosting-based approach to utilize the spectral features as additional information for the strong model.

As future work, we plan to evaluate the Boosting-based approach using other remote sensing datasets which contain spatial and hyperspectral information, and covering a higher spatial area. As another future work, we intend to adapt the framework to handle with different sensors, e.g., LIDAR, which contain elevation information from the objects.

ACKNOWLEDGMENT

The authors would like to thank the financing institute CNPq, CAPES, Fapemig. We are also grateful to Cooxupé and Rubens Lamparelli due to the support related to agricultural aspects and the remote sensing dataset.

REFERENCES

[1] A. Ghiyamat and H. Z. Shafri, "A review on hyperspectral remote sensing for homogeneous and heterogeneous forest biodiversity assessment," *International Journal of Remote Sensing*, vol. 31, no. 7, pp. 1837–1856, 2010.

[2] T. Schmid, M. Koch, and J. Gumuzzio, "Multisensor approach to determine changes of wetland characteristics in semiarid environments (central spain)," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 11, pp. 2516–2525, 2005.

[3] Y. Lanthier, A. Bannari, D. Haboudane, J. R. Miller, and N. Tremblay, "Hyperspectral data segmentation and classification in precision agriculture: A multi-scale analysis," *Geoscience and Remote Sensing Symposium, IEEE International*, vol. 2, pp. II–585, 2008.

[4] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 103, no. 9, pp. 1560–1584, 2015.

[5] M. D. Mura, S. Prasad, F. Pacifici, P. Gamba, and J. Chanussot, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1585–1601, 2015.

[6] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[7] I. R. Farah, W. Boulila, K. S. Ettabaa, and M. B. Ahmed, "Multiapproach system based on fusion of multispectral images for land-cover classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 12, pp. 4153–4161, 2008.

[8] S. Delalieux, P. J. Zarco-Tejada, L. Tits, M. A. Jimenez Bello, D. S. Intrigliolo, and B. Somers, "Unmixing-based fusion of hyperspatial and hyperspectral airborne imagery for early detection of vegetation stress," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 2571–2582, 2014.

[9] B. Huang, H. Song, H. Cui, J. Peng, and Z. Xu, "Spatial and spectral image fusion using sparse matrix factorization," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 52, no. 3, pp. 1693–1704, 2014.

[10] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 8, pp. 2207–2218, 2006.

[11] C. M. Gevaert and F. J. García-Haro, "A comparison of starfm and an unmixing-based algorithm for landsat and modis data fusion," *Remote Sensing of Environment*, vol. 156, pp. 34–44, 2015.

[12] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, "Unmixing-based multisensor multiresolution image fusion," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 3, pp. 1212–1226, 1999.

[13] R. Zurita-Milla, J. G. Clevers, and M. E. Schaepman, "Unmixing-based landsat tm and meris fr data fusion," *Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 453–457, 2008.

[14] J. Amorós-López, L. Gómez-Chova, L. Alonso, L. Guanter, J. Moreno, and G. Camps-Valls, "Regularized multiresolution spatial unmixing for envisat/meris and landsat/tm image fusion," *Geoscience and Remote Sensing Letters*, vol. 8, no. 5, pp. 844–848, 2011.

[15] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 1844–1868, 2014.

[16] C. Lanaras, E. Baltsavias, and K. Schindler, "Advances in hyperspectral and multispectral image fusion and spectral unmixing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 40, no. 3, p. 451, 2015.

[17] J. Marcello, A. Medina, and F. Eugenio, "Evaluation of spatial and spectral effectiveness of pixel-level fusion techniques," *Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 432–436, 2013.

[18] J. Zhang, "Multi-source remote sensing data fusion: status and trends," *International Journal of Image and Data Fusion*, vol. 1, no. 1, pp. 5–24, 2010.

[19] I. Amro, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, "A survey of classical methods and new trends in pansharpening of multispectral images." *Journal on Advances in Signal Processing*, vol. 2011, p. 79, 2011.

[20] R. Gharbia, A. T. Azar, A. E. Baz, and A. E. Hassanien, "Image fusion techniques in remote sensing," *arXiv preprint arXiv:1403.5473*, 2014.

[21] G. Camps-Valls and L. Bruzzone, *Kernel methods for remote sensing data analysis*. John Wiley & Sons, 2009.

[22] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 93–97, 2006.

[23] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[24] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, "Multisource composite kernels for urban-image classification," *Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 88–92, 2010.

[25] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 10, pp. 3780–3791, 2010.

[26] L. Gómez-Chova, G. C. Valls, L. Bruzzone, and J. C. Maravilla, "Mean map kernel methods for semisupervised cloud classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 1, pp. 207–220, 2010.

[27] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multitemporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 107, pp. 50–63, 2015.

[28] W. Li, S. Prasad, and J. E. Fowler, "Decision fusion in kernel-induced spaces for hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 52, no. 6, pp. 3399–3411, 2014.

[29] G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, C. Thiel, J. Inglada, E. Christophe, J. Chanussot *et al.*, "Decision fusion for the classification of hyperspectral data: Outcome of the 2008 grs-s data fusion contest," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 11, pp. 3857–3865, 2009.

[30] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot *et al.*, "Multi-modal change detection, application to the detection of flooded areas: outcome of the 2009–2010 data fusion contest," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 1, pp. 331–342, 2012.

[31] J. Wang, C. Li, and P. Gong, "Adaptively weighted decision fusion in 30 m land-cover mapping with landsat and modis data," *International Journal of Remote Sensing*, vol. 36, no. 14, pp. 3659–3674, 2015.

[32] W. Liao, R. Bellens, A. Pizurica, S. Gautama, and W. Philips, "Combining feature fusion and decision fusion for classification of hyperspectral and lidar data," *Geoscience and Remote Sensing Symposium, IEEE International*, pp. 1241–1244, 2014.

[33] J. A. dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. d. S. Torres, and A. X. Falcao, "Multiscale classification of remote sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 10, pp. 3764–3775, 2012.

[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition*, vol. 1, pp. I–511, 2001.

[35] R. Lotufo, A. X. Falcão, F. Zampirolli *et al.*, "Ift-watershed from gray-scale marker," *Computer Graphics and Image Processing*, pp. 146–152, 2002.

[36] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 102–109.

[37] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the fourth ACM international conference on Multimedia*, 1997, pp. 65–73.

[38] M. J. Swain and D. H. Ballard, "Color indexing," *International Conference on Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[39] M. Unser, "Sum and difference histograms for texture classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 1, pp. 118–125, 1986.

[40] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," *Computer Vision and Pattern Recognition*, pp. 762–768, 1997.

[41] F. A. Faria, J. A. Dos Santos, A. Rocha, and R. da S Torres, "A framework for selection and fusion of pattern classifiers in multimedia recognition," *Pattern Recognition Letters*, vol. 39, pp. 52–64, 2014.

[42] E. F. de Andrade Jr, A. de Albuquerque Araújo, and J. A. dos Santos, "A multiclass approach for land-cover mapping by using multiple data sensors," *Iberoamerican Congress on Pattern Recognition*, pp. 59–66, 2015.