Classification of spatially enriched pixel time series with convolutional neural networks

Mohamed Chelali*, Camille Kurtz*, Anne Puissant† and Nicole Vincent*
*Université de Paris, LIPADE, Paris, FRANCE
†Université de Strasbourg, LIVE, Strasbourg, FRANCE
{firstname.lastname}@{u-paris,unistra}.fr

***Abstract*—Satellite Image Time Series (SITS), MRI sequences, and more generally image time series, constitute $2D+t$ data providing spatial and temporal information about an observed scene. Given a pattern recognition task such as image classification, considering jointly such rich information is crucial during the decision process. Nevertheless, due to the complex representation of the data-cube, spatio-temporal features extraction from $2D+t$ data remains difficult to handle. We present in this article an approach to learn such features from this data, and then to proceed to their classification. Our strategy consists in enriching pixel time series with spatial information. It is based on Random Walk to build a novel segment-based representation of the data, passing from a $2D+t$ dimension to a $2D$ one, without loosing too much spatial information. Such new representation is then involved in an end-to-end learning process with a classical $2D$ Convolutional Neural Network (CNN) in order to learn spatio-temporal features for the classification of image time series. Our approach is evaluated on a remote sensing application for the mapping of agricultural crops. Thanks to a visual attention mechanism, the proposed $2D$ spatio-temporal representation makes also easier the interpretation of a SITS to understand spatio-temporal phenomenons related to soil management practices.**

I. Introduction

An image time series generally takes the form of an ordered set of images, sensed over the same spatial scene at different times. Its acquisition can make use of different sensor sources to obtain a larger data series with short time interval between two images. These $2D+t$ data carry rich information about the temporal evolution of a scene. In remote sensing, many constellations of satellites acquire images with high spatial, spectral and temporal resolutions around the world leading to Satellite Image Time Series (SITS). Such data are now largely (freely) available and are involved in various applications related to Earth observation: understanding environmental evolution, studying the causes of particular (natural or artificial) changes, and predicting future evolution... Spectral and spatial dimensions, coupled with temporal information, enable in particular the analysis of complex patterns related to land cover mapping applications (e.g. agricultural zones, urban areas) or the identification of land use changes (e.g. urbanization, deforestation), leading ultimately to the production of accurate land-cover maps of a territory [1].

The joint consideration of the temporal and the spatial dimensions of the $2D+t$ data-cube remains a major issue when analyzing image time series. Classical methods for SITS analysis are actually mainly based on temporal information [2] studied at pixel level. In some specific applications, this may not be sufficient to get satisfactory results. Taking directly into account spatio-temporal information can make easier the recognition of different complex land cover classes (e.g. agricultural crop rotation, peri-urban areas), prone to confusions when a single image (at a particular date) is used.

In this article, we focus on the problem of spatio-temporal features extraction for the classification of image time series, using an end-to-end (deep) learning paradigm. A novel spatio-temporal representation of image time series is defined, making it possible to consider classical $2D$ Convolutional Neural Network (CNN) frameworks. Our contribution relies on a transformation to represent $2D+t$ data as $2D$ images, without loosing too much spatial information. It is based on the construction of a set of $(1D)$ segments using a Random Walk paradigm to decrease the spatial dimension of the data. The $1D$ temporal dimension is attached to the $1D$ spatial segments to build the $2D$ representation. A CNN is then fed with this new representation in order to learn spatio-temporal features with only $2D$ filters, (involving at the same time temporal and spatial information), that are used to classify image time series. The proposed spatio-temporal $2D$ representation also leads to a novel way of structuring a SITS facilitating its interpretation. To benefit from this, we integrate in our approach a visual attention strategy that allows us to better understand spatio-temporal phenomena occurring in the scene.

The remainder of this article is organized as follow. Section II presents related works for SITS analysis. In Section III, we introduce the proposed $2D$ spatio-temporal representation of the image time series for a CNN analysis. Section IV presents an experimental study, in the remote sensing domain, focusing on the classification of agricultural crops. Section V explains the proposed attention mechanism, involved in our framework, for visual explanations from CNN. Finally, conclusion and perspectives will be found in Section VI.

II. Related works

Satellite images allow the monitoring of large-scale ground surfaces observed from space. Thanks to novel satellite constellations, these images are now available along several months with a high acquisition rate. Such data may improve our understanding of environmental evolution and changes, which can be of different types, origins and duration [3].

The oldest methods proposed to analyze such $2D + t$ data were designed to process single images from image stacks. Various measurements per pixel were considered, on each image, as independent features and involved in supervised procedures. In such approaches, the date of the measurements was ignored in the feature space. Bi-temporal analysis methods were then considered to analyze changes occurring between two observations [4], [5].

Another category of approaches is more directly dedicated to the processing and analysis of image time series, such as multi-date classification methods. For example, we can cite radiometric trajectory analysis [6], exploiting the notion that land cover can vary through time (e.g. vegetation evolution, seasons [7]). Other approaches take into account the time dimension by considering dedicated time series analysis methods [8]. Here, each pixel is considered as time ordered (and aligned) series of measurements, leading to the notion of temporal pixel, and the changes in the measurements through time are analyzed to find temporal (symbolic or statistical) patterns.

