COMBINING MULTISCALE FEATURES FOR CLASSIFICATION OF HYPERSPECTRAL IMAGES: A SEQUENCE-BASED KERNEL APPROACH

Yanwei Cui, Laetitia Chapel, Sébastien Lefèvre

Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France
{yanwei.cui, laetitia.chapel, sebastien.lefevre}@irisa.fr

ABSTRACT

Nowadays, hyperspectral image classification widely copes with spatial information to improve accuracy. One of the most popular way to integrate such information is to extract hierarchical features from a multiscale segmentation. In the classification context, the extracted features are commonly concatenated into a long vector (also called stacked vector), on which is applied a conventional vector-based machine learning technique (*e.g.* SVM with Gaussian kernel). In this paper, we rather propose to use a sequence structured kernel: the spectrum kernel. We show that the conventional stacked vector-based kernel is actually a special case of this kernel. Experiments conducted on various publicly available hyperspectral datasets illustrate the improvement of the proposed kernel *w.r.t.* conventional ones using the same hierarchical spatial features.

Index Terms— Spectrum kernel, hierarchical features, multiscale image representation, hyperspectral image classification

1. INTRODUCTION

Integration of spatial information paves the way for improved accuracies in hyperspectral image classification [1], as the use of spatial features extracted from image regions produces spatially smoother classification maps [2]. A common approach to extract such features is to rely on multiscale representations, *e.g.* through (extracted) attribute profiles [3] or hierarchical spatial features [4]. In this framework, features from multiple scales are extracted to model the context information around the pixels through different scales. Features computed at each scale are then concatenated into a unique (long) stacked vector. Such a vector is then used as input into a conventional classifier like SVM. Representative examples of this framework include [4, 5, 6]. While defining kernels on stacked vectors is a simple and standard way to cope with hierarchical spatial features, it does not take into account the specific nature (*i.e.* hierarchical) of the data.

Indeed, hierarchical spatial features extracted from hyperspectral images can rather be viewed as a sequence of data, for which structured kernels are commonly applied in other fields. Among the existing sequence structured kernels, the spectrum kernel based on subsequences of various lengths has been successfully applied in various domains, *e.g.* biology for protein classification [7, 8] or nature language processing for text classification [9]. Its relevance for hyperspectral image classification remains to be demonstrated and is the main objective of this paper. Indeed, by applying the spectrum kernel onto hierarchical spatial features, we can explicitly take into account the hierarchical relationships among regions from different scales. To do so, we construct kernels on various lengths of subsequences embedded in the whole set of hierarchical spatial features instead of modeling this set as a single stacked vector, the latter actually being a particular case of the sequence kernel. Furthermore, we also propose an efficient algorithm to compute the spectrum kernel with all possible lengths, thus making realistic to apply such a kernel on hyperspectral images.

The paper is organized as follows. We first briefly recall some background on hierarchical image representation (Sec. 2). We then detail the concept of spectrum kernel (Sec. 3), and introduce an efficient algorithm for its computation. Evaluation of proposed method is detailed in Sec. 4, before giving a conclusion and discussing future works.

2. HIERARCHICAL IMAGE REPRESENTATION
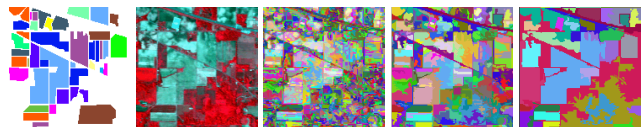
Fig. 1: The image representation of `Indian Pines` at different segmentation scales. From left to right: ground truth, false-color image, fine level (2486 regions), intermediate level (278 regions), coarse level (31 regions).

Hierarchical image representation describes the content of an image from fine to coarse level (as illustrated in Fig. 1) through a tree structure, where the nodes represent the image regions at different levels and the edges model the hierarchical relationships among those regions. Such representation

is commonly used in the GEOgraphic-Object-Based Image Analysis (GEOBIA) framework [10] and can be constructed with hierarchical segmentation algorithms, *e.g.* HSeg [11].

