Inducing Contextual Classifications With Kernel Functions Into Support Vector Machines

Rogério Galante Negri, Erivaldo Antônio da Silva, and Wallace Casaca

Abstract—**Kernel functions have revolutionized theory and practice in the field of pattern recognition, especially to perform image classification. Besides giving rise to nonlinear variants of the well-known support vector machine (SVM), these functions have also been successfully used to classify nonvectorial data (e.g., graphs and collection of sets), in which customized metrics are created to precisely measure the similarity among such contextual data entities. This letter introduces two context-inspired kernel functions as new SVM-driven methods for remote sensing image classification. In contrast to the existing SVM-based approaches that assume only multiattribute vectors as representative features in a high-dimensional space, the proposed models formally establish comparisons between the entire sets of context-given data, thus employing these contextual measurements to drive the classification. More precisely, stochastic distances as well as hypothesis tests are conveniently handled and "kernelized" to build our models. A complete battery of experiments involving both remote sensing and real-world images is conducted to validate the performance of the proposed kernels against various well-established SVM-based methods.**

Index Terms—**Context, image classification, Kernel functions.**

I. INTRODUCTION

KERNEL functions play an important role in pattern recognition, especially when one intends to accomplish data classification. In fact, this class of functions forms the basis of the nonlinear extensions for several popular linear models, in particular the well-established support vector machine (SVM). In addition to being effective and mathematically well posed, kernel functions also allow the use of their prior linear counterparts on a variety of data classification problems, including such cases whereby no vectorial representations are available (e.g., textual content, collections of sets, and graphs). In summary, kernel operators are seen as a flexible and powerful tool for tuning a certain classification method to fit the input data and not the inverse, as they may lead to alternatives specially designed to obtain more refined results and customizations on the addressed problem [1].

Considering the data classification context, investigations toward achieving more accurate and robust methods remain

challenging. Thus, the inclusion of contextual information into the classification process may lead to new results in terms of accuracy and data readability. There are a few works that exploit this concept, by introducing kernels that are able to embed contextual clues into the classification task. For example, Camps-Valls *et al.* [2] and Gurram and Kwon [3] first paved the way for kernel operators as contextual models to properly address the specific case of hyperspectral image classification.

This letter presents two novel kernels for contextual remotely sensed image classification. The kernels are designed in terms of stochastic distances and statistical hypothesis tests, which take advantage of the robustness and versatility provided by the probability theory. An extensive battery of experiments with remote and nonremote sensing images is conducted to numerically assess and validate the proposed kernel functions. These experiments include a study case with a synthetic aperture radar (SAR) image, a full data set of optical remote sensing scenes, and a standard real-world image. As a baseline, the SVM method is adopted as the kernel machine in our analysis. The kernels are equipped on the SVM and then compared against two contextual SVM-derived methods: Markov random fields and smoothing-based models. Finally, comparisons with classical SVM are also provided.

II. IMAGE CLASSIFICATION, SVM, AND KERNELS

Formally, a classifier is a function $F : \mathcal{X} \rightarrow \mathcal{Y}$ that assigns an element $\mathbf{x}$ from the attribute space $\mathcal{X}$ to a specific class ω_i listed on $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$, $c \in \mathbb{N}^*$, with class labels varying in $\mathcal{Y} = \{1, 2, \ldots, c\}$. It means that $\mathbf{x}$ corresponds to a certain class ω_y, $y \in \mathcal{Y}$, i.e., $y = F(\mathbf{x})$.

Focusing on the image classification problem, the classifier F is evaluated on the attribute vector $\mathbf{x}$ associated with a given pixel s from the target image $\mathcal{I}$, which is defined on a support set $\mathcal{S} \subset \mathbb{N}^2$. While the expression $\mathcal{I}(s) = \mathbf{x}$ states that pixel $s \in \mathcal{S}$ has its attributes characterized by $\mathbf{x} \in \mathcal{X}$, the neighborhood of s can be mathematically represented as

$$\mathcal{V}_\rho(s) = \{t \in \mathcal{S} : 0 \leq md(s, t) \leq \rho\} \qquad (1)$$

where ρ accounts for the *neighborhood influence radius*, and $md(s, t) = \max\{|s_1 - t_1|, |s_2 - t_2|\}$, with $s = (s_1, s_2)$, $t = (t_1, t_2)$ being the spatial positions of pixels s and t. Note that the set $\mathcal{V}_\rho(s)$ allows for incorporating contextual information into a given classification model, as the resulting classification will be induced by the attribute vectors $\mathbf{x}_i$ so that $\mathcal{I}(t) = \mathbf{x}_i$, $t \in \mathcal{V}_\rho(s)$. In practice, the context of s will be conveyed by $\mathcal{V}_\rho(s)$. Therefore, methods considered as "contextual" are those ones that formally embed the context of the image pixels into the classification pipeline, i.e., the

neighborhood structures, which are represented as a full set of pixels, not properly as a vector.