Some researches propose to represent the SITS into a new space before performing its analysis. For example "frequency-domain" approaches include spectral analysis, wavelet analysis while "time-domain" approaches involve auto-correlation and cross-correlation analysis [9]. Certain methods extract more discriminative "hand-crafted" features in a new enriched space [10]–[12] and the classification is operated into this enhanced space. Concerning the classification step, classical approaches measure similarity between any new sample – a temporal pixel (that can be enriched with the "hand-crafted" features mentioned previously) and the training set. For example, the Euclidean distance, or the Dynamic Time Wrapping [13] measure, coupled to a nearest neighbor algorithm, can be used to assign the label of the most similar class.

Few years ago, deep learning strategies have been considered for remote sensing image classification, producing land-cover maps at a large scale. In majority, CNNs are considered to deal with the spatial domain by applying $2D$ convolutions to the data [14]. Note that $1D$ convolutions can be applied in the temporal domain when dealing with image time series [2]. To benefit from the advantages of $2D$ CNNs, an idea is also to transform a $1D$ signal data (e.g. the time) to a $2D$ representation, compatible with CNN architectures pre-trained on large image datasets. This has been already studied in the domain of sound classification with time-frequency representations [15] or on-line handwritten recognition [16].

Recurrent Neural Networks (RNN) such as Long-Short Term Memory (LSTM) represent another type of deep learning architecture that is designed for temporal data. They have been used successfully in [17], [18] for particular earth observation applications. In this context, deep approaches outperform classical classification algorithms such as Random Forest [19], but they do not directly take into account the spatial dimension of the images as they consider (temporal) pixels in an independent manner. Some approaches have been proposed to consider both the temporal and the spatial dimensions of the $2D + t$ data-cube. A common strategy is to train two models, one for spatial dimensions and one for temporal dimension, then to fuse their results at the decision level [20]. Authors of [21] propose an hybrid architecture applied at region level. It first encodes the spatial information, and then combines it with temporal features extracted by a self-attention module. In video analysis, spatio-temporal features can be learned directly using $3D$ CNNs [22] but such strategy requires the learning of a huge number of parameters to define a model. In [23], we introduced a first strategy to classify a SITS using a classical $2D$ CNN model by proposing a representation of image time series that embed simultaneously the temporal and the spatial dimensions of the data. We considered a single image representations based on a space filling curve, such as Hilbert curve, one dimension is spatial and the other concerns time. The pixels are ordered but each pixel has only two neighbors. The neural network processes with $2D$ convolutions temporal and spatial information at the same time. This approach was used at region of interest (ROI) level and only a signle $2D$ representation per ROI was producible, limiting the number of training examples per class and then limiting the learning capacity of deep architecture.

On the one hand, pixel-based state-of-the-art approaches, such as $1D$ temporal convolutions [2], make it possible to learn efficient models because many annotated examples are available but they do not take spatial information into account. On the other hand, our previous approach [23] takes into account spatio-temporal information but the number of examples for learning remains limited. In this article, our first contribution is to consider an intermediate point of view where we enrich temporal pixels with spatial information, also leading to a new spatio-temporal representation, combining the advantages of the two approaches.

The main criticism addressed to systems relying on deep neural networks is the lack of result interpretability. Some works have partially answered this by introducing the notion of Class Activation Map (CAM) to highlight the attention zones of a CNN [24]. Note that CAM is only employable with models that use a global average pooling layer before the softmax classifier [24]. CAM strategies have been improved in [25], by GradCAM and GradCAM++, which are based on a guided back-propagation. The notion of CAM was also used successfully for temporal convolutions of time series [26]. To benefit from these strategies, helping to understand the network decision making, our second contribution involves such attention mechanism for visual explanations and interpretation of our generated $2D$ spatio-temporal representations. This enables to select temporal interval in the data-space more adapted to the classification, ultimately improving the model, and also to investigate the spatial uniformity of the data.

III. Enriching temporal pixels with spatial information for SITS classification

In this section, we present our strategy in order to classify SITS with spatio-temporal features. Since our system is dedicated to the classification of particular objects of interest (e.g. agricultural crops), the initial input data may be an image

centered over a specific object, an image patch, or only the connected pixels of a region of interest (ROI), modeled as a polygon. In any case, we will use the term "image" and SITS to refer to the input data.

In our applicative context, one problem is the low number of annotated ROIs sharing the same label in a SITS. Since we consider the pixel level is too local to bare spatial information, we propose to build a new richer spatio-temporal representation of the data and we also want to get enough elements to perform a significant learning with a deep architecture.

The global process is divided into three parts. First we present (Section III-A) the strategy of data representation before going more deeply in the details, then the learning phase (Section III-B) enables to define the spatio-temporal features that are most significant according to the classification task. Finally in (section III-C), we propose the last on-line step of decision making that is based on the use of previous learning step. Figure 1 illustrates the workflow of our system.

A. *Data representation*

We first explain how to transform the original $2D + t$ data to less complex $2D$ spatio-temporal representations.

