Article

Hyperspectral Image Dimensionality Reduction via Maximum Information Tensor Bands Selection for Classification with Convolutional Neural Networks

Josué López [1,*], Deni Torres [1], Clement Atzberger [2], Andrea González [1] and Israel Yañez [3]

[1] Center for Research and Advanced Studies of the National Polytechnic Institute, Telecommunications Group, Av del Bosque 1145, Zapopan 45017, Mexico; deni.torres@cinvestav.mx; andrea.gonzalez@cinvestav.mx
[2] University of Natural Resources and Life Science, Institute of Geomatics, Peter Jordan 82, Vienna 1180, Austria; clement.atzberger@boku.ac.at
[3] Polytechnic University Juventino Rosas, Networks and Telecommunications Ingenering Department, Miguel Hidalgo 102, Comunidad de Valencia 38253 Juventino Rosas, Guanajuato, Mexico; jyanezptc@upjr.edu.mx
* Correspondence: josue.lopez@cinvestav.mx

Version February 4, 2021 submitted to Remote Sens.

Abstract: Tensor-based decomposition for compression of high dimensionality datasets have been widely used in recent years in several research areas, including Multi- and Hyperspectral Images (MSI and HSI) processing. On the other hand, Convolutional Neural Networks (CNNs) are specialized kind of Artificial Neural Networks (ANNs) for processing data that has a known grid-like topology and belongs to the set of natural numbers, such as image data. Compression of the input data of a CNN induce poor performance in tasks, such as classification and semantic segmentation. In this paper, Tucker-based models are employed to reduce the dimensionality in the spectral domain of MSIs and HSIs to reduce the complexity of a pixel-wise classification CNN, while preserving high performance. We propose a framework, based on information theory, that performs a characterization of a spectral dataset, by computing the entropy and the probability distribution function of the spectral bands, as well as a quantification of orthogonality and divergence of the compressed data, to define the dimensionality of the CNN input tensor. Besides, we propose an alternative Tucker approximation with non-negativity and integer constraints called Integer Approximation Non-negative Tucker Decomposition (IANTD). Experimental results demostrate...

Keywords: entropy; hyperspectral imagery; tensor decomposition

---

1. Introduction

Dimensionality reduction of dataset for machine learning algorithms has been one of the most active research areas in recent years []. The introduction of tensor-based models for these kind of tasks inspired a change in several areas, such as image processing [].

Particularly, Remote Sensing (RS) image processing is focused on detecting and monitoring physical features about areas of interest, by analyzing the reflectance of materials over the earth surface. Some remote sensors use filters to separate the reflectance of an object in different wavelength ranges []. These sensors generate the well-known Multi- (MSI) and Hyper-spectral Images (HSI), which lead to high performance in image processing tasks such as detection, classification and segmentation []. Besides, in the last few years the use of spectral data has grown exponentially in other fields as medical analysis [], biomedical [], and, in RS fields as agriculture [], natural disaster prevention [], security affairs [], among others [].

In the last decade, many supervised classification and segmentation algorithms were developed, with the aim of taking advantage of the spatial and spectral data features provided by RS MSI and HSI. Support Vector Machines (SVM) [], k-Nearest Neighbors (k-NN) [] and Convolutional Neural Networks (CNN) [] are examples of the aforementioned. Spectral image processing in artificial intelligence algorithms increase drastically the execution time [], which forces to have robust computer equipment to achieve time competitive results.

Within the aim of reducing high-dimensionality of spectral images, some authors developed dimensionality reduction strategies, by selecting the most salient spectral bands [], and by maximum information and minimum redundancy criteria, based on entropy and mutual information metrics []. These approaches have the advantage of preserving the original domain of MSIs and HSIs. Other works opted for matrix factorization methods, such as Principal Component Analysis (PCA) [] and Singular Values Decomposition (SVD) []. Recently, tensor-based factorization algorithms have proven to be advantageous over those based on matrices []. Nevertheless, changing the input data domain of a machine learning model could lead to a drop in its performance [].

In this work, we propose an innovative spectral imagery dimensionality reduction method from the perspective of maximum information analysis, in which the spectral signatures are carried to a new domain through tensor decompositions. This has the aim of achieving high performance in pixel-wise classification CNNs with low dimensional data. The proposed method can be seen as a three stages process. In the first stage, the original spectral image is factorized, through a non-negative Tucker-based model, transforming the spectral signatures from pixel reflectance domain into *tensor signatures*. In the second stage, a band selection is developed based on an entropy criterion to reduce data dimensionality. Lastly, the compressed tensor is used as input to a pixel-wise classification CNN meant to keep high classification performance. The experimental results on public available dataset show that the proposed framework decrease considerably the computational complexity of the classification CNN with a 10x speedup in testing execution time. Besides, this approach achieves competitive performance, $\pm 2.1\%$ Pixel Accuracy (PA), with lower data dimensionality compared with previous works.

1.1. Previous works

In recent years, several researchers have developed methods to reduce computational complexity of machine learning algorithms [], specially, for HSI classification with Deep Learning (DL) Artificial Neural Networks (ANNs) []. The crucial factor addressed in this work is, to achieve MSI and HSI compression reducing the high computational complexity, without decreasing classification performance in CNNs.

The first methods used for HSI compression was band selection. Li et al. in [23] proposed a band selection method from the perspective of spectral shape similarity analysis. Saliency of spectral bands was another popular approach. Wang Q. et al. [22] proposed to eliminate the drawbacks of traditional salient band selection methods by manifold ranking. More recently, P. Wang et al [] introduced image fusion for feature reduction with joint sparsity model. Besides, other researchers focused their efforts in compressing HSI spectral bands from an information theory point of view. Arecent example is Tschannerl et al. [], who proposed a band selection algorithm following the Maximum-Information-Minimum-Redundancy (MIMR) criterion that maximises the information carried by individual features of a subset and minimizes redundant information between them, as done in [].

Later, matrix decomposition methods were used, such as PCA in [**?**], and even non-negative matrix decomposition methods [**?**]. Nevertheless, matrix-based methods are limited to data representations in 2-dimensional spaces. Spectral imagery have data structures as 3rd-order arrays. This 2-way view produces considerable loss in information, and in turn, in further processing performance. In 2015 Zhang et al. [24] were pioneers in experimenting with multilinear algebra-based decompositions on hyperspectral images.

Table 1. Related work to compression and classification of spectral imagery.

Author & year	Data	Decomposition	Compression	Classifier
Li, S. et al. [23] (2014)	HSI	-	Band selection	SVM
Zhang, L. et al. [24] (2015)	HSI	TKD	Spatial-Spectral	-
Wan, Q. et al. [22] (2016)	HSI	-	Band selection	SVM/kNN/CART
Tong L. et al. [] (2017)	HSI	NMF	Unmixing	-
Chien, J. et al. [] (2017)	RGB	TFNN	Spatial-Spectral	TFNN
Dewa, M. et al. [] (2018)	HSI	PCA	Spectral	PCA
Xu, Y. et al. [] (2018)	HSI	-	-	CNN
Li, J. et al. [28] (2019)	MSI	NTD-CNN	Spatial-spectral	-
An, J. et al. [27] (2019)	HSI	T-MLRD	Spatial-spectral	SVM/1NN
An, J. et al. [29] (2019)	HSI	TDA	Spatial-spectral	SVM/1NN
Sayeh, M. et al [] (2019)	HSI	NTD	Spatial-Spectral	3D-CNN
Lopez, J. et al. [] (2020)	MSI	TKD	Spectral	FCN
Sayeh, M. [] (2020)	HSI	BG-NTD	Spatial-Spectral	MLR
Our framework	**MSI/HSI**	**NTD-1**	**Spectral**	**CNN**

On the other hand, instead of HSI, MSIs was a good alternative due to the small number of spectral bands, which also produce competitive classification performance [11], [18], [21] and [?]. However, the need to increase performance forced researchers to use data with higher number of spectral bands, which ease classification of materials hard to differentiate [26], [27], [29], [?] and [?].

Recently, a work close to our research was published. Sayeh [?] proposed a framework where discriminative features are extracted applying Non-negative Tensor Decomposition (NTD) technique to the image tensor. The factorized components indicate the spectral signatures and 2D abundance maps of the constituent materials. Different to our framework, they compress HSI in the spatial and spectral domain, while our approach preserve the architecture of the image by compressing only the spectral domain and with non-negativity constraints in the core tensor of the TKD. In addition, we introduce information metrics to reinforce the rank estimation method proposed in [] and we propose an integer approximation to take advantages of the kindness of CNNs. Table 1 summarizes some related papers, which deal with the compression-classification issue.