Techniques devoted to classifying images vary in many different aspects as to define $F : \mathcal{X} \to \mathcal{Y}$ as well as to evaluate F on $\mathcal{I}$. For example, methods that rely on supervised learning take information from a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \ldots, c\}$. The mapping created by F between $\mathcal{X}$ and $\mathcal{Y}$ represents the knowledge acquired from $\mathcal{D}$. A good representative of this kind of approach is the well-known SVM, which has become popular within the remote sensing community [4]. This method accomplishes data classification by computing a hyperplane with a larger separating margin. Such a hyperplane is given by the geometric place for which the following function is set equal to zero:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + \text{bias} \qquad (2)$$

where $\mathbf{w}$ represents an orthogonal vector to the hyperplane and the quotient $|\text{bias}|/\|\mathbf{w}\|$ accounts for the distance between the origin of the attribute space and the hyperplane. The tuning parameters $\mathbf{w}$ and bias are obtained by solving an optimization problem built from the training set $\mathcal{D}$. For a detailed discussion concerning the formality and training issues of SVM, see [5].

An attractive manner to increase the assertiveness of SVM-derived classifications is to embed the input patterns (i.e., feature vectors) into a more appropriate space with better separability. Indeed, this can be implicitly done by applying kernel functions. Kernels are used to modify the inner product between input patterns in (2), acting directly on the corresponding optimization problem to train the SVM [5]. Furthermore, they also enable the SVM and other existing kernel methods to address problems, wherein the input data are not necessarily modeled as a vectorial representation (e.g., textual content, arbitrary sets, and graphs).

In more mathematical terms, $K : \mathcal{X}^2 \to \mathbb{R}$ is called a kernel function if K is symmetric with right to the input elements and satisfies Mercer's theorem conditions [1]. Despite its versatility and well posedness, defining new kernels while satisfying Mercer's conditions is not a straightforward task in practice. An alternative for successfully creating a kernel is to take general models of kernels, for example, the radial basis function (RBF) [1]

$$K(\mathbf{x}_u, \mathbf{x}_v) = g(d(\mathbf{x}_u, \mathbf{x}_v)) \qquad (3)$$

where $d : \mathcal{X}^2 \to \mathbb{R}$ is a metric and $g : \mathbb{R} \to \mathbb{R}$ is a strictly positive function.

The study and development of more appropriate kernels for remotely sensed image classification have attracted considerable attention over the past two decades. A pioneering study concerning contextual classifications driven by kernel functions is presented in [2], and later improved upon in [3]. Both methodologies integrate the spatial context, delimited by an influence radius, with an average-guided kernel operator. As a result, their outputs tend to produce excessive blurring, similar to the use of smoothing filters. Still on kernel models, Kondor and Jebara [6] combine the concept of stochastic distance and kernels, being their approach later extended in [7] to properly cope with region-based classifications on remote sensing images.

This letter proposes two novel kernel operators which conceptually embed the notion of measuring contextual data

into the image classification procedure. The kernels bear many attractive features such as the ability to promote context-guided classifications, solid mathematical foundation, and good adaptability to fit the SVM onto the available data.

In contrast to the existing SVM-based methods that only assume the input data as the set of multiattribute vectors in a vector space, the designed kernels formally establish comparisons between sets of abstract data instances in the sense of a metric space. In other words, the neighborhood structures of the pixels are interpreted as statistic models, not necessarily as a well-structured feature vector. These models are provided as input to the SVM, enabling the classifier to perform contextual classifications while still exploiting the spatial variability of the pixels neighborhoods in a deeper and more effective way. As a result, the classification tends to be more precise and discriminative, since the data to be labeled are represented in terms of their local neighborhood patterns.