From pixels to segments: Most methods consider a ROI covered by a SITS as a set of independent pixels sharing the same label. The time is taken into account as the pixel is represented by the evolution of its intensity values along time. There are quite a lot of pixels, enabling the learning phase at pixel level. This property seems important and we want to keep it in mind. Besides, some spatial information is needed to improve the final results. Then, we propose to replace the pixels by a $1D$ curve segment containing the pixel. Such segment contains neighbouring pixels figuring the spatial relation between pairs of pixels and even further, depending on the segment length L. The pixels in the segment are ordered by the curvilinear abscissa in the curve (see top of Figure 1).

Segment construction strategies: In our case, the construction of the segments is based on a Random Walk (RW) process. RW is a mathematical model with a random iterative system with Markovian properties. The RW is used to construct a random segment in a $2D$ space with length L, noted $RW(L)$. The initialization of the first point of the segment is done randomly on the $2D$ image. For each next point of the segment, 8 or 4 directions are possible, according to the considered connexity. In our case we chose the 8 connexity.

So, we propose to consider the image (respectively the SITS), as a set of N_p independent segments (respectively temporal segments) sharing the same class label. The N_p segments can be interlaced and can pass through the same pixel several times but according to different directions. This is a strength of the method since different spatial configurations can be captured.

B. *Segment labeling*

A CNN model is used in order to label the segments. We have two types of information: time and space. Besides it is well admitted that the learning of a CNN system, dedicated to

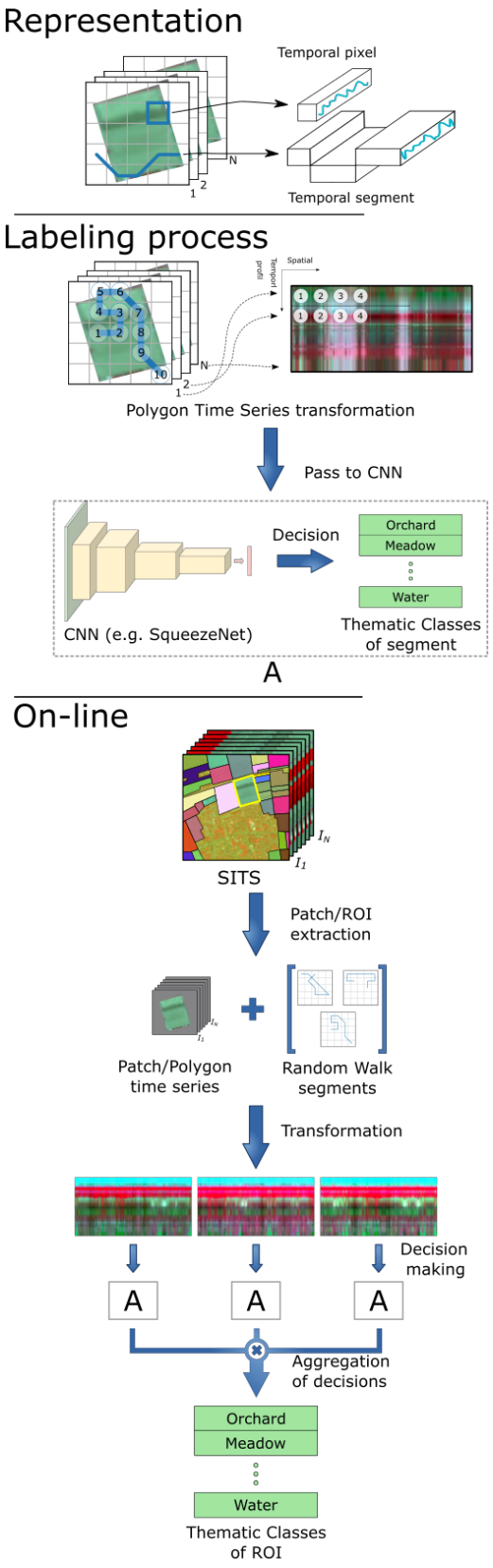


Fig. 1. Flowchart of our method for image time series deep classification based on a planar spatio-temporal data representation obtained from Random Walk based segments. (top) data representation, (center) learning / labeling phase enabling to define the spatio-temporal features and (bottom) on-line (i.e. decision making) phase of the classification process.

classification task, can benefit from a good initialization, e.g. using the weights computed from the IMAGENET classification database. From these two remarks, we have chosen first to transform the segments evolving along time in $2D$ images, and then to label the $2D$ representations thanks to a classical $2D$ CNN. These two steps are presented hereinafter.

From segments to $2D$ representations: Each $1D$ segment can be associated with a $2D$ structure. In the abscissa, X axis, is considered the index of the pixel in the segment (from the initial pixel) and in the ordinate, Y axis, is considered the evolution of the intensity of the pixels over time. This leads to a novel $2D$ representation composed of N rows (N is the number of images in the SITS) in the temporal domain and L columns (L is the length of the considered segment) in the spatial domain. This image can then be interpreted as a (partial) spatio-temporal $2D$ representation of the $2D+t$ SITS.

Such representations will be used as input of a learning process, the segments classes are the labels of the annotated input image they belong to (see center of Figure 1).

CNN model (architecture): Thanks to a limited set of labeled ROIs in SITS, it is possible to build a large set of segment representations as several different segments (i.e. N_p) can be extracted from a single ROI (at least as many as the number of pixels contained in the ROI from the SITS).