Let n_1 be a pixel of the image. Through hierarchical image representation, we write n_i the nested image regions at level $i = 2, ..., p_{\max}$, with region at lower levels always being included in higher levels *i.e.* $n_1 \subseteq n_2 ... \subseteq n_{p_{\max}}$. The context information of pixel n_1 can be then described by its ancestor regions n_i at multiple levels $i = 2, ..., p_{\max}$. More specifically, one can define the context information as a sequence $S = \{n_1, ..., n_{p_{\max}}\}$ that encodes the evolution of the pixel n_1 through the different levels of the hierarchy. Each n_i is described by a D-dimensional feature $\boldsymbol{x}_i$ that encodes the region characteristics *e.g.* spectral information, size, shape, etc.

3. SPECTRUM KERNELS

3.1. Definition

The spectrum kernel is an instance of kernels for structured data that allows the computation of similarities between contiguous subsequences of different lengths [7, 8]. Originally designed for symbolic data, we propose here an adaptation to deal with hierarchical representations equipped with numerical features.

Contiguous subsequences can be defined as $s_p = (n_t, n_{t+1} ..., n_{t+p})$, with $t \geq 1, t + p \leq p_{\max}$ and p being the subsequence length. Fig. 2 gives an example of a sequence and enumerates all its subsequences s_p.
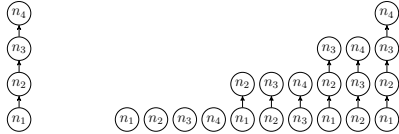
Fig. 2: A sequence S (left) and all its subsequences s_p (right).

The spectrum kernel measures the similarity between two sequences S, S' by summing up kernels computed on all their subsequences. Let $S_p = \{s_p \in S \mid |s_p| = p\}$ be the set of subsequences with a specific length p, the spectrum kernel can be written as:

$$
\begin{aligned}
K(S, S') &= \sum_p \omega_p \, K(S_p, S'_p) \\
&= \sum_p \omega_p \sum_{s_p \in S_p, s'_p \in S'_p} K(s_p, s'_p) \,,
\end{aligned}
\tag{1}
$$

where the p-spectrum kernel $K(S_p, S'_p)$ is computed between the set of subsequences of length p, and is further weighted by parameter ω_p. In other words, it only allows the matching of subsequences with same length. The kernel between two subsequences $K(s_p, s'_p)$ is defined as the product of atomic kernels computed on individual nodes $k(n_{t+i}, n'_{t'+i})$, with i

denoting the position of nodes in the subsequence, following an ascending order $0 \leq i \leq p - 1$:

$$
K(s_p, s'_p) = \prod_{i=0}^{p-1} k(n_{t+i}, n'_{t'+i}) \,.
\tag{2}
$$

$K(S, S')$ in Eq. (1) suffers a common issue for structured kernels: the kernel value highly depends on the length of the sequences, as the number of compared substructures greatly increases with the length of sequence. One can mitigate this problem by normalizing the kernel as:

$$
K^*(S, S') = \frac{K(S, S')}{\sqrt{K(S, S)}\sqrt{K(S', S')}}.
\tag{3}
$$

In the sequel, we only use the normalized version K^* of the kernel (written K for the sake of simplicity).

3.2. Weighting

Several common weighting schemes [8] can be considered:

- $\omega_p = 1$ if $p = q$ and $\omega_p = 0$ otherwise, yielding to a q-spectrum kernel considering only subsequence with a given length q: $K(S, S') = \sum_{s_q \in S_q, s'_q \in S'_q} K(s_q, s'_q)$;

- $\omega_p = 1$ for all p, leading to a constant weighting with all lengths of subsequences;

- $\omega_p = \lambda^p$ with $\lambda \in (0, 1)$, an exponentially decaying weight *w.r.t.* the length of the subsequences.