III. INDUCING CONTEXTUAL ANALYSIS INTO KERNELS

A. Jeffries-Matusita Kernel

A simple and effective manner to describe neighborhood content as a valid contextual information is to consider such content as data instances. Consequently, the acquired data will no longer behave as usual, due to the absence of a regular set of attribute points $\mathcal{X}$ that formally represents a pixel neighborhood (i.e., collection of vectors $\mathcal{V}_\rho(s)$) as a single feature vector. To deal with this criticism, one may set $d(\cdot, \cdot)$ as a stochastic distance in (3) so that comparisons between neighborhoods $\mathcal{V}_\rho(s_i)$ and $\mathcal{V}_\rho(s_j)$ would make sense.

Stochastic distances are viewed in probability theory as a powerful tool to quantify the separability between different sets of data, therein measuring how far apart their corresponding probability distributions are from each other [8]. Assuming that $\mathcal{V}_\rho(s_i)$ and $\mathcal{V}_\rho(s_j)$ satisfy multivariate Gaussian distributions, the so-called *Jeffries-Matusita* measure (JM) [9] was taken as a stochastic distance in (3). More precisely, the similarity between the neighbors of s_i and s_j is computed as follows:

$$JM(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j)) = 2(1 - e^{-B(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j))}) \qquad (4)$$

where $B(\cdot, \cdot)$ is the *Bhattacharyya* distance between two multivariate Gaussian models [5].

Notice that the straight substitution of the JM function into (3) does not lead to a genuine kernel because the triangular inequality may not be strictly verified, thus hampering the use of (4) as an authentic metric. To validate this requirement, a sufficiently large constant is conventionally added to the JM expression when the inputs $\mathcal{V}_\rho(s_i)$ and $\mathcal{V}_\rho(s_j)$ are structurally different. Once the values of JM are bounded by $[0, 2]$, the constant 2 arises as a suitable choice to be incorporated into (4), thereby establishing the definitive JM kernel function as follows:

$$K_{JM}(\mathcal{V}(s_i), \mathcal{V}(s_j)) = e^{-\gamma \overline{JM}(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j))} \qquad (5)$$

where $\gamma \in \mathbb{R}_+$ is a regularization factor and $\overline{JM}$ is given as

$$\overline{JM}(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j)) = \begin{cases} 0, & \text{if } \mathcal{V}_\rho(s_i) = \mathcal{V}_\rho(s_j), \\ JM(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j)) + 2, & \text{o\textbackslash w.} \end{cases} \qquad (6)$$

B. Hypothesis Testing Kernel

Following Section III-A, where neighborhoods are also inputs and (3) is taken to originate a kernel function, the semantic data contained in $\mathcal{V}_\rho(s_i)$ match the valid content in $\mathcal{V}_\rho(s_j)$ by computing a p-value from some hypothesis tests. Among the various well-known hypothesis tests, a convenient choice is Student's t-test, because it compares the mean values of two data samples with distinct variances.

Let $p_{st}(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j); k)$ be a function that returns p-values from a Student's t-test so that these values are used to compare the neighborhood structures of $\mathcal{V}_\rho(s_i)$ and $\mathcal{V}_\rho(s_j)$. The constant k denotes the kth band of the image $\mathcal{I}$, while b sets the total number of bands in $\mathcal{I}$. Then, the following similarity measure can be formulated:

$$P(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j)) = \frac{1}{b} \sum_{k=1}^{b} (1 - p_{st}(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j); k)). \quad (7)$$

Notice from the p-value of the Student's test that values closer to 1 indicate greater correlations between neighborhoods. In addition, to precisely measure the similarity with p-values, the quantity $(1 - p_{st}(\cdot, \cdot; \cdot))$ is intentionally accommodated within 7. This guarantees that the identity property will hold, that is, identical inputs will return a null distance when measured. Since $P(\cdot, \cdot)$ is symmetric and it satisfies the identity property, function (7) is indeed a distance. Therefore, as in Section III-A, $P(\cdot, \cdot)$ is redesigned as a metric by summing constant 1 when $\mathcal{V}_\rho(s_i) \neq \mathcal{V}_\rho(s_j)$.