As a classifier, we have chosen SqueezeNet [27] but any other $2D$ CNN model could be used. SqueezeNet has interesting properties, few parameters with same accuracy level as the AlexNet model on the IMAGENET dataset. The training of the model is then faster. The architecture of SqueezeNet differs from a classical CNN. It introduces a new module called Fire composed of a squeeze layer using 1×1 convolution filters followed by expand layer that contains a mix of 1×1 and 3×3 convolution filters. Also, its classifier part is based on a global average pooling over feature maps, potentially decreasing the overfitting effect. The CNN model is trained with the $2D$ spatio-temporal representations obtained from each input image time series from the training set.

C. Decision making

As already mentioned, our input data are raster polygons representing particular objects of interest in SITS. Each input data is associated with a set of N_p segments, N_p is consequently a parameter of the method. Thanks to the classifier described in Section III-B, a class label is predicted for each segment with some probability (see bottom of Figure 1). We proceed by taking average of the returned probabilities by the model for the N_p segments of the polygon and we affect the class label with the highest probability ensuring a unique decision per ROI.

IV. APPLICATION: AGRICULTURAL CROPS CLASSIFICATION

This experimental study is focused on a remote sensing application, the classification of agricultural crop-fields from SITS with low spatial resolution. The goal is to discriminate within four agricultural thematic classes: meadows, vineyards, traditional orchards and intensive orchards. The automatic identification of these classes is a complex task since these agricultural crop-fields are subject to many agricultural practices depending on the season and the territory management policy. In order to differentiate these classes, spatio-temporal features carry useful information to discriminate the agricultural practices, such as tree organizations, mowing, etc.

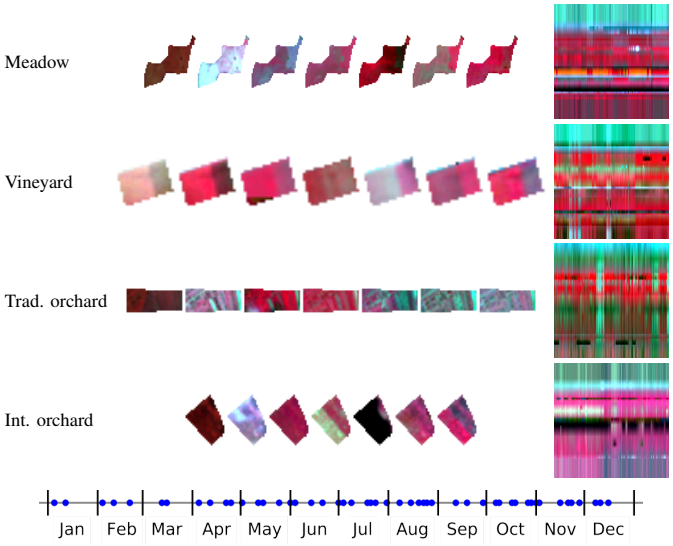


Fig. 2. Temporal evolution of four different agricultural crop-fields (a meadow, a vineyard, a traditional orchard and an intensive orchard) according (left) to their initial $2D + t$ representation from a SITS and (right) to our $2D$ spatio-temporal representation. The last line shows the distribution of the images from the SITS (2017).

A. Material

We study SITS provided by the satellite Sentinel-2 containing $N = 50$ optical images sensed in 2017 over the same geographical area (Alsace, France – tiles 32ULU, 32ULV). Figure 2 displays the temporal distribution of the SITS. The images have been corrected and orthorectified by the French Theia program[1] to be radiometrically comparable. We also dispose of the cloud, shadow and saturation masks associated with each image. A pre-processing step was applied on the images with a linear interpolation on masked pixels to fill the missing values in the SITS.

For each image, only three bands are kept which are near-infrared (Nir), red (R) and green (G). The blue band (B) is considered as useless in the literature to discriminate different kinds of agricultural fields and is also sensitive to atmospheric effects. All these bands have a spatial resolution of 10 meters.

The used reference data are extracted from the (freely distributed) RPG[2], which is the agricultural parcel delineations. Some examples of polygons are represented in Figure 2. These polygons have been corrected by photo-interpretation to ensure a good delimitation of the parcels. The reference data used in our experiment are the semantic labels of these polygons corresponding to the four classes.

[1]https://theia.cnes.fr/
[2]http://professionnels.ign.fr/rpg

TABLE I
DATA SUMMARY; (SECOND COL.) INITIAL NUMBER OF POLYGONS PER CLASS; (THIRD COL.) AVERAGE NUMBER OF PIXELS PER POLYGON; (FOUR LAST COL.) NUMBER OF SPATIO-TEMPORAL SEGMENTS.

Classes	# poly.	$\overline{area}$	# Spatio-temp. rep. for *RW* (N_p)			
			10%	20%	50%	70%
Meadows	1 045	250±338	26 110	51 688	128 424	179 914
Vineyards	562	50±47	3 060	5 821	14 137	19 853
Trad. orchards	136	154±305	2 146	4 222	10 474	14 672
Int. orchards	191	129±115	2 564	5 027	12 414	17 408
Total	1 934	–	33 880	66 758	165 449	231 847

B. Data preparation

From the SITS and the polygons representing the ROIs, we extract segments to build the proposed $2D$ spatio-temporal representations. As already mentioned, one interest of our approach is to make it possible to generate a high number N_p of representations per ROI, depending of the initialization of the segments. Thanks to the RW process, segments can capture different spatial configurations per ROI (by passing through a same pixel in different directions) in the $2D$ representations. We have experimentally chosen to extract from each ROI a number N_p of segments depending on the ROI area (i.e. the number of pixels composing a polygon). We experimentally use 10%, 20%, 50% and 70% of the size of each polygon. Table I displays the number of instances of polygons per class and the number of (temporal) segments built from these data. We also indicate the mean area (noted $\overline{area}$) of polygons for each class.