It should be noted here that when using Gaussian kernel for the atomic kernel

$$
k(n_i, n'_i) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|^2) \,,
\tag{4}
$$

the kernel computed on the stacked vector $\boldsymbol{z} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p_{\max}})$ comes down to the $p_{\max}$-spectrum kernel:

$$
\begin{aligned}
K(s_{p_{\max}}, s'_{p_{\max}}) &= \prod_{i=1}^{p_{\max}} \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|^2) = \\
\exp\left(\sum_{i=1}^{p_{\max}} (-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|^2)\right) &= \exp(-\gamma \|\boldsymbol{z} - \boldsymbol{z}'\|^2) \,.
\end{aligned}
\tag{5}
$$

3.3. Kernel computation

We propose here an efficient computation scheme to iteratively compute all the p-spectrum kernels in a single run, yielding a complexity of $O(p_{\max} p'_{\max})$. The basic idea is to iteratively compute the kernel on subsequences s_p and s'_p using previously computed kernels on subsequences of length $(p - 1)$. The atomic kernel $k(n_i, n'_{i'})$ thus needs to be computed only once, avoiding redundant computing.

We define a three-dimensional matrix M of size $p_{\max} \times p'_{\max} \times \min(p_{\max}, p'_{\max})$, where each element $M_{i,i',p}$ is defined as:

$$M_{i,i',p} = k(n_i, n'_{i'})(M_{i-1,i'-1,p-1}) \ . \tag{6}$$

where $M_{0,0,0} = M_{0,i',0} = M_{i,0,0} = 1$ by convention. The kernel value for the p-spectrum kernel is then computed as the sum of all the matrix elements for a given p:

$$K(S_p, S'_p) = \sum_{i,i'=1}^{p_{\max},p'_{\max}} M_{i,i',p} \ . \tag{7}$$

4. EXPERIMENTS

4.1. Datasets and design of experiments

We conduct experiments on 6 standard hyperspectral image datasets: `Indian Pines`, `Salinas`, `Pavia Centre` and `University`, `Kennedy space center (KSC)` and `Botswana`, considering a *one-against-one* SVM classifier (using the Java implementation of LibSVM [12]).

We use Gaussian kernel as the atomic kernel $k(\cdot, \cdot)$. Free parameters are determined by 5-fold cross-validation over potential values: the bandwidth γ (Eq. (4)) and the SVM regularization parameter C. We also cross-validate the different weighting scheme parameters: $q \in \{1, \ldots, p_{\max}\}$ for the q-spectrum kernel and $\lambda \in (0, 1)$ for the decaying factor.

4.2. Results and analysis

We randomly pick $n = \{10, 25, 50\}$ samples per class from available ground truth for training, and the rest for testing. In the case of small number of pixels per class in `Indian Pines` dataset (total sample size for a class less than $2n$), we use half of samples for training.

Hierarchical image representations are generated with HSeg [11] by increasing the region dissimilarity criterion α. Parameter α is empirically chosen: $\alpha = [2^{-2}, 2^{-1}, ..., 2^8]$, leading to a tree that covers the whole scales from fine to coarse (top levels of whole image are discarded as they do not provide any additional information). Hierarchical levels $\alpha = \{2^2, 2^4, 2^6\}$ of `Indian Pines` are shown in Fig. 1 as the fine, intermediate, coarse level for illustration. Features $\boldsymbol{x}_i$ that describe each region are set as the average spectral information of the pixels that compose the region.

4.2.1. Comparison with state-of-the-art algorithms

We compare our sequence-based kernel with state-of-the-art algorithms that take into account the spatial information relying on multiscale representation of an image: i) spatial-spectral kernel [2] that uses area filtering to obtain the spatial features (the filtering size is fixed so as to lead to the best accuracy); ii) attribute profile [3], using 4 first principal components with automatic level selection for the area attribute

and standard deviation attribute as detailed in [13]; iii) hierarchical features stored on a stacked vector [4, 5, 6]. For comparison purposes, we also report the pixel-based classification overall accuracies. All results are obtained by averaging the performances over 10 runs of (identical among the algorithms) randomly chosen training and test sets.