Finally, taking the radial basis equation (3) as premise, our Student's t-test-based kernel operator is defined as

$$K_{ST}(\mathcal{V}(s_i), \mathcal{V}(s_j)) = e^{-\gamma \overline{P}(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j))} \quad (8)$$

where $\gamma \in \mathbb{R}_+$ is a tuning constant and $\overline{P}$ is the metric as follows:

$$\overline{P}(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j)) = \begin{cases} 0, & \text{if } \mathcal{V}_\rho(s_i) = \mathcal{V}_\rho(s_j), \\ 2 - P(\mathcal{V}_\rho(s_i), \mathcal{V}_\rho(s_j)), & \text{otherwise.} \end{cases} \quad (9)$$

IV. EXPERIMENTS AND RESULTS

The performance of our context-inspired kernels is assessed through an expressive set of comparisons involving three well-established SVM classification methods on three application scenarios. The first one aims at classifying a full data set of very high-resolution optical remote sensing images. Then, a realistic case study for land use/cover classification from a radar image is discussed as part of our comparative analysis. Finally, experiments run on a typical real-world image are also provided to conclude this section. Implementation aspects such as the training and test (ground truth) samples, as well as the classes used in our experiments are shown in Table I.

The outputs computed by the pair of methods SVM "plus" JM and Student's t-test kernels are named here as SVM+KJM and SVM+KST, while the acronyms SVM+ICM and SVM+Mode denote the context-based algorithms of SVM "plus" Markov random fields (with the iterated conditional modes algorithm) [10] and the statistical mode filter. SVM in a basic pixel-based version is also included as a baseline in our tests. The ρ parameter in (1) is empirically taken as 1–3, resulting in spatial windows of sizes 3×3, 5×5, and 7×7 for the context-based methods. The "one-against-all" multiclass

TABLE I
NUMBER OF PIXELS FOR TRAINING AND TEST SAMPLES WITH RESPECT TO LULC CLASSES FOR THE SIRI-WHU DATABASE, TAPAJÓS AREA, AND LENA

| | Data sets | | Tapajós | | Lena |
Class	Train/Test†	Class	Train/Test	Class	Train/Test
LAND	160/6000	AG	512/14395	SK	232/4016
NONL	160/6000	BS	512/14207	HT	348/4909
-	-	PF	768/14136	FE	232/4312
†Estim. average value		PS	768/14/678	HA	232/1903

strategy [5] is adopted in all evaluated methods, while the RBF is applied to "kernelize" the pixel-based SVM and the variants SVM+Mode and SVM+ICM.

To numerically validate the results, the kappa coefficient is used as a quality measure. This metric gauges the agreement level between the resulting classification and the expected labels with respect to a set of test samples with ground-truth-labeled areas [11]. The kappa variances are also tabulated, in order to properly conduct hypothesis testing. The results of the statistical analysis are generated in the sense of the bilateral confidence interval of 95% to compare the kappa values.

Finally, to train the methods and maximize their classification accuracies, convenient choices are made toward determining optimized values for C and γ. The selection is performed by applying an exhaustive grid search process with 10-fold cross validation for the parameters that yield the best accurate results with respect to the training samples, as in [12]. Motivated by [12], C ranges over $\{1, 10, 100, 1k, 10k\}$, while the kernel parameter $\gamma \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$. The *number of iterations* and *minimum percentage change* are estimated in a similar manner.

A. Classifying Full Collections of Aerial Images

We first run all the evaluated techniques on a comprehensive data set of remote sensing images. This collection of images, called *SIRI-WHU data set*, contains more than 2400 aerial scenes made publicly available by Zhao *et al.* [13]. The full data set offers a dozen land classes covering different urban areas in China. Therefore, we take this data set as a valid benchmark to verify the assertiveness rate of the methods for two representative land classes: *industrial* and *coastal zones* (see the frames illustrating both zones in Fig. 1).

Table II summarizes the kappa values and their variances for SVM, SVM+ICM, SVM+Mode, and our techniques, SVM+KJM and SVM+KST, for the above-described land classes of aerial scenes. One can observe from the tabulated scores that the SVM+KJM variant outperforms the existing context-based methods in all the measurements, attesting to its accuracy in addressing the entire sets of aerial images such as the ones found in [13]. Next, SVM+KST achieves the second best scores for both the kappa and variance measurements, followed by SVM+Mode and SVM+ICM.