In this study, we analyze the importance of the pixel spatial relationships, highlighting the interest of enriching temporal pixels with spatial information. Two factors are studied:

1) the interest of considering a RW strategy to build continuous strings of pixels instead of generating totally random strings of pixels, leading to "naive" representations noted $\mathfrak{R}$;
2) the impact of the length L of the segments. This enables to evaluate the impact of adding more spatial information to learn spatio-temporal features instead of considering single $0D$ pixels, as this is the case in temporal-based approaches. We used $L = 50$ and 100.

The SqueezeNet model takes as input images of size 224×224. When building the 224×224 $2D$ images from the segments of length L, if $L < 224$, we center them horizontally and the rest of columns are fixed to zero values.

For the temporal dimension (Y axis), we fill the 224 values by applying a linear interpolation in the STIS on time information. We assume that the temporal information between two consecutive dates is monotonic and linear. The interpolation is then done by considering that we only have 224 days in the year so that one day is done with about 39 hours. For the initial date (beginning of the year 224 days), we affect the temporal information of the first date in the SITS. For the last date (end of the year 224 days), we affect the last known temporal information in the SITS.

The data normalization is a linear transform based on the maximum and the minimum values of the dataset after values are limited with 2% (or 98%) percentile, as proposed in [2].

C. Learning and validation protocol

The experiments are validated using a five-fold cross validation strategy. Each time, we split the dataset into three subsets at polygon level with sizes of 60%, 20% and 20% representing respectively training, validation and test sets. The CNN model is then trained and evaluated five times. In the end, we report the average overall accuracy (OA) of the five splits and indicate the standard-deviation (STD).

The model is trained using *Adam* optimizer with a learning rate of 10^{-6} and default values of the other parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$, as suggested in [28]) with a batch size of 128. We use an early stopping technique with a patience number of 20. As the dataset is relatively unbalanced, a weighted cross entropy loss function was considered to avoid a potential overfitting towards the majority classes.

In this study, the CNN is initialized with weights obtained with the IMAGENET database in a classification problem (i.e. the ILSVRC challenge) and then fine-tuned with our data.

D. Results and discussion

The $2D$ spatio-temporal representations are used to feed the chosen CNN. For the totally random strategy $\mathfrak{R}$, we use the segment length $L = 100$ as first baseline.

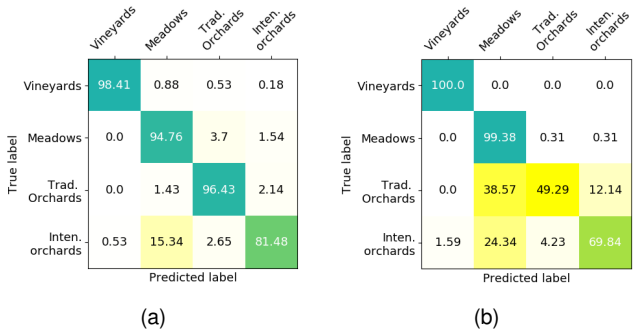
The classification results (overall accuracy) with spatio-temporal representations are reported in Table II. We remark that with less percentage (10%) of temporal segments, all scores are in the same range. But when we increase this percentage, all the scores are increased except for the totally random strategy $\mathfrak{R}(100)$. This highlights the interest of considering a RW strategy that conserves a part of the spatial relationships among pixels to build continuous strings of pixels instead of generating totally random strings of pixels. The next result is related to the impact of the length L of the segments ($RW(50)$ vs. $RW(100)$). From Table II, we observe that the scores are slightly higher when integrating more pixels in the spatio-temporal representation, as this was expected.

The last result is related to the number N_p of spatio-temporal representations that we build for one given polygon to increase the learning set. We notice, as expected, that the higher this number is, the higher the capacity of the system to discriminate the different classes is.

To highlight the interest of our approach, we compare the obtained scores with those obtained with TempCNN [2] and the method presented in [20]. TempCNN is dedicated to the classification of time series, where $1D$ convolutions are applied in the temporal domain. The used filter size is 11 since we consider interpolated dates [2]. Note that the TempCNN model is proposed in [2] with different architectures (depths of the network). In Table II we report the results from the best architecture. The method presented in [20] by Di Mauro *et al.* is designed to learn spatio-temporal representation of pixels time series by using two models associated respectively

TABLE II
CLASSIFICATION RESULTS (OVERALL ACCURACY – OA AND STANDARD DEVIATION – STD) WITH OUR SPATIO-TEMPORAL REPRESENTATIONS.