First of all, in Tab. 1, we can see that the overall accuracies are highly improved when spatial information is included. Using hierarchical features computed over a tree (stacked vector or any version of the spectrum kernel) yields competitive results compared with state-of-the-art methods. By applying the proposed spectrum kernel on the hierarchical features rather than a kernel on a stacked vector, the results are further improved: best results for `Indian Pines`, `Salinas`, `Pavia Centre`, `KSC` and `Botswana` datasets are obtained with a spectrum kernel. We can observe that attribute profiles perform better for `Pavia University`. This might be due to the kind of hierarchical representation used, *i.e.* min and max-trees in the case of attribute profiles instead of HSeg in our case. Besides, the popularity of these profiles as well as the `Pavia` dataset result in optimizations of the scale parameters for years. However, the proposed spectrum kernel is not limited at all to the HSeg representation, and it is thus possible to apply it to min- and max-trees and attribute features. This will be explored in future studies.
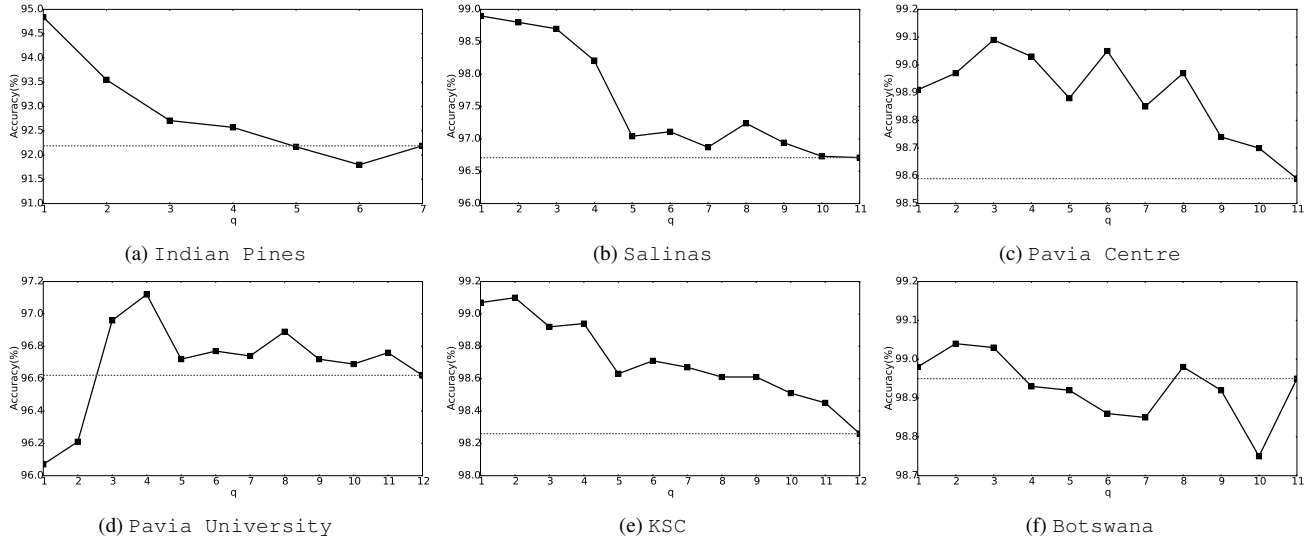
4.2.2. Impact of the weighting scheme

We study the impact of the different weighting schemes introduced in Sec. 3.2. Fig. 3 shows that the stacked vector ($q = p_{\max}$) does not lead to the best performances, and that the best scale q can not be determined beforehand as it depends on the dataset. For most setups, combination of different scales (constant weighting or decaying factor) allows the improvement of the accuracies. However, the best weighting scheme again depends on the considered dataset, and this calls for a more extensive study of weighting strategies.