1.2. Motivation

Dimensionality reduction in HSI processing is still a challenging issue. The recent rise of DL-ANN in the image processing area has led researchers to find ways to preserve the spatial-spectral information of HSI in lower dimensionality than the raw data. Several methods have been proposed in the last five years. However, band selection, matrix factorization, as well as tensor-based methods proposed previously perform dimensionality reductions that, to obtain competitive performance, still require more than twenty percent of the data [], which still maintain high computational complexity.

In this work, we address the HSIs dimensionality reduction issue by applying NTD to project original data into a new tensor domain, where the spatial-spectral features are contained in a lower dimensionality tensor than the raw data. With the aim of keeping high classification performance with the compressed data, this work presents an entropy based strategy, to avoid high information loss, while reducing data dimensionality and, in turn, computational load of pixel-wise classification CNNs. This work has three particular motivations: 1) to reduce the computational load of CNNs, 2) to avoid overfitting by reducing redundant information and excess features, and 3) to keep high classification performance.

1.3. Contribution

Unlike previous works, we address the problem of HSI dimensionality reduction by a tensor factorization and entropy approach. The main contribution of this work can be summarized as follows. i) We propose a HSI band selection strategy combined with a tensor decomposition approach, where the original spectral bands are transformed into new *tensor images* by a non-negative Tucker-1 decomposition (NTD-1), and select a set of tensor bands with the highest entropies to be the input of a pixel-wise classification CNN. ii) Besides, this work also presents an analysis, starting from the Linear Mixing Model (LMM), to relate the transformation of the raw data spectral signatures into the new tensor signatures formed in the tensor slices generated by the NTD-1. iii) Lastly, due to high imbalance presented in the most of the available datasets, we present the classification performance evaluation of the framework proposed with metrics considering the impact of class imbalanced.

The remainder of this work is organized as follows. Section 2 introduces tensor algebra notation and basic concepts to familiarize the reader with the symbology used in this paper. Section 3 describes the problem statement and the framework proposed in this work. Experimental results are presented in Section 4. Finally, Sections 5 and 6 present the discussions, comparisons and conclusions based on the experimental results obtained in the cases studied.

2. Preliminaries

As defined formally in [], an Nth-order tensor is an element of the tensor product of N vector spaces, each of which has its own coordinate system. A tensor can be seen as a multidimensional array. The order of a tensor is the number of dimensions, also known as modes, i.e., an N-order tensor $\mathfrak{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is an N-dimensional array, which elements $x_{i_1,i_2,\ldots,i_n}$ are indexed by $i_n \in 1, 2, \ldots, I_n$ for $1 \leq n \leq N$. Throughtout this paper, the mathematical notation used by Kolda et al. [17] has been adopted. Table 2 summarize this notation.

2.1. Tensor decompositions (TDs)

As an extension of the SVD [], two main specific tensor decompositions can be considered, Tucker Decomposition (TKD) [] and CANDECOMP/PARAFAC (CP) []. There are many other tensor decompositions, INDSCAL, PARAFAC2, CANDELINC, DEDICOM, PARATUCK2, among others [17]. Furthermore, there are also nonnegative variants of all of the above. With the aim of preserving particular characteristics of hyperspectral images for pixel-wise classification, this study is limited to use decompositions based on the Tucker model.

2.1.1. Tucker Decomposition (TKD)

For the particular case of third-order tensors, the TKD [17] can be formally formulated as follows [?]. Given a third-order tensor $\mathfrak{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and three positive indices J_1, J_2 and J_3, find a core tensor $\mathfrak{G} \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ and three component matrices called factor matrices $\mathbf{U}^{(1)} \in \mathbb{R}^{I_1 \times J_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{I_2 \times J_2}$ and $\mathbf{U}^{(3)} \in \mathbb{R}^{I_3 \times J_3}$ which perform the following approximate decomposition:

$$\mathfrak{X} = \mathfrak{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} + \mathcal{E} \tag{1}$$

where $\mathcal{E} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ denotes the approximation error. The core tensor $\mathfrak{G}$ preserves the level of interaction for each factor or projection matrix $\mathbf{U}^{(n)}$. The factor matrices are commonly considered orthogonal, but in Tucker models with non-negativity constraints, that is not necessarily imposed [?]. These matrices can be seen as the principal components in each mode [17] (see Figure 1). J_n represents the number of components in the decomposition, i.e., the rank $-(R_1, R_2, R_3)$. The core tensor is computed by the multilinear projection

$$\mathfrak{G} = \mathfrak{X} \times_1 \mathbf{U}^{(1)\mathrm{T}} \times_2 \mathbf{U}^{(2)\mathrm{T}} \times_3 \mathbf{U}^{(3)\mathrm{T}} \tag{2}$$

where $\mathbf{U}^{(n)\mathrm{T}}$ denotes the transpose matrix of $\mathbf{U}^{(n)}$ for $n = 1, \ldots, N$. Hence, the approximation $\hat{\mathfrak{X}}$ of the tensor decomposition is given by

$$\hat{\mathfrak{X}} = \mathfrak{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}. \tag{3}$$

Table 2. Tensor algebra notation summary

$\boldsymbol{\mathcal{A}}, \mathbf{A}, \mathbf{a}, a$	Tensor, matrix, vector and scalar respectively
$\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$	N-order tensor of size $I_1 \times \cdots \times I_N$.
$a_{i_1 \ldots i_N}$	An element of a tensor
$\mathbf{a}_{:i_2 i_3}$, $\mathbf{a}_{i_1 : i_3}$, and $\mathbf{a}_{i_1 i_2 :}$	Column, row and tube fibers of the third order tensor $\boldsymbol{\mathcal{A}}$
$\mathbf{A}_{i_1 ::}$, $\mathbf{A}_{:i_2 :}$, $\mathbf{A}_{::i_3}$	Horizontal, lateral and frontal slices of the third order tensor $\boldsymbol{\mathcal{A}}$
$\mathbf{A}^{(n)}$, $\mathbf{a}^{(n)}$	A matrix/vector element from a sequence of matrices/vectors
$\mathbf{A}_{(n)}$	Mode-n matricization of a tensor. $\mathbf{A}_{(n)} \in \mathbb{R}^{I_n \times \prod_{m \neq n} I_m}$
$\mathbf{a}^{(\mathbf{1})} \circ \cdots \circ \mathbf{a}^{(\mathbf{N})}$	Outer product of N vectors
$\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle$	Inner product of two tensors
$\boldsymbol{\mathcal{A}} \times_n \mathbf{U}$	n-mode product of tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J \times I_n}$ along axis n
$\boldsymbol{\mathcal{A}} * \mathbf{U}$	Tensor / matrix Hadamard product
$\text{rank}_n(\boldsymbol{\mathfrak{X}})$	column rank of $\mathbf{X}_{(n)}$. If $R_n \equiv \text{rank}_n(\boldsymbol{\mathfrak{X}})$, then $\boldsymbol{\mathfrak{X}}$ has a rank $-(R_1, \ldots, R_N)$ tensor

and the reconstruction error ξ can be computed by the Mean Square Error (MSE) given by

$$\xi(\hat{\boldsymbol{\mathfrak{X}}}) = ||\boldsymbol{\mathfrak{X}} - \hat{\boldsymbol{\mathfrak{X}}}||_F^2 \tag{4}$$

and $|| \cdot ||_F^2$ represents the Frobenius norm. To compute the $\text{rank}_n(\boldsymbol{\mathfrak{X}})$ approximation of a tensor, it can be used iterative algorithm such as ALS, HALS or HOOI, commonly using HOSVD initialization, minimizing the cost function given in equation 4 [**?**].

2.1.2. Non-negative Tucker Decomposition (NTD)

The NTD is a decomposition based on the Tucker model. It is a tensor factorization method with non-negativity constraints []. For the third-order case, the NTD, as defined by Cichocky [15], can be formulated as follows. Given a third-order tensor $\boldsymbol{\mathfrak{X}} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$ find a core tensor $\boldsymbol{\mathfrak{G}} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$ and the factor matrices $\mathbf{U}^{(1)} \in \mathbb{R}_+^{I_1 \times J_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}_+^{I_2 \times J_2}$ and $\mathbf{U}^{(3)} \in \mathbb{R}_+^{I_3 \times J_3}$ which performs the approximation given in Eq. (1), minimizing the cost function given in equation 4 by an iterative algorithm.