Lengths L of the segments	percentage	OA	STD
$\Re(100)$	10	91.97	1.11
	20	91.72	2.72
	50	92.29	0.60
	70	91.75	0.61
RW(50)	10	91.07	2.53
	20	93.80	1.57
	50	94.06	1.44
	70	94.80	1.57
RW(100)	10	92.50	1.05
	20	93.20	0.65
	50	94.21	1.19
	70	94.64	0.80
Comparative methods			
TempCNN [2]	–	92.98	0.89
Di Mauro et al. [20]	–	91.28	0.53



(a) (b)

Fig. 3. Confusion matrices: (a) with our method $RW(100)$ where N_p corresponds to 70% of segments per polygon; (b) with TempCNN [2].

with the temporal and the spatial domains. Temporal features are extracted by applying $1D$ convolutions while spatial ones are extracted using a MLP that takes as input, hand-crafted features computed at ROI level. In the experiment we use the same features as in [20], in particular the average and the standard deviation of pixel time series. The final decision is taken by a fully connected layer that considers the concatenation of the features extracted by the two models. We use the same learning and validation protocol as explained in Section IV-C.

The bottom part of Table II reports the obtained results with TempCNN [2] and the method presented in [20]. We notice that the best obtained scores with our method are higher than those obtained by competitors. This highlights, in our applicative context, the benefit of considering a classical $2D$ CNN model for classifying $2D+t$ images thanks to our spatio-temporal representations.

To get deeper in the results, Figure 3 shows the confusion matrices obtained with our best model and the best TempCNN [2] for a class-by-class analysis. We can notice that both approaches tend to confuse meadows and orchards, despite the consideration of spatio-temporal features in our approach. This motivates the use in our system of an attention mechanism in order to better understand why the learned spatio-temporal features as well as the network lead to such errors.

V. ATTENTION MECHANISM FOR VISUAL EXPLANATIONS

One of the most important criticism that is done to the systems relying on neural network is their inability to explain why they have produced their decision. Systems that are looking for objects in an image, producing different labels, are able to focus their attention on some regions of the input image where the element of interest may lie, e.g. a car or a pedestrian. In our thematic application, the aim is to classify some ROIs in an automatic fashion according to different labels. The difference that may appear between ROIs are related both to the vegetation cycle and to the interference of the human farmers adding fertilizer, or mowing some parts of the crop. Then, we want to have a deeper insight in the way the labels are determined. Contrary to image analysis classification, here we are processing synthetic images we have built to feed the network where the ordinates indicate a date but the abscissa are related to pixels depending on path considered in the crops. We aim to study in an independent way, the temporal and spatial aspects.

For this purpose we have considered the class activation maps $CAM(S)$ associated with the last layer of the convolutional part of the CNN used to label the temporal segments S issued from a ROI (with GradCAM++ [25]).

A. Temporal attention

First, we focus on the temporal aspect. The vertical axis indicates the same time scale for all the ROIs that are aligned and can then be compared. The top of the image is related to spring, the central part is related to the summer season and the bottom part to the winter season. The attention study will then be focusing here, on the evolution along vertical axis.

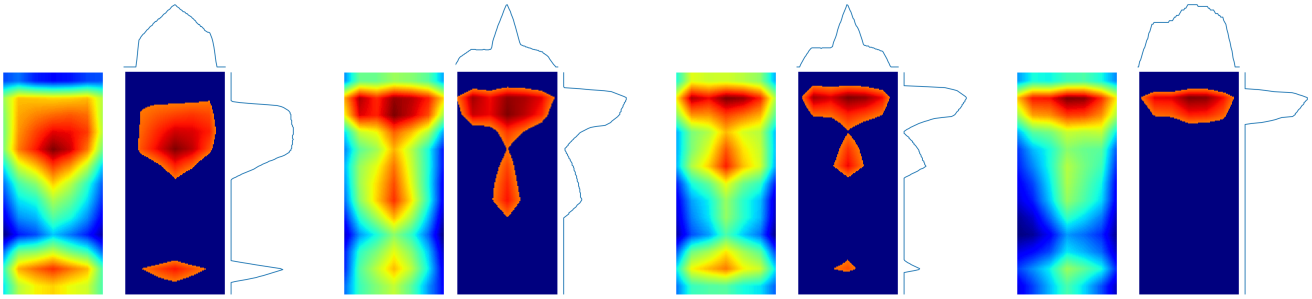
More precisely let us note $a^S_{i,j}$ the value of attention associated with the i^{th} pixel of a segment S at date j in $CAM(S)$. Accumulated along the segment, we obtain, by adding the attention map values on each row, a vector X:

$$X_j = \sum_{i=1}^{L} a_{i,j} \tag{1}$$

Each component X_j is associated with a date in the year and the size of X is N.

For a set of ROI samples belonging to different classes, the X vectors can be computed associated with each class X^k or to the whole classes. The principle of our approach is to select the dates that are most involved in taking the decision, those for which the attention values are high. Then, the X vectors can be embedded in $\{0,1\}^N$ giving AX vectors. The element transformed to 1 are those with high values and 0 correspond with those with low values. The AX vectors are obtained from the X vectors thanks to a Otsu thresholding.