One can note that the quality increment is more expansive when the neighborhood size (#Neigh.) comes up from 3×3 to 5×5. Another observed aspect is that this size enlargement implies in better classifications, but this is not always a learned behavior, as reported in the bottommost rows of Table II.
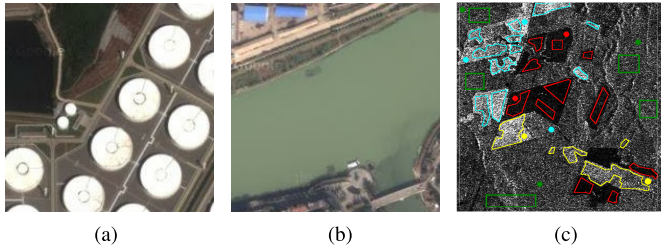
Fig. 1. (a) and (b) Two illustrative aerial scenes taken from SIRI-WHU database. (c) Examined areas and LULC samples (■ AG, ■ BS, ■ PF, and ■ PS) used for training and accuracy evaluations in our case study with a SAR image.

TABLE II
QUANTITATIVE COMPARISON OF CONTEXTUAL SVM-BASED METHODS WHEN ASSESSED ON THE REMOTE SENSING DATA SET [13]. BOLD AND GRAY VALUES INDICATE THE BEST AND SECOND BEST SCORES

		Industrial land class		Coastal land class	
	#Neigh.	Kappa	Variance	Kappa	Variance
SVM	-	0.6609	0.2231	0.7540	0.1308
SVM+ICM	-	0.8606	0.1907	0.9129	0.1155
SVM+Mode	3×3	0.8348	0.1990	0.8947	0.1185
	5×5	0.8701	0.1934	0.9177	0.1219
	7×7	0.8816	0.1903	0.9236	0.1287
SVM+KJM	3×3	0.8928	0.1630	0.9366	0.0738
	5×5	0.9389	0.1380	0.9636	0.0581
	7×7	**0.9503**	**0.1240**	**0.9793**	**0.0314**
SVM+KST	3×3	0.8734	0.1696	0.9218	0.0880
	5×5	0.8874	0.1758	0.9415	0.0827
	7×7	0.8929	0.2036	0.9262	0.1214

B. Case Study Description and Validation

A case study covering a multiclass classification case using a SAR image captured by the ALOS-PALSAR sensor is given next. This SAR image, acquired on April 23, 2007, at 600×600 pixels and with a 20-m resolution after applying a 3×3 multilook process in HH and HV amplitude polarization, corresponds to an area of the Amazon situated in the south of Santarém, state of Pará, Brazil, more precisely around the Tapajós National Forest. Its area's geographic coordinates are $3^o8'20''$ S and $54^o55'33''$ W. Field work conducted in September 2009 identified the following land use and land cover (LULC) types: Primary Forest (PF), Pasture (PS), Bare Soil (BS), and Agriculture (AG) [see Fig. 1(c), where the polygonal regions indicate the test samples and colored points the training sets].

1) SAR Image Classification (Quantitative Assessments):
From the kappa and variance values listed in Table III, the SVM+KJM technique delivers the best scores, followed by SVM+KST. Indeed, the SVM+KJM and SVM+KST variants differ substantially from the other for both quality measures. For instance, the kappa achieved by the pair SVM+KST for the minimum neighborhood size of 3×3 matches the kappa obtained by SVM+ICM and SVM+Mode with a 5×5 spatial window. Concerning hypothesis testing, the SVM+Mode results under the sizes of 5×5 and 7×7 are statistically equivalent. Similar conclusions can be verified from the bottommost rows of the SVM+KST variant. Finally, one can check that the accuracies of SVM+KJM, SVM+KST, and SVM+Mode rise as the neighborhood dimensions increase.

TABLE III
QUANTITATIVE COMPARISON OF SVM-BASED METHODS WHEN APPLIED TO THE SAR IMAGE OF TAPAJÓS AREA AND LENA'S PICTURE. BOLD AND GRAY VALUES INDICATE THE BEST AND SECOND BEST SCORES

		Tapajós area		Lena's picture	
	#Neigh.	Kappa	Var. (10^{-5})	Kappa	Var. (10^{-5})
SVM	-	0.4918	0.6740	0.4506	0.2668
SVM+ICM	-	0.6265	0.5628	0.5390	0.2235
SVM+Mode	3×3	0.5691	0.5924	0.5317	0.2359
	5×5	0.5951	0.5721	0.5425	0.2224
	7×7	0.6022	0.5677	0.5363	0.2194
SVM+KJM	3×3	0.7623	0.4494	0.6441	0.2292
	5×5	0.8680	0.2751	0.7520	0.1769
	7×7	**0.9044**	**0.2057**	**0.8039**	**0.1458**
SVM+KST	3×3	0.7086	0.5260	0.5856	0.2325
	5×5	0.7795	0.4200	0.6544	0.2207
	7×7	0.7763	0.4244	0.6364	0.2270