Non-negativity constraints lead minimization problem to converge to a non optimal local minima []. Yong et al. [] first proposed NTD and developed multiplicative updating algorithms for learning a Tucker decomposition of a nonnegative tensor with restricting a core tensor and mode matrices to be nonnegative. Projection matrices and core tensor updating rule can be written as

$$\mathbf{U}^{(n)} \leftarrow \mathbf{U}^{(n)} * \frac{\mathbf{X}_{(n)} \mathbf{G}_U^{(n)\mathrm{T}}}{\mathbf{U}^{(n)} \mathbf{G}_U^{(n)} \mathbf{G}_U^{(n)\mathrm{T}}} \tag{5a}$$

$$\boldsymbol{\mathfrak{G}} \leftarrow \boldsymbol{\mathfrak{G}} * \frac{\boldsymbol{\mathfrak{X}} \times_1 \mathbf{U}^{(1)\mathrm{T}} \times_2 \mathbf{U}^{(2)\mathrm{T}} \times_3 \mathbf{U}^{(3)\mathrm{T}}}{\boldsymbol{\mathfrak{G}} \times_1 \mathbf{U}^{(1)\mathrm{T}} \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)\mathrm{T}} \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)\mathrm{T}} \mathbf{U}^{(3)}} \tag{5b}$$

where $\mathbf{G}_U^{(n)} = \left[\boldsymbol{\mathfrak{G}} \times_{m \neq n} \mathbf{U}^{(m)} \right]_{(n)}$ denotes the encoding variable. This algorithm shows efficient performance and some advantages comparing with other algorithms proposed for computing NTD, such as local ALS, HALS, Alpha, Beta, HOOI, etc [].

3. Proposed framework

3.1. Problem Statement

Let $\boldsymbol{\mathfrak{X}} \in \mathbb{N}_0^{I_1 \times I_2 \times I_3}$ be a spectral image represented as a third-order tensor, and $\mathbf{Y} \in \mathbb{C}^{I_1 \times I_2}$ its corresponding ground truth, where $\mathbb{C}$ denotes the set of classes of interest. Find a tensor $\hat{\boldsymbol{\mathfrak{G}}} \in \mathbb{R}^{I_1 \times I_2 \times B}$, with lower dimensionality than the raw data, $B < I_3$, to reduce computational load of pixel-wise classification CNNs Θ, while preserving high performance at the prediction output $\hat{\mathbf{Y}}$.

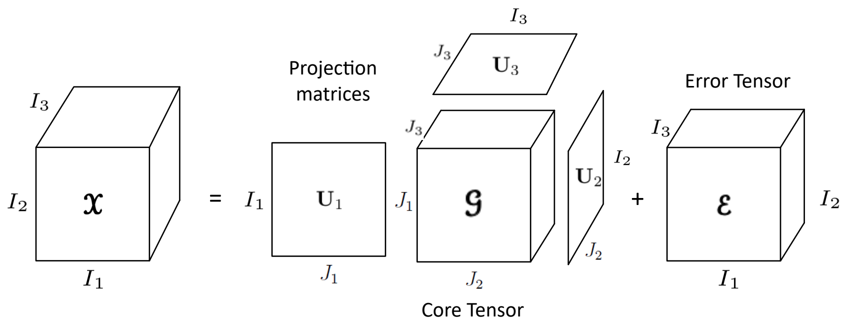


Figure 1. Tucker decomposition for a third-order tensor.

3.2. Mathematical Definition

The problem statement described above can be mathematically defined as the optimization problem

$$
\begin{array}{lll}
\min\limits_{\hat{\mathbf{Y}}} & ||\mathbf{Y}-\hat{\mathbf{Y}}||_F^2 = \min\limits_{\hat{\mathfrak{G}}} ||\mathbf{Y}-\Theta(\hat{\mathfrak{G}})||_F^2 & \\
\text{subject to} & \hat{\mathfrak{G}} \subset \mathfrak{G} & \text{subtensor of the core tensor} \\
 & H(\hat{G}_1) \leq H(\hat{G}_2) \leq \cdots \leq H(\hat{G}_B) & \text{bands sorted in entropy decreasing order} \\
 & D_{JS}(\mathfrak{X}||\hat{\mathfrak{G}}) \leq D_{JS_S} & \text{divergence information measure,} \\
\text{where} & \hat{\mathfrak{G}} \in \mathbb{R}^{J_1\times J_2\times B} & \text{and } B < J_3\text{, which is computed by} \\
 & B = \min(b) | H(G_b) > T\max(\mathbf{h}) & \text{number of selected bands.}
\end{array}
\tag{6}
$$

and the core tensor is found by the NTD-1 mathematically defined as follows

$$
\begin{array}{lll}
\min\limits_{\mathfrak{G},\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)}} & ||\mathfrak{X} - \mathfrak{G}\times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}||_F^2 & \\
\text{subject to} & \mathbf{U}^{(1)} = \mathbf{U}^{(2)} = \mathbf{I} & \text{Tucker-1 model} \\
 & \mathbf{U}^{(3)} \in \mathbb{R}_+^{I_3\times J_3} & \text{non-negative projection matrix,} \\
 & \mathfrak{G} \in \mathbb{R}_+^{J_1\times J_2\times J_3} & \text{non-negative constraints,} \\
 & \text{rank}_n(\mathfrak{X}) = \text{rank}_n(\mathfrak{G}) & J_n = I_n \quad \text{for} \quad n = 1,2,3\text{, and} \\
 & \xi(\hat{\mathfrak{X}}) \leq \xi_s & \text{representativity measure.}
\end{array}
\tag{7}
$$

The following subsections describe the big picture of the framework proposed in this work, which is summarized in three steps: tensor decomposition, band selection and classification.

3.3. Methodology

Given a spectral image $\mathfrak{X} \in \mathbb{N}_0^{I_1\times I_2\times I_3}$, where I_1, I_2 represents its spatial dimensionality, I_3 the number of spectral bands, and $\mathbb{N}_0$ the set of natural numbers including 0, a 3-mode fiber $\mathbf{x}_k \in \mathbb{R}^{I_3}$ represents the spectral signature at pixel k for $k = 1,2,\ldots,I_1 I_2$, and can be represented by the Linear Mixing Model (LMM) as follows

$$
\mathbf{x}_k = \mathbf{M}\boldsymbol{\alpha}_k + \boldsymbol{\eta} \tag{8}
$$

where $\boldsymbol{\alpha}_k \in \mathbb{R}^C$ represents the abundance vector at pixel k, $\mathbf{M} \in \mathbb{R}^{I_3\times C}$ denotes the endmember matrix, and $\boldsymbol{\eta} \in \mathbb{R}^{I_3}$ an additive noise vector. The abundance vectors $\boldsymbol{\alpha}_k$ must always satisfy two properties: i) the non-negativity, $\alpha_{i_3} \geq 0$ for $i_3 = 1,\ldots,I_3$, and ii) the sum-to-one restriction, $\sum_{i_3=1}^{I_3} \alpha_{i_3} = 1$. Additionally, the nonnegativity property must be satisfied by the endmember matrix as well.

As studied in [], nonnegative decompositions are regarded to be part-based data representations leading the factorization result to fit the requirement of spectral unmixing. If $\mathbf{U}$ fulfill the nonnegativity property, and $\mathbf{g}_k$ satisfy nonnegativity and sum-to-one properties, this decomposition can be seen as a linear spectral unmixing, where $\mathbf{U}$ may be seen as the endmembers matrix and $\mathbf{g}_k$ the contribution vector at pixel $k = 1, 2, \ldots, I_1 I_2$. Nevertheless, non-negative tensor decomposition with additional constraints may lead to non optimal local minima and slower convergence [].