The most important times of the year that enable the discrimination are corresponding to the dates where the X^k components are high. Then we consider dates where at least one class has an attention value sufficiently high to build a



(a) Vineyards (b) Meadows (c) Traditional orchards (d) Intensive orchards

Fig. 4. (Left column) Mean of attention maps. Only the central 100 columns are depicted since we considered the $RW(100)$ spatio-temporal representations (the rest of columns are fixed to zero value). (Right column) Otsu thresholds after projecting the attention values on horizontal and vertical axes.

temporal mask M. Finally, instead of comparing the time series all around the year, they can be shorten, taking into account only selected times in M:

$$M = (M_j)_j = \begin{cases} 1 & \text{if } \exists k \quad AXk_j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Figure 4 illustrates the obtained attention maps on the four different agricultural classes.

By restricting our classification problem to a 2-class study (trad. orchards vs intensive orchards), we observe experimentally that from these attention maps, the most discriminative period is between time 1 and 120. Refining the learning of our model on this temporal range, we obtained a more accurate result than with all the year range. More precisely, the decrease in errors is about 7%. Another conclusion of this study is that temporal attention enables to understand, that spring season was more significant than the rest of the year to discriminate the classes.

B. Spatial attention

On the horizontal axis, the different information at one abscissa lead to interpretations that cannot be compared since each segment is constructed using a RW process. Nevertheless, here we aim to use attention to highlight which (spatial) part of a crop-field is more discriminative to take a decision.

For one polygon belonging to the SITS, N_p $2D$ spatio-temporal representations are built (from N_p different segments S) and are used to classify the polygon by aggregating all the probabilities obtained for each $2D$ representation by the CNN classifier. We can then associate with each $1D$ RW segment S, the class probability obtained from the CNN and compute the class activation maps $CAM(S)$ as presented in Figure 4. Each of the L columns is associated with a pixel i of the segment and the value in the column $a^S_{i,j}$ indicates the attention given to this pixel at time j in order to classify it in one of the K classes. This value is depending on the date. By considering an optimistic approach, we associate with a pixel, the highest value of $CAM(S)$ as $Y_i = \max_{j=1}^{N} a_{i,j}$.

To be able to extract a spatial attention we then embed the pixel in the spatial domain according to the decision of the CNN with respect to the path S. As we explore N_p paths in a polygon, in the original $2D$ image space a pixel can be covered more than once with different labels. K spatial attention maps are built, involving the N_p attention maps associated with each path. We define each of the K spatial attention SAk in an optimistic way. Among all the paths going through the pixel we take the highest probability value or the paths giving k as a conclusion class $c(S)$:

$$SAk(x,y) = \max_{S} Y^k_{(x,y)} \ / \ c(S) = k \text{ and } p(x,y) \in S \tag{3}$$

where (x, y) indicates the spatial coordinates of a pixel p. This leads for each ROI of the SITS to K spatial attentions SAk showing the spatial attention of our system and to each class.

Finally, we can fuse these results in one spatial semantic map. The label of each pixel is the label of the class where the spatial attention is maximum:

$$SA(x,y) = \arg\max_{k=1}^{K} SAk(x,y) \tag{4}$$

In our case, we obtain a 4-color spatial semantic map since we consider four classes.

Figure 5 illustrates the obtained spatial semantic maps on two meadows. The first one has been misclassified as a traditional orchard. In fact, on the spatial map it can be seen this is due to the lower part of the ROI. When looking at a high resolution image of this area (taken from Google map) we notice that this bottom part contains trees. This highlights the interest of considering this visual attention strategy to better understand the errors made by our systems based on the image tested. Moreover, in the second example we consider a meadow that has been well classified. Nevertheless, in the spatial attention map some green pixels appear indicating a possible traditional orchard. We observe on the associated high resolution image they mostly are in zones where some tree is present. In particular here we highlight the spatial heterogeneity of the ROI, due to specific agricultural practices which can lead to classification errors in agricultural crops.

VI. Conclusion

In this article, we present a method to classify an image time series based on a spatio-temporal representation. This representation aims to reduce the structure of the data from

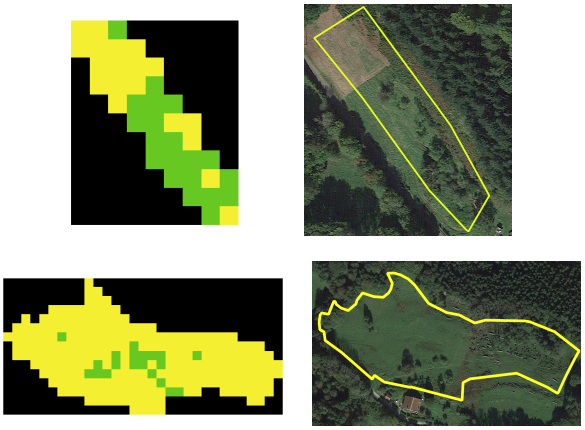


Fig. 5. Spatial semantic maps on meadows with our best model obtained with $RW(100)$. Yellow color represents a "meadow" decision while green color represents "trad. orchard" decision.