2) SAR Image Classification (Qualitative Assessments):
Fig. 2 portrays for the investigated area, the visual results obtained by the techniques when the neighborhood is fixed to 5×5. Although the context-guided methods produce better (noiseless) partitions when compared with the SVM pixelwise approach, the use of the proposed JM kernel leads to clearer characterization of the PS and AG segments, as its output is more similar to the ground-truth classifications outlined in Fig. 1(c). Note also that, similar to the SVM+KJM method, the pair SVM+KST produces satisfactory classifications, even for the PS and AG patterns.

C. Contextual Classifications on Real-Word Images

Fig. 3 illustrates the capability of SVM+KJM and SVM+KST methods to classify images that commonly appear in everyday life, even for images that have been contaminated by noise. So, to establish a valid benchmark to assess the classification results, the segments identified and annotated as classes in the Lena's picture [see Fig. 3(a)] are: SK (skin), HT (hat), FE (feather) and HA (hair).

Visually inspecting the results, one can see that both the KJM and KST kernels achieve more accurate and consistent outcomes, mainly regarding the assertiveness level for the FE and SK classes, as indicated by the large areas demarcated by blue and red pixels as shown in Fig. 3(e) and (f). This is also confirmed when one visualizes the quality measures listed in Table III, where the kappas for SVM+KJM and SMV+KST are greater than those obtained by the evaluated methods in almost all the comparative scenarios with different neighborhoods.

D. Computational Aspects and Timings

To perform our experiments, an Intel *Core i7* processor with 16 GB of RAM running a Linux operating system has been used. The implementations were coded using IDL (Interactive Data Language) programming language.

The computational timings of the examined algorithms are reported in Fig. 4. Although the SVM+KJM and SVM+KST methods produce more accurate classifications, they also take longer than the others. To address this, more efficient implementations (e.g., GPU and parallel architectures) can be used to alleviate the computational burden of the algorithms.
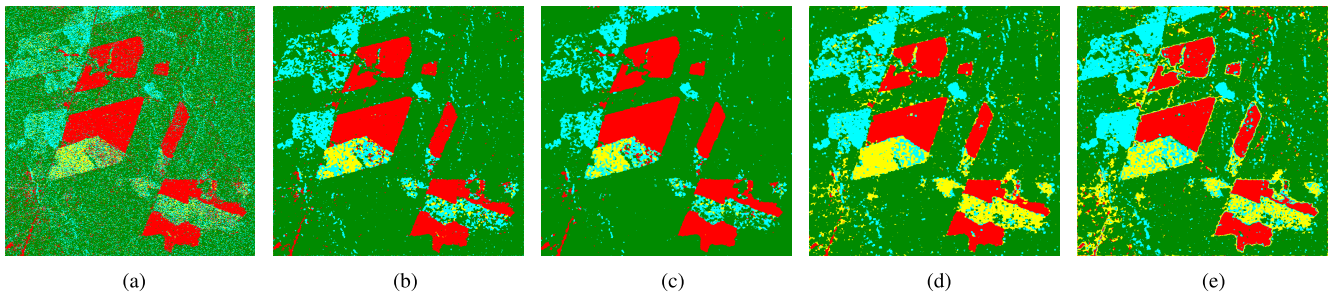
Fig. 2.   Classification results obtained from the evaluated methods when applied to the Tapajós study area. The classes are ▮ AG, ▮ BS, ▮ PF, and ▮ PS. (a) SVM. (b) SVM+ICM. (c) SVM+Mode (5×5). (d) SVM+KJM (5×5). (e) SVM+KST (5×5).
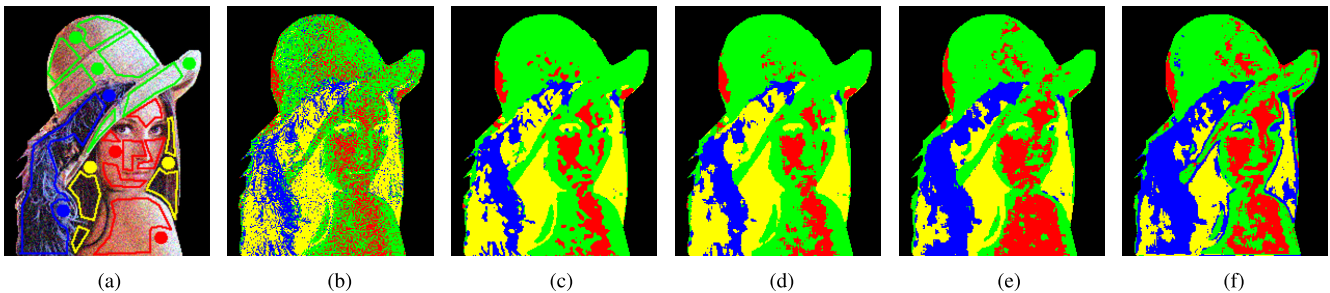