In this work, NTD-1 is used to preserve neighboring pixel correlation and to transform spectral signatures into new *tensor signatures*. This way, classification algorithms may reach high performance with lower number of tensor bands selected. We compute the decomposition setting the projection matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ in Eq. 1 as the identity matrix as

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{I} \times_2 \mathbf{I} \times_3 \mathbf{U} + \mathcal{E} = \mathcal{G} \times_3 \mathbf{U} + \mathcal{E}. \tag{9}$$

The core tensor can be approximated by

$$\mathcal{G} = \mathcal{X} \times_3 \mathbf{U}^{+} \Leftrightarrow \mathbf{G}_{(3)} = \mathbf{U}^{+}\mathbf{X}_{(3)} \tag{10}$$

where $\mathbf{U}^{+} \in \mathbb{R}^{J_3 \times I_3}$ denotes the pseudoinverse matrix. Each 3-mode fiber of $\mathcal{G}$, denoted $\mathbf{g}_k \in \mathbb{R}^{J_3}$ can be computed by

$$\mathbf{g}_k = \mathbf{U}^{+}\mathbf{x}_k \tag{11}$$

Since the LMM is a lineal transformation that maps the endmembers into the pixel space of the HSI, and the TKD $\mathcal{X} \rightarrow \mathcal{G}$ is multilinear but linear in each mode, then for 3-mode

$$\mathbf{g}_k = \mathbf{U}^{+}\mathbf{M}\boldsymbol{\alpha}_k + \mathbf{U}^{+}\boldsymbol{\eta} \tag{12}$$

and this equation can be written as

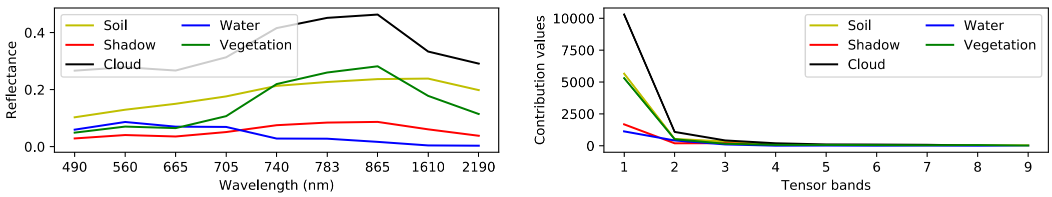
$$\mathbf{g}_k = \mathbf{M}'\boldsymbol{\alpha}_k + \boldsymbol{\eta}' \tag{13}$$

where $\mathbf{M}' = \mathbf{U}^{+}\mathbf{M}$ represents the equivalent endmember matrix in the new tensor bands domain, and $\boldsymbol{\eta}' = \mathbf{U}^{+}\boldsymbol{\eta}$ the additive noise component. In case of $\mathbf{U}$ is orthogonal, $\mathbf{M}' = \mathbf{U}^{\mathrm{T}}\mathbf{M}$ and $\boldsymbol{\eta}' = \mathbf{U}^{\mathrm{T}}\boldsymbol{\eta}$. Figure 2 shows the transformation from the original spectral signatures to the tensor signatures generated by the decomposition.

Using the SVD as initialization of the NTD-1 algorithm, these new tensor signatures generally follows the ordering property stated in [31] as

$$\|\mathcal{G}_{i_3=1}\|_F \geq \|\mathcal{G}_{i_3=2}\|_F \geq \cdots \geq \|\mathcal{G}_{i_3=I_3}\|_F. \tag{14}$$

where $\|\mathcal{G}_{i_3}\|_F$ denotes the Frobenius norm of the I_3 tensor bands. This property can be visualized in Figure 3 as three dimensional heat map of all the core tensor components values. Hence, the 3-mode slices of $\mathcal{G}$ are regarded as new tensor images (see Figure 4), which carry spatial-spectral information from the original image.



(a) **(b)**

Figure 2. Visual comparison between a) Spectral signatures of a single image from the normalized dataset sentinel-2 with five classes of interest and b) tensor signatures after NTD-1.

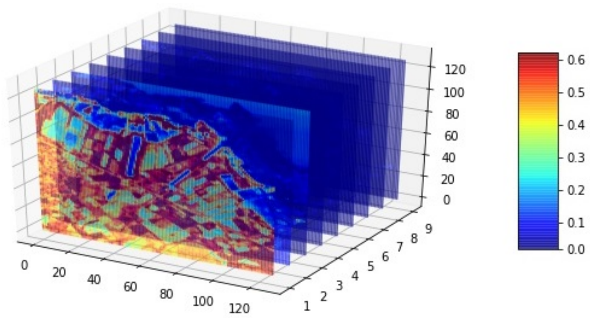
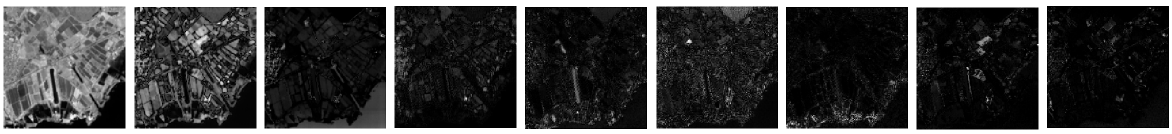


Figure 3. Core tensor heat map applying NTD-1 to a single image from the normalized dataset setinel-2.



(a) **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)** **(i)**

Figure 4. Tensor bands of the NTD-1 applied to a single image from the Sentinel-2 dataset.

3.4. Dimensionality reduction

Under the principle of matrix-based factorization [], the larger eigenvalues in the core tensor are considered to keep more intrinsic information of the raw data []. Considering that each band of the core tensor is an image, it is possible to quantify its uncertainty, by the Shannon entropy H computed as

$$H(G_b) = -\sum_{g \in \Omega} p(g) \log p(g) \tag{15}$$

where G_b denotes the 3-mode slice (tensor band) b of $\mathcal{G}$ as discrete random variable for $b = 1, \ldots, J_3$, with probability space $(\Omega, \Sigma, p(g_b))$, where $\Omega = \{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, 1\}$, n denotes the number of bits needed to store the maximum value of $\mathcal{G}$, $\Sigma = \{\sigma_1, \sigma_2, \ldots\}$ and $\sigma_n \subset \Omega$, and probability mass function $p(g_b) = Pr\{G_b = g_b\}$.

Let $\mathbf{h}$ be a vector of the entropies of each tensor band sorted in decreasing order, the number of tensor bands selected B to be the input of the CNN is determined by

$$B = \min(b) | H(G_b) > T \max(\mathbf{h}) \tag{16}$$

where T denotes a given threshold value. From Eq. 16, the number of selected bands depends on the entropy of the given tensor. The smaller the value of T, the larger the number of selected bands and, in turn, the more detailed features inputs the CNN. Figure 5 shows that each spectral band of the original

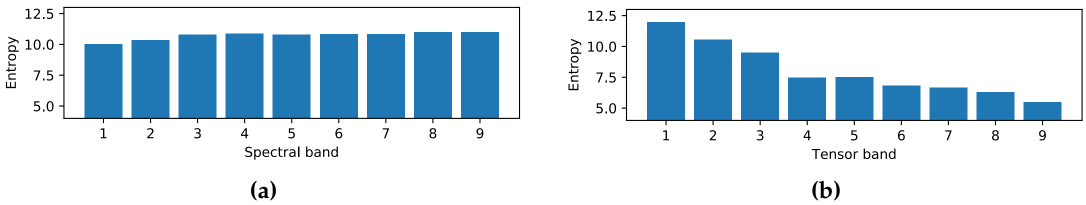


Figure 5. Entropy of Sentinel-2 dataset, a) raw data, and b) NTD-1 core tensor core tensors.

Sentinel-2 data have relative high entropy values (close to ten), while NTD-1 tensor images have this high values only in the first three bands.

This band selection process generates a lower dimensional tensor $\hat{\mathbf{\mathfrak{G}}} \in \mathbb{R}^{J_1 \times J_2 \times B}$ with $B < I_3$ tensor bands selected. We use the Jensen Shanon Divergence (JSD) as information metric to measure how representative from the original data is $\hat{\mathbf{\mathfrak{G}}}$, and is computed as

$$D_{JS}(\mathcal{X}\|\hat{\mathcal{G}}) = \frac{1}{2}D_{KL}(\mathcal{X}\|\mathcal{M}) + \frac{1}{2}D_{KL}(\hat{\mathcal{G}}\|\mathcal{M}) \tag{17}$$

where $D_{JS}(\mathcal{X}\|\hat{\mathcal{G}})$ represents the JSD between the probability distributions of the raw data $\mathcal{X}$ and the reduced tensor $\hat{\mathcal{G}}$, $\mathcal{M} = \frac{\mathcal{X}+\hat{\mathcal{G}}}{2}$ is the mean of the two probability distributions, and $D_{KL}(\cdot)$ denotes the Kullback-Leibler divergence, which is a asymmetric version of the JSD and it is computed as

$$D_{KL}(\mathcal{X}\|\hat{\mathcal{G}}) = \sum_{i=1}^{I} p(x_i) \log \frac{p(x_i)}{p(\hat{g}_i)} \tag{18}$$

where $p(x_i)$ and $p(\hat{g}_i)$ represent the probability of the i-th element at distributions $\mathcal{X}$ and $\hat{\mathcal{G}}$ respectively. The use of this metric has the purpose of controlling high information loss rate. Large divergences drive to not optimal classification performance. Due to non negative input data, non negative decompositions present lower divergence between raw data and core tensor than not restricted decomposition. Besides, preserving spatial properties, as done by the NTD-1 used in this work, reach high similitude with the original data. Figure 6 compares normal TKD-1 and NTD-1 decomposition for variable $\text{rank}_n(\mathbf{\mathfrak{G}})$.