$2D+t$ to $2D$ without loosing too much the spatial relationship of pixels and the temporal one. Then, these new representation images are used to feed a classical CNN to perform a classification of temporal segments, corresponding to temporal pixels enriched with spatial information. With the proposed representation, the applied $2D$ convolutions lead to a spatio-temporal features extraction process. The trained filters have weights linked to the temporal evolution and others linked to spatial evolution. Finally, the combination of both, carries information on spatio-temporal evolution. By considering $2D$ convolutions on this kind of images, we can also benefit of a pre-trained model from ImageNet on a similar classification problem. The proposed representation of a SITS also facilitates the interpretation of the data-cube to better understand spatio-temporal patterns related to soil management practices. We embed into our system a visual attention mechanism that allows us to observe which are the best time ranges used by the CNN to make its decisions, and the spatial heterogeneity of the crop-fields, to ultimately improve the model.

ACKNOWLEDGMENT

This work is funded by the French ANR (17-CE23-0015).

REFERENCES

[1] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, "Operational high resolution land cover map production at the country scale using satellite image time series," *Remote Sens*, vol. 9, no. 1, pp. 95–108, 2017.

[2] C. Pelletier, G. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sens*, vol. 11, no. 5, pp. 523–534, 2019.

[3] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Digital change detection methods in ecosystem monitoring: A review," *Int J Remote Sens*, pp. 1565–1596, 2004.

[4] R. Johnson and E. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int J Remote Sens*, vol. 19, no. 16, pp. 411–426, 1998.

[5] L. Bruzzone and D. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans Geosci Remote Sens*, vol. 38, no. 3, pp. 1171–1182, 2000.

[6] J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor, "Detecting trend and seasonal changes in satellite image time series," *Remote Sens Environ*, vol. 114, no. 1, pp. 106–115, 2010.

[7] C. Senf, P. Leitao, D. Pflugmacher, S. Van der Linden, and P. Hostert, "Mapping land cover in complex mediterranean landscapes using landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery," *Remote Sens Environ*, vol. 156, pp. 527–536, 2015.

[8] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Min Knowl Discov*, vol. 31, no. 3, pp. 606–660, 2017.

[9] L. Andres, W. Salas, and D. Skole, "Fourier analysis of multi-temporal AVHRR data applied to a land cover classification," *Int J Remote Sens*, vol. 15, no. 5, pp. 1115–1121, 1994.

[10] P. Ravikumar and V. S. Devi, "Weighted feature-based classification of time series data," in *CIDM, Procs.*, 2014, pp. 222–228.

[11] F. Petitjean, C. Kurtz, N. Passat, and P. Gançarski, "Spatio-temporal reasoning for the classification of satellite image time series," *Patt Rec Letters*, vol. 33, no. 13, pp. 1805–1815, 2012.

[12] M. Chelali, C. Kurtz, A. Puissant, and N. Vincent, "Urban land cover analysis from satellite image time series based on temporal stability," in *JURSE, Procs.*, 2019, pp. 1–4.

[13] F. Petitjean, J. Inglada, and P. Gançarski, "Satellite image time series analysis under time warping," *IEEE Trans Geosci Remote Sens*, vol. 50, no. 8, pp. 3081–3095, 2012.

[14] B. Huang, K. Lu, N. Audebert, A. Khalel, Y. Tarabalka, J. Malof, and A. Boulch, "Large-scale semantic classification: Outcome of the first year of inria aerial image labeling benchmark," in *IGARSS, Procs.*, 2018, pp. 6947–6950.

[15] B. Bozkurt, I. Germanakis, and Y. Stylianou, "A study of time-frequency features for cnn-based automatic heart sound classification for pathology detection," *Comp in Bio and Med*, vol. 100, pp. 132–143, 2018.

[16] C. R. Pereira, D. R. Pereira, G. H. Rosa, V. H. C. de Albuquerque, S. A. T. Weber, C. Hook, and J. P. Papa, "Handwritten dynamics assessment through convolutional neural networks: An application to parkinson's disease identification," *Artif Intell Medicine*, vol. 87, pp. 67–77, 2018.

[17] M. Russwurm and M. Korner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *EarthVision@CVPR, Procs.*, 2017, pp. 1496–1504.

[18] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci Remote Sens Lett*, vol. 14, no. 10, pp. 1685–1689, 2017.

[19] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller, "Deep learning for time series classification: A review," *Data Min Knowl Discov*, vol. 33, no. 4, pp. 917–963, 2019.

[20] N. Di Mauro, A. Vergari, T. M. A. Basile, F. G. Ventola, and F. Esposito, "End-to-end learning of deep spatio-temporal representations for satellite image time series classification," in *DC@PKDD/ECML, Procs.*, 2017, pp. 1–8.

[21] V. Sainte Fare Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *CVPR, Procs.*, 2020.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatio-temporal features with 3D convolutional networks," in *ICCV, Procs.*, 2015, pp. 4489–4497.

[23] M. Chelali, C. Kurtz, A. Puissant, and N. Vincent, "Image time series classification based on a planar spatio-temporal data representation," in *VISAPP, Procs.*, 2020, pp. 276–283.

[24] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR, Procs.*, 2016, pp. 2921–2929.

[25] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV, Procs.*, 2018, pp. 839–847.

[26] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller, "Deep learning for time series classification: a review," *Data Min Knowl Discov*, vol. 33, no. 4, pp. 917–963, 2019.

[27] F. Iandola, M. Moskewicz, K. Ashraf, S. Han, W. Dally, and K. Keutzer, "Squeezenet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size," *arXiv*, vol. abs/1602.07360, 2016.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR, Procs.*, 2015.