Fig. 3.   Classification results obtained from the evaluated methods on a real-world picture (Lena's picture). The classes are ▮ SK, ▮ HT, ▮ FE, and ▮ HA. (a) Training/test samples. (b) SVM. (c) SVM+ICM. (d) SVM+Mode (5×5). (e) SVM+KJM (5×5). (f) SVM+KST (5×5).
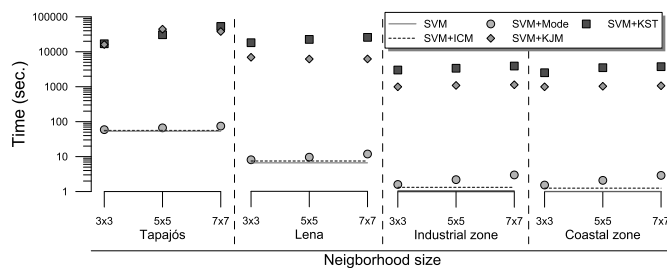

Fig. 4.   Computational timings for all the evaluated methods.

V. CONCLUSION

This letter proposes two context-inspired kernel functions as new SVM methods for remote sensing image classification. In contrast to the existing SVM-based approaches that exploit the pixel context as an attribute vector, the proposed JM and Student's t-test kernels allow for computing the similarity between the entire sets of neighborhoods, thus leading to the use of such contextual patterns to drive the classification task. Besides implicitly embedding contextual data into kernel functions, the SVM+KJM and SVM+KST techniques are found to be robust when dealing with complex aerial images. This behavior can be verified from the experiments involving full collections of aerial scenes and the case study with an Amazon's area, where both kernels performed better than others, thus achieving high scores while still making the regions classified in the images more noticeable. In summary, flexibility and high accuracy render the proposed methods two very attractive image classification approaches in the context of remote sensing.

As future work, parallel architectures will be used to reduce the cost of the methods. Another goal is to exploit other statistical tests to address different applications in this context.

REFERENCES

[1] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2002.

[2] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[3] P. Gurram and H. Kwon, "Contextual SVM using Hilbert space embedding for hyperspectral classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1031–1035, Sep. 2013.

[4] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[5] A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition*, 3rd ed. New York, NY, USA: Wiley, 2011.

[6] R. Kondor and T. Jebara, "A kernel between sets of vectors," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 1–8.

[7] R. G. Negri, L. V. Dutra, S. J. S. Sant'Anna, and D. Lu, "Examining region-based methods for land cover classification using stochastic distances," *Int. J. Remote Sens.*, vol. 37, no. 8, pp. 1902–1921, 2016.

[8] L. Castañeda, V. Arunachalam, and S. Dharmaraja, *Introduction to Probability and Stochastic Processes With Applications*. New York, NY, USA: Wiley, 2012.

[9] J. A. Richards, *Remote Sensing Digital Image Analysis*, 5th ed. Berlin, Germany: Springer, 2013.

[10] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," in *Pattern Recognition and Machine Intelligence* (Lecture Notes in Computer Science), vol. 3776. Berlin, Germany: Springer, 2005, pp. 260–265.

[11] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data*. Boca Raton, FL, USA: CRC Press, 2009.

[12] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, Res. Paper, 2016. [Online]. Available: http://www.mdpi.com/2072-4292/8/2/157

[13] B. Zhao, Y. Zhong, L. Zhang, and B. Huang, "The Fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.*, vol. 8, no. 2, pp. 157–177, 2016. [Online]. Available: http://www.mdpi.com/2072-4292/8/2/157