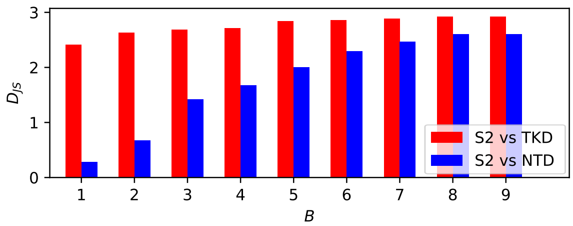


Figure 6. JSD between core tensor and raw data.

3.5. Pixel-wise classification CNN

Let $\hat{\mathbf{\mathfrak{G}}} \in \mathbb{R}^{I_1 \times I_2 \times B}$ be the lower dimensionality core tensor, with B selected bands, and $\mathbf{Y} \in \mathbb{C}^{I_1 \times I_2}$ its corresponding ground truth, where $\mathbb{C}$ denotes the set of C different classes. $\hat{\mathbf{\mathfrak{G}}}$ and $\mathbf{Y}$ form the input tuple to the CNN classifier denoted as Θ, which produce a prediction matrix $\hat{\mathbf{Y}} \in \mathbb{C}^{I_1 \times I_2}$, i.e.,

$$(\hat{\mathbf{\mathfrak{G}}}, \mathbf{Y}) \xrightarrow{\Theta} \hat{\mathbf{Y}}. \tag{19}$$

Generally, CNNs are composed by a set of convolutional, Rectified Linear Unit (ReLU) and pooling/unpooling layers. The ReLU activation function generates activation maps, which identify features of a specific class in the image. At last layer, the activation maps are introduced to a softmax function, which output a probability distribution tensor $\mathbf{\mathfrak{P}}$ over C different classes. Let $\mathbf{\mathfrak{Z}} \in \mathbb{R}^{I_1 \times I_2 \times C}$ be a tensor with the set of activation maps at last layer and $\mathbf{z}_k$ a 3rd-mode fiber of $\mathbf{\mathfrak{Z}}$, then

$$\gamma_k(\mathbf{z}) = \frac{e^{z_k}}{\sum_{c=1}^{C} e^{z_c}} \tag{20}$$

where $\gamma_k(\cdot)$ denotes the softmax function and k the element of the output vector. Each fiber $\mathbf{p}_{i_1 i_2:}$ is the predicted probability distribution at pixel i_1, i_2, wich has a wide relation with the contribution parameter $\boldsymbol{\alpha}_{i_1 i_2:}$ in 13. The highest probability or contribution value indicates the truth or predicted class respectively. Figure 7 shows the big picture summarizing the framework proposed.

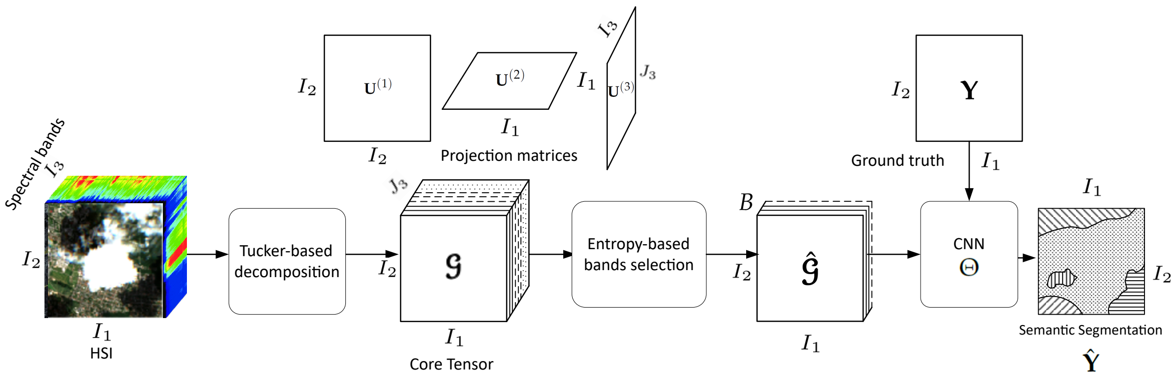


Figure 7. Big picture of the framework proposed.

4. Experimental Results

Our framework was implemented in Python code using the open source machine learning library for tensor learning Tensorly []. The CNN classification architecture used to evaluate the framework proposed in this work is Segnet [], implemented using tensorflow library []. Table 3 shows the hyperparameters of the CNN set by cross-validation and software / hardware specifications.

Table 3. Experiments' software and hardware specifications.

Hyperparameters	Software/Hardware
learning rate: 1×10^{-3}	Platform: Python 3.7
epochs: 100	AI Framework: Tensorflow 1.13
optimizer: Adam []	GPU: NVIDIA GeForce GTX 1050 Ti
initialization: Xavier []	Processor: Intel core i7
kernel dimensions: 3×3	RAM: 8GB
Activation Function: ReLU / Softmax	SSD: 128GB / HDD: 1TB

4.1. Classification metrics

On the other hand, the datasets used for experiments in this work are strongly imbalanced. For this reason, we selected the metrics based on the works of Luque et al. [], Chicco et al. [] and Grandini et al. [], where they assess the impact of the imbalance and propose a set of metrics with lower bias in function of the imbalance and depending on the confusion matrix.

In order to fair evaluate performance of the classifier and to compare our work with others of the state of the art, we selected three main performance evaluation metrics. Pixel Accuracy (PA) is used to compute a ratio between the amount of correctly classified pixels and the total number of pixels. Despite this metric is hihgly biased for multiclass imbalanced dataset, it is one of the most popularly used in the state-of-art. Given a confusion matrix $\mathbf{M} \in \mathbb{N}^{C \times C}$ relating the True Positive (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), PA is computed by Eq. 21. Cohen's Kappa coefficient and Matthews Correlation Coefficient (MCC) are alternative measures less affected by the imbalance issue [**?**]. Kappa and MCC are computed by 25 and 23 respectively.

Additionally, we use PA and MCC considering class imbalance as proposed in []. Each metric μ can be expressed as a function $\mu = \mu(\lambda_{PP}, \lambda_{NN}, \delta)$, where $\lambda_{PP} = \frac{TP}{TP+FN}$, $\lambda_{NN} = \frac{TN}{TN+FP}$ and $\delta = 2\frac{TP+FN}{TP+FN+FP+TN} - 1$.

Table 4. Table of metrics used to evaluate CNN classification performance.

Metric	Formula	Imbalance metric
PA	$\frac{TP+TN}{TP+TN+FP+FN}$ (21)	$\lambda_{PP}\frac{1+\delta}{2} + \lambda_{NN}\frac{1-\delta}{2}$ (22)
MCC	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ (23)	$\frac{1}{2}\left(\frac{\lambda_{PP}+\lambda_{NN}-1}{\sqrt{[\lambda_{PP}+(1-\lambda_{NN})]\frac{1-\delta}{1+\delta}[\lambda_{NN}+(1-\lambda_{PP})\frac{1+\delta}{1-\delta}]}}+1\right)$ (24)
Kappa	$\kappa = \frac{\rho_o - \rho_e}{1-\rho_e}$ (25)	-

4.2. Cases study

For this work, one multispectral dataset and three popular hyperspectral dataset were selected. The datasets were obtained from the European Space Agency Sentinel-2 database and from the Hyperspectral Remote Sensing Scenes web page.

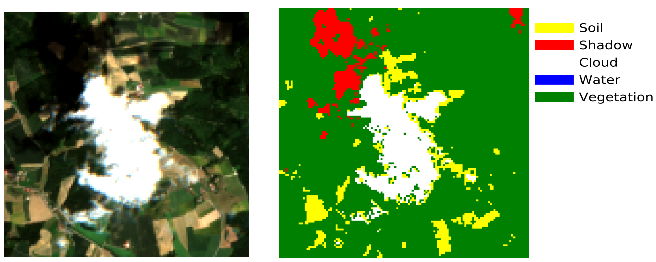
Table 5 summarizes the datasets used in this work, as well as their spatial and spectral characteristics, the number of classes and samples.

Table 5. Summary of the different dataset used for experiments in this work.

Dataset	Spatial dimensions	Bands	Classes	Samples
Sentinel-2 CNNMSI	128×128	9	5	1,802,240
Indian Pines	145×145	220	16	10,249
Salinas	512×217	224	16	53,785
Pavia University	610×340	103	9	40,076

4.2.1. Case A: Sentinel-2 dataset

This dataset proposed by Lopez et al. [?] is composed of 110 RS Sentinel-2 scenarios from central Europe. It has 100 scenarios as the training set and 10 scenarios for testing, all of them with 128×128 pixels with spatial resolution of $20m^2$ and 9 spectral bands in the range $490 - 2190nm$. The labels are semi-manually assigned for five classes of interest: vegetation, soil, water, clouds and shadows. Data are available in the link Sentinel-2 Dataset.