5. CONCLUSION

In this paper, we propose to use the spectrum kernel for applying machine learning on hierarchical features for hyperspectral image classification. The proposed kernel considers the hierarchical features as a sequence of data and exploits the hierarchical relationship among regions at multiple scales by constructing kernels on various lengths of subsequences. The method exhibits better performances than state-of-the-art algorithms for all but one tested dataset. We also show that combining different scales allows the improvement of the accuracies, but the way to combine them should be further explored. The use of optimal weights thanks to the multiple kernel learning framework [14] is the next step of our work.

Table 1: Mean (and standard deviation) of overall accuracies (OA) computed over 10 repetitions using n training samples per class for 6 hyperspectral image datasets. c stands for constant weighting, q for the q-spectrum kernel and λ for the decaying weight. Best results are boldfaced.

			Indian Pines				
n	pixel only	Spatial-spectral	Attribute profile	Stacked vector	Spectrum kernel-c	Spectrum kernel-q	Spectrum kernel-λ
10	54.89 (2.10)	72.03 (2.52)	64.37 (2.87)	73.21 (2.60)	78.70 (4.88)	80.19 (4.48)	**80.19 (3.40)**
25	66.04 (1.59)	84.02 (1.31)	76.71 (2.60)	84.90 (2.42)	89.16 (2.89)	**91.36 (1.57)**	89.46 (3.61)
50	72.99 (0.10)	90.82 (2.07)	84.57 (1.45)	92.19 (0.86)	94.12 (1.18)	**94.76 (1.09)**	94.48 (1.20)
			Salinas				
n	pixel only	Spatial-spectral	Attribute profile	Stacked vector	Spectrum kernel-c	Spectrum kernel-q	Spectrum kernel-λ
10	83.87 (1.96)	87.72 (1.88)	91.89 (1.73)	89.17 (2.95)	**93.18 (1.70)**	91.16 (2.65)	91.44 (2.71)
25	88.13 (1.22)	92.93 (0.98)	95.99 (1.11)	94.86 (1.58)	**97.28 (1.62)**	97.04 (1.28)	97.02 (1.57)
50	88.86 (1.22)	94.34 (0.81)	97.39 (0.45)	96.71 (0.70)	98.51 (0.89)	**98.81 (0.70)**	97.93 (1.22)
			Pavia Centre				
n	pixel only	Spatial-spectral	Attribute profile	Stacked vector	Spectrum kernel-c	Spectrum kernel-q	Spectrum kernel-λ
10	93.37 (3.59)	95.69 (0.73)	96.03 (0.91)	95.94 (1.01)	96.14 (1.61)	96.56 (1.09)	**96.71 (0.97)**
25	96.13 (0.48)	96.99 (0.48)	97.59 (0.27)	97.85 (0.53)	97.93 (0.55)	**97.96 (0.59)**	97.93 (0.57)
50	96.98 (0.52)	98.10 (0.34)	98.59 (0.24)	98.59 (0.48)	98.83 (0.39)	98.92 (0.37)	**99.04 (0.31)**
			Pavia University				
n	pixel only	Spatial-spectral	Attribute profile	Stacked vector	Spectrum kernel-c	Spectrum kernel-q	Spectrum kernel-λ
10	69.00 (5.68)	76.74 (5.26)	**88.69 (4.06)**	83.30 (3.75)	84.34 (5.14)	84.43 (6.13)	85.10 (6.65)
25	79.81 (1.42)	87.92 (3.36)	**95.17 (1.84)**	92.95 (3.29)	93.70 (2.56)	93.98 (1.91)	93.98 (2.22)
50	84.72 (1.32)	93.27 (1.29)	**97.52 (0.86)**	96.62 (1.06)	97.20 (0.97)	96.76 (1.11)	96.66 (1.84)
			KSC				
n	pixel only	Spatial-spectral	Attribute profile	Stacked vector	Spectrum kernel-c	Spectrum kernel-q	Spectrum kernel-λ
10	86.56 (1.33)	90.96 (2.12)	90.61 (0.63)	92.75 (1.71)	93.98 (1.29)	**94.18 (0.87)**	94.01 (1.15)
25	91.27 (0.84)	97.16 (0.16)	95.53 (0.71)	97.32 (0.45)	**97.85 (0.63)**	97.45 (0.86)	97.82 (0.66)
50	93.67 (0.58)	98.46 (0.29)	97.41 (0.49)	98.26 (0.37)	99.13 (0.34)	99.00 (0.40)	**99.15 (0.23)**
			Botswana				
n	pixel only	Spatial-spectral	Attribute profile	Stacked vector	Spectrum kernel-c	Spectrum kernel-q	Spectrum kernel-λ
10	87.72 (2.42)	92.62 (1.40)	92.17 (1.32)	94.16 (1.41)	**94.66 (1.62)**	94.59 (1.69)	94.63 (1.54)
25	91.89 (0.67)	96.65 (0.69)	95.35 (0.91)	97.71 (0.72)	**97.99 (0.48)**	97.79 (0.55)	97.90 (0.79)
50	94.03 (0.60)	97.74 (0.52)	96.83 (0.64)	98.95 (0.53)	**99.10 (0.50)**	98.97 (0.42)	98.99 (0.45)