(a) (b)

Class	Dataset	Training	Testing
Soil	420,934	374,792	46,142
Shadow	45,080	42,281	2,799
Cloud	251,692	235,546	16,146
Water	202,753	195,512	7,241
Vegetation	881,781	790,269	91,512

Figure 9 & Table 5. Sentinel-2 dataset a) True color image and b) Ground truth. Table) Samples per class.

In this dataset Sentinel-2 images are in Level-2A ESA product type, which provides images in top-of-atmosphere reflectance integer units. This ease the dataset normalization dividing it by 10000 to obtain reflectance values in 0 to 1 range []. Another important consideration is the dataset imbalance, which can be seen in Table 6. Vegetation and soil are the classes with higher positive imbalance

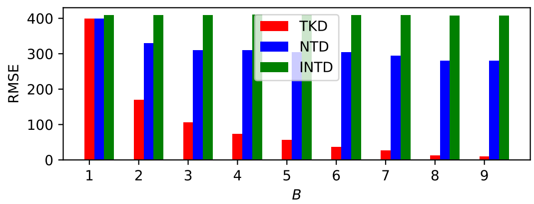


Figure 10. Reconstruction error for variable number of tensor bands.

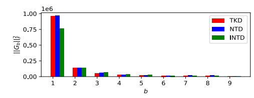


Figure 11. Tensor band Frobenius norm.

encompassing 48.92% and 23.35% of the samples respectively. On the other hand, shadow class has only 2.5% of the whole dataset. Comparing with vegetation and soil, there is a significant imbalance among them. For this reason, performance evaluation metrics used in this work have been selected considering class imbalance.

This dataset can be characterized by information theory metrics. Figure 5 show the entropy level of the raw data spectral bands, where it can be seen that each band has high entropy. On the contrary, tensor bands generated in the Tucker-based decompositions generally carry the highest entropy to the frontal bands and have a decreasing behavior.

The five classes of interest in this dataset show spectral signatures easy to discern, since their reflectance are considerably distanced one from the others in more than three wavelengths, as can be seen in Figure 2. When a Tucker-based decomposition is applied to the raw data, these signatures behaves differently. It is worth noting that, the signatures in the new tensor band domain are easy diferenciable in the first frontal bands, but they are highly correlated after the third one.

The low spectral dimensionality of Sentinel-2 dataset makes simpler the analysis of the distance metrics used in this work. Figure 10 shows the reconstruction error comparing the three decompositions used in this work for variable number of 3rd-mode tensor dimensionality. As it can be expected, the larger the number of bands retained, the lower the reconstruction error. Also, non-negative and integer restricted decompositions produce slower error decrease than the no constrained TKD.

On the other hand, Figure 6 shows that compressed tensors with lower number of tensor bands selected, with respect to the raw data, have higher level of representativity than those with higher dimensionality. In this case, TKD cores look less representative due to negative values, which diverge widely from the positive ones in the raw data. Nevertheless, under Lathauwer criteria for all orthogonality [31], TKD present a much higher orthogonality, which is totally related to the freedom of the decomposition to find solutions in a wider domain. Figure 11 shows the norm criterion, which is generally satisfied in the three core tensors. However, as it can be anticipated, NTD does not fulfilled the inner product criterion, and in turn, projection matrices have low orthogonality degree.

In terms of complexity, the execution time may be a metric to quantify the complexity reduction degree. Figure 12 shows that, as the number of tensor bands selected increases, the execution time rises exponentially.

Finally, Figure 13 shows the PA for the three decomposition varying the number of selected bands. As point of comparison, dotted lines indicates the PA with the raw data as input to the CNN. TKD and IANTD are highly competitive with less than 25% of the original tensor dimensionality, while NTD proves to get better results, even than the raw data, with only 1 tensor band. This can also be seen in Table 6, where it is summarized the performance evaluation under the three metrics selected.

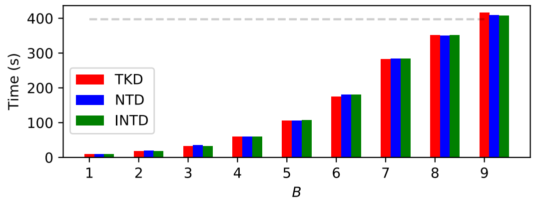


Figure 12. Execution time for variable compressed tensor dimensionality.

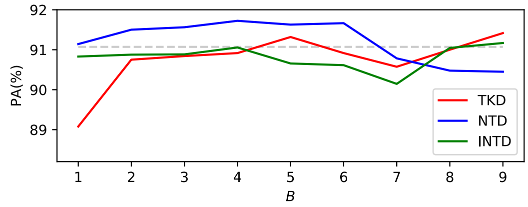


Figure 13. PA for variable number of selected bands.

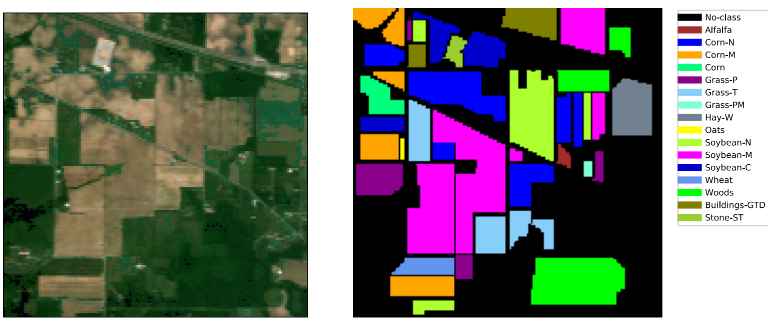
Table 6. Quantitative results[1] for the Sentinel-2 test dataset running in a NVIDIA GeForce GTX 1050 Ti GPU, Intel core i7 processor, 8 Gb RAM, SSD 128 Gb, and HDD 1 Tb. Decomposition recontruction error, average processing time per scenario, PA and Kappa's coefficient results for $J_3 = 1, \ldots, 9$.

J_3	TKD			NTD			INTD		
	PA	κ	**MCC**	**PA**	κ	**MCC**	**PA**	κ	**MCC**
1	0.9069	0.8466	0.8481	0.9076	0.8493	0.8511	0.9002	0.8355	0.8365
2	0.9005	0.8364	0.8376	0.9155	0.8610	0.8626	0.9043	0.8424	0.8439
3	0.9056	0.8446	0.8454	0.8734	0.7956	0.8010	0.8999	0.8366	0.8384
4	0.8822	0.8081	0.8113	0.8675	0.7863	0.7926	0.8615	0.7776	0.7844
5	0.8968	0.8326	0.8346	0.8228	0.7222	0.7409	0.8703	0.7905	0.7956
6	0.8635	0.7805	0.7857	0.8637	0.7814	0.7875	0.8821	0.8084	0.8122
7	0.8973	0.8332	0.8357	0.8699	0.7905	0.7964	0.8635	0.7782	0.7820
8	0.8795	0.8061	0.8122	0.8544	0.7647	0.7709	0.9088	0.8503	0.8512
9	0.8696	0.7908	0.7983	0.9197	0.8675	0.8680	0.9057	0.8450	0.8461

[1] Raw data: PA = 0.8578, κ = 0.7709 and MCC = 0.7768.

4.2.2. Case B: Indian Pines

This dataset is a scene produced by AVIRIS in North-western Indiana and consists of 145×145 pixels and 224 spectral bands in the wavelength range $0.4 - 2.5\mu m$. The Indian Pines scene contains two-thirds agriculture, and one-third forest or other natural perennial vegetation. There are two major dual lane highways, a rail line, as well as some low density housing, other built structures, and smaller roads. Since the scene is taken in June some of the crops present, corn, soybeans, are in early stages of growth with less than 5% coverage. The ground truth available is designated into sixteen classes and is not all mutually exclusive. Indian Pines data are available at Indian Pines dataset. Figure 14a shows the true color, Figure 14b the ground truth and Table 4 shows the number of samples for each class.



(a) **(b)**

Class	Samples
Alfalfa	46
Corn-N	1428
Corn-M	830
Corn	237
Grass-P	483
Grass-T	730
Grass-PM	28
Hay-W	478
Oats	20
Soybean-N	972
Soybean–M	2455
Soybean-C	593
Wheat	205
Woods	1265
Building-GTD	386
Stone-ST	93

Figure 14 & Table 7. Indian Pines dataset, a) True color image and b) Ground truth. Table) Samples per class

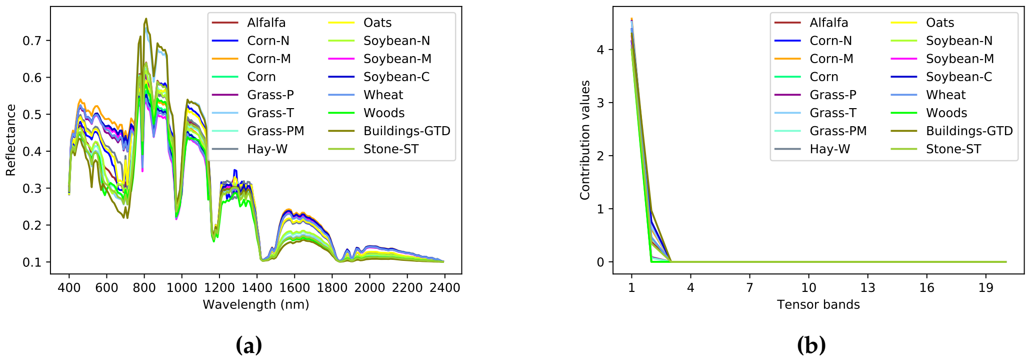


Figure 15. Behavior of the 16 classes of the Indian Pines dataset, a) in the spectral domain (spectral signatures) and b) in the the tensor bands domain after IANTD.

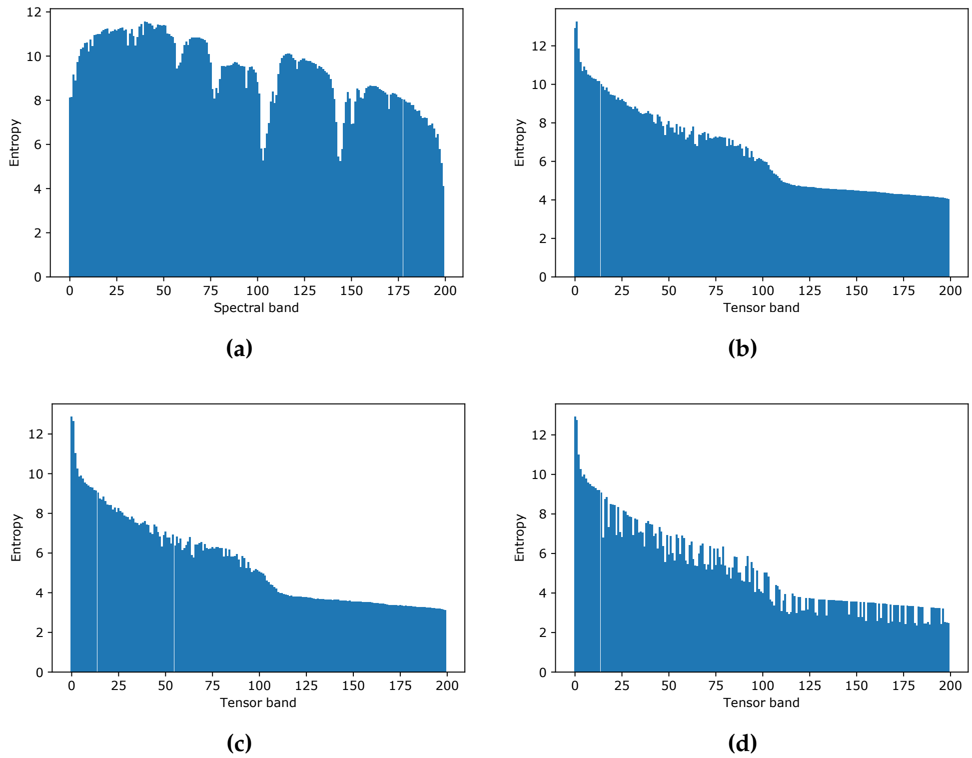
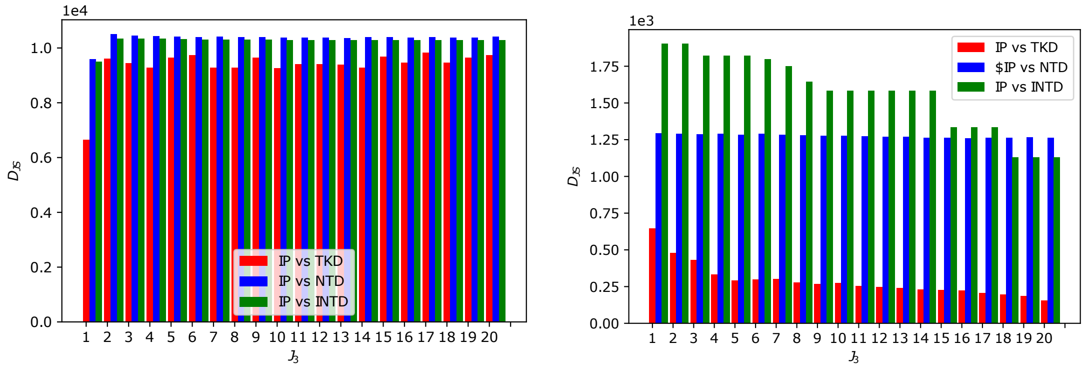


Figure 16. Indian Pines entropy of each band in the a) original dataset, and b) TKD, c) NTD and d) INTD core tensors.



(a) **(b)**

Figure 17. JSD between a) TKD / NTD / INTD core tensor vs input dataset, and b) TKD / NTD / INTD reconstruction vs input dataset.

As shown by the results obtained by measuring the divergence between the original dataset and its reconstruction for each decomposition model, non-negative decompositions produce a reconstruction error greater than the TKD, see Table **??**. Furthermore, the fall of the error is very slow as the 3-rank of the decomposition increases.

Table 8. Quantitative results[1] for the Indian Pines dataset running in a NVIDIA GeForce GTX 1050 Ti **GPU!** (**GPU!**), Intel core i7 processor, 8 Gb RAM, SSD 128 Gb, and HDD 1 Tb. Decomposition recontruction error, average processing time per scenario, PA and Kappa's coefficient results for $J_3 = 1, \ldots, 10$.

J_3	**TKD**				**NTD**				**INTD**			
	ξ	**Time (s)**	**PA (%)**	κ	ξ	**Time (s)**	**PA (%)**	κ	ξ	**Time (s)**	**PA (%)**	κ
1	375.365	15.83	82.93	0.8207	375.36	16.21	86.41	0.8252	2965.49	15.72	86.75	0.7799
2	140.6	32.83	92.51	0.9020	67.53	31.55	92.87	0.8844	2965.49	39.63	91.82	0.8972
3	116.63	56.56	88.03	0.8946	343.33	57.32	92.31	0.9074	2957.91	62.31	93.25	0.8427
4	105.57	92.13	91.76	0.8766	343.43	97.10	90.83	0.8534	2951.48	98.94	88.39	0.8855
5	98.85	156.21	88.53	0.8981	343.33	151.23	92.75	0.8597	2938.82	164.32	89.91	0.8478
6	92.52	298.80	83.99	0.8653	339.64	301.09	89.90	0.8990	2929.61	313.21	92.36	0.7913
7	87.41	520.13	89.21	0.8973	337.89	515.63	92.74	0.8523	2895.02	535.08	88.94	0.8540
8	79.53	715.69	88.33	0.8561	335.90	704.21	89.86	0.8599	2876.32	732.12	92.15	0.8469
9	76.15	881.21	89.97	0.8853	335.91	876.36	92.03	0.8891	2866.45	901.35	92.01	0.8603
10	72.67	934.78	88.61	0.8871	335.74	910.84	92.93	0.8864	2854.12	978.54	91.95	0.8462

[1] For the original Indian Pines dataset: Time = 878.09 s, PA = 91..22%, κ = 0.9040.

5. Discussion and Comparison

In this work, the hyperspectral input dataset is decomposed by a Tucker-based decomposition model to transform them from the spectral bands domain (wavelength) to a new tensor bands domain. The decompositions are restricted to preserve the spatial domain and to compress the spectral domain. Figure 15 it can be seen how the endmembers of the materials of interest behave in a way that, from a salient band point of view, the first new tensor bands are able to provide enough information to a CNN to differentiate diverse materials. On the other hand,From the information theory point of view, the entropy computed for each original and core tensors band reinforce this assertion (See Figure 16).

Unlike previous works, the introduction of information metrics in this work aids to trade off the empirical setting of the multirank TKD parameters. Although the process is still semi-empirical, it is based on metrics that quantify the amount of information and the divergence from the original data. It is worth noting that, in this work, the compression is developed only in the spectral domain, but the basis of the proposal can also be applicable for other kinds of decomposition.