(a) `Indian Pines`   (b) `Salinas`   (c) `Pavia Centre`

(d) `Pavia University`   (e) `KSC`   (f) `Botswana`

Fig. 3: The overall accuracies of q-spectrum kernel with different lengths q using $n = 50$ training samples per class. Results for the stacked vector with Gaussian kernel are shown in dashed line.

6. REFERENCES

[1] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and James Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[2] Mathieu Fauvel, Jocelyn Chanussot, and Jon Atli Benediktsson, "A spatial–spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognition*, vol. 45, no. 1, pp. 381–392, 2012.

[3] Mauro Dalla Mura, Jon Atli Benediktsson, Björn Waske, and Lorenzo Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975–5991, 2010.

[4] Lorenzo Bruzzone and Lorenzo Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2587–2600, 2006.

[5] Sébastien Lefèvre, Laetitia Chapel, and François Merciol, "Hyperspectral image classification from multiscale description with constrained connectivity and metric learning," in *6th International Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, 2014.

[6] Lian-Zhi Huo, Ping Tang, Zheng Zhang, and Devis Tuia, "Semisupervised classification of remote sensing images with hierarchical spatial similarity," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 1, pp. 150–154, 2015.

[7] Christina Leslie, Eleazar Eskin, and William Stafford Noble, "The spectrum kernel: A string kernel for SVM protein classification," *Pacific symposium on biocomputing*, vol. 7, pp. 566–575, 2002.

[8] Alexander J. Smola and S.V.N. Vishwanathan, "Fast kernels for string and tree matching," in *Advances in Neural Information Processing Systems*, 2003, pp. 585–592.

[9] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins, "Text classification using string kernels," *The Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[10] Geoffrey J. Hay and Guillermo Castilla, *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*, chapter Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline, pp. 75–89, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[11] James C. Tilton, "Image segmentation by region growing and spectral clustering with a natural convergence criterion," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 1998, vol. 4, pp. 1766–1768.

[12] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[13] Pedram Ghamisi, Jón Atli Benediktsson, Gabriele Cavallaro, and Antonio Plaza, "Automatic framework for spectral–spatial classification based on supervised feature extraction and morphological attribute profiles," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2147–2160, 2014.

[14] Mehmet Gönen and Ethem Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.