Qualitative results (Figures **??**–**??**) and quantitative results in Figure **??** present the performance evaluation of the CNN, based on PA, comparing the three models based on TKD. Comparing with results shown in previous works [?], [?], [?], the proposed INTD overcomes unsupervised classification algorithms, as well as decomposition without non-negativity and integer restrictions. While it is true that the PA metric is not the best for an unbalanced dataset, it is a good starting point for a general comparison. Nevertheless, the kappa coefficients results, shown in Figure **??**, show greater stability in classification for the TKD, but as the value of J_3 increases, the NTD and INTD improve their performance, while the TKD fall. this can be attributed to the phenomenon of overfitting.

Tables 6 and 8 allow us to make a fair comparison among the Tucker-based decompositions. First of all, as expected, the TKD reconstruction error decrease faster than the approximation with non-negativity and integer constraints. On the other hand, the analyisis of PA and the Kappa's coefficient, in combination with the entropy, give a measure linked to the diminution of execution time in favor of the NTD and the proposed INTD approximation.

6. Conclusions

In this work, we had the purpose of improving the features of a multi- or hyperspectral image, while reducing dimensionality and, in turn, the computational complexity of a classification CNN. From the results presented above and the analysis of each metric, it has been shown that the constraints imposed to a decomposition model, as the NTD and INTD, produce an improvement in classification metrics of CNN. It is worth noting that, depending on the model of the classifier, the TD should be limited to provide characteristics that aids the classifier to improve its performance.

Results shown in Figure 17 we can conclude that the proposed integer non-negative approximation

6.1. Open issues

-
-
-

Author Contributions: Conceptualization, J.L.; formal analysis, D.T.; investigation, J.L.; methodology, J.L., D.T., and C.A.; resources, C.A.; software, J.L.; supervision, D.T. and C.A.; validation, D.T. and C.A.; writing—original draft, J.L. and D.T.

Funding: This work was supported by the National Council of Science and Technology CONACYT of Mexico under grant XXXXXXX.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tempfli, K.; Huurneman, G.; Bakker, W.; Janssen, L.; Feringa, W.; Gieske, A.; Grabmaier, K.; Hecker, C.; Horn, J.; Kerle, N.; et al. *Principles of Remote Sensing: An Introductory Textbook*, 4th ed.; ITC: Geneva, Switzerland, 2009.
2. He, Z.; Hu, J.; Wang, Y. Low-rank tensor learning for classification of hyperspectral image with limited labeled sample. *IEEE Signal Process.* **2017**, *145*, 12–25.
3. Richards, A.; Xiuping, J.J. Band selection in sentinel-2 satellite for agriculture applications. In *Remote Sensing Digital Image Analysis*, 4th ed.; Springer-Verlag: Berlin, Germany, 2006.
4. Zhang, T.; Su, J.; Liu, C.; Chen, W.; Liu, H.; Liu, G. Band selection in sentinel-2 satellite for agriculture applications. In Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7–8 September 2017.
5. Xie, Y.; Zhao, X.; Li, L.; Wang, H. Calculating NDVI for Landsat7-ETM data after atmospheric correction using 6S model: A case study in Zhangye city, China. In Proceedings of the 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010.
6. Gao, B. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 1–6.
7. Ham, J.; Chen, Y.; Crawford, M.; Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501.

8. Hearst, Marti A. Support Vector Machines. *IEEE Intell. Syst. J.* **1998**, *13*, 18–28.
9. Huang, X.; Zhang, L. An SVM Ensemble Approach Combining Spectral, Structural, and Semantic Features for the Classification of High-Resolution Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 257–272.
10. Delalieux, S.; Somers, B.; Haest, B.; Spanhove, T.; Vanden Borre, J.; Mucher, S. Heathland conservation status mapping through integration of hyperspectral mixture analysis and decision tree classifiers. *Remote Sens. Environ.* **2012**, *126*, 222—231.
11. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77.
12. Pirotti, F.; Sunar, F.; Piragnolo, M. Benchmark of machine learning methods for classification of a sentinel-2 image. In Proceedings of the XXIII ISPRS Congress, Prague, Czech Republic, 12–19 July 2016.
13. Mateo-García, G.; Gómez-Chova, L.; Camps-Valls, G. Convolutional neural networks for multispectral image cloud masking. In Proceedings of the IGARSS, Fort Worth, TX, USA, 23–28 July 2017.
14. Guo, X.; Huang, X.; Zhang, L.; Zhang, L.; Plaza, A.; Benediktsson, J. A. Support Tensor Machines for Classification of Hyperspectral Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3248–3264.
15. Cichocki, A.; Mandic, D.; De Lathauwer, L.; Zhou, G.; Zhao, Q.; Caiafa, C.; Phan, H. Tensor Decompositions for Signal Processing Applications: From two-way to multiway component analysis. *IEEE Signal Process. Mag.* **2015**, *32*, 145–163.
16. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer Verlag: New York, NY, USA, 2002.
17. Kolda, T.; Bader, B. Tensor Decompositions and Applications. *SIAM Rev.* **2009**, *51*, 455–500.
18. Lopez, J.; Santos, S.; Torres, D.; Atzberger, C. Convolutional Neural Networks for Semantic Segmentation of Multispectral Remote Sensing Images. In Proceedings of the LATINCOM, Guadalajara, Mexico, 14–16 November 2018.
19. European Space Agency. Available online: https://sentinel.esa.int/web/sentinel/missions/sentinel-2 (accessed on 15 July 2019).
20. Kemker, R.; Kanan, C. Deep Neural Networks for Semantic Segmentation of Multispectral Remote Sensing Imagery. *arXiv* **2017**, arXiv:abs/1703.06452.
21. Hamida, A.; Benoît, A.; Lambert, P.; Klein, L.; Amar, C.; Audebert, N.; Lefèvre, S. Deep learning for semantic segmentation of remote sensing images with rich spectral content. In Proceedings of the IGARSS, Fort Worth, TX, USA, 23–28 July 2017.
22. Wang, Q.; Lin, J.; Yuan, Y. Salient Band Selection for Hyperspectral Image Classification via Manifold Ranking. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 1279–1289.
23. Li, S.; Qiu, J.; Yang, X.; Liu, H.; Wan, D.; Zhu. Y. A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search. *Eng. Appl. Artif. Intell.* **2014**, *27*, 241–250.
24. Zhang. L.; Zhang, L.; Tao, D.; Huang, X.; Du, B. Compression of hyperspectral remote sensing images by tensor approach. *Neurocomputing* **2015**, *147*, 358–363.
25. Astrid, M.; Lee, Seung-Ik. CP-decomposition with Tensor Power Method for Convolutional Neural Networks compression. In Proceedings of the BigComp, Jeju, Korea, 13–16 February 2017.
26. Chien, J.; Bao, Y. Tensor-factorized neural networks. *IEEE Trans. Neural Networks Learn. Syst.* **2018**, *29*, 1998–2011.
27. An, J.; Lei, J.; Song, Y.; Zhang, X.; Guo J. Tensor Based Multiscale Low Rank Decomposition for Hyperspectral Images Dimensionality Reductio. *Remote Sens.* **2019**, *11*, 1485.
28. Li, J.; Liu, Z. Multispectral Transforms Using Convolution Neural Networks for Remote Sensing Multispectral Image Compression. *Remote Sens.* **2019**, *11*, 759.
29. An, J.; Song, Y.; Guo, Y.; Ma, X.; Zhang, X. Tensor Discriminant Analysis via Compact Feature Representation for Hyperspectral Images Dimensionality Reduction. *Remote Sens.* **2019**, *11*, 1822.
30. Absil, P.-A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 2007.
31. De Lathauwer, L.; De Moor, B.; Vandewalle, J. On the best rank-1 and rank-(R 1 ,R 2 ,···,R N) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1324–1342.
32. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*, 1st ed.; MIT Press, 2016.
33. Sheehan, B. N.; Saad, Y. Higher Order Orthogonal Iteration of Tensors (HOOI) and its Relation to PCA and GLRAM. In Proceedings of the 7th SIAM International Conference on Data Mining, Minneapolis, MN, USA, 26–28 April 2007.

34. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
35. De Lathauwer, L.; De Moor, B.; Vandewalle, J. A Multilinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1253–1278.
36. Rodes, I.; Inglada, J.; Hagolle, O.; Dejoux, J.; Dedieu, G. Sampling strategies for unsupervised classification of multitemporal high resolution optical images over very large areas. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012.

Sample Availability: Samples of the compounds ...... are available from the authors.