Dense Bag-of-Temporal-SIFT-Words for Time Series Classification

Adeline Bailly, Simon Malinowski, Romain Tavenard, Thomas Guyet, Laetitia Chapel

Dense Bag-of-Temporal-SIFT-Words for Time Series Classification

Adeline Bailly[1,4], Simon Malinowski[2], Romain Tavenard[1], Laetitia Chapel[3], and Thomas Guyet[4]

[1] Université de Rennes 2, IRISA, LETG-Rennes COSTEL, Rennes, France
[2] Université de Rennes 1, IRISA, Rennes, France
[3] Université de Bretagne-Sud, IRISA, Vannes, France
[4] Agrocampus Ouest, IRISA, Rennes, France

Abstract. The SIFT framework has shown to be accurate in the image classification context. In [1], we designed a Bag-of-Words approach based on an adaptation of this framework to time series classification. It relies on two steps: SIFT-based features are first extracted and quantized into words; histograms of occurrences of each word are then fed into a classifier. In this paper, we investigate techniques to improve the performance of Bag-of-Temporal-SIFT-Words: dense extraction of keypoints and normalization of Bag-of-Words histograms. Extensive experiments show that our method significantly outperforms most state-of-the-art techniques for time series classification.

Keywords: time series classification, Bag-of-Words, SIFT, dense features, BoTSW, D-BoTSW

1 Introduction

Classification of time series has received an important amount of interest over the past years due to many real-life applications, such as medicine [24], environmental modeling [7], speech recognition [12]. A wide range of algorithms have been proposed to solve this problem. One simple classifier is the k-nearest-neighbor (kNN), which is usually combined with Euclidean Distance (ED) or Dynamic Time Warping (DTW) similarity measure. The combination of the kNN classifier with DTW is one of the most popular method since it achieves high classification accuracy [20]. However, this method has a high computation cost which makes its use difficult for large-scale real-life applications.

Above-mentioned techniques compute similarity between time series based on point-to-point comparisons. Classification techniques based on higher level structures (*e.g.* feature vectors) are most of the time faster, while being at least as accurate as DTW-based classifiers. Hence, various works have investigated the extraction of local and global features in time series. Among these works, the Bag-of-Words (BoW) approach (also called Bag-of-Features) consists in representing documents using a histogram of word occurrences. It is a very common technique in text mining, information retrieval and content-based image retrieval

because of its simplicity and performance. For these reasons, it has been adapted to time series data in some recent works [2, 3, 14, 21, 24]. Different kinds of features based on simple statistics, computed at a local scale, are used to create the words.

In the context of image retrieval and classification, scale-invariant descriptors have proved their accuracy. Particularly, the Scale-Invariant Feature Transform (SIFT) framework has led to widely used descriptors [17]. These descriptors are scale and rotation invariant while being robust to noise. In [1], we built on this framework to design a BoW approach for time series classification where words correspond to quantized versions of local features. Features are built using the SIFT framework for both detection and description of the keypoints. This approach can be seen as an adaptation of [22], which uses SIFT features associated with visual words, to time series. In this paper, we improve our previous work by applying enhancement techniques for BoW approaches, such as dense extraction and BoW normalization. To validate this, we conduct extensive experiments on a wide range of datasets.

This paper is organized as follows. Section 2 summarizes related work, Section 3 describes the proposed Bag-of-Temporal-SIFT-Words (BoTSW) method and its improved version (dense extraction and BoW normalization, D-BoTSW), and Section 4 reports experimental results. Finally, Section 5 concludes and discusses future work.

2 Related work

Our approach for time series classification builds on two well-known methods in computer vision: local features are extracted from time series using a SIFT-based approach and a global representation of time series is produced using Bag-of-Words. This section first introduces state-of-the-art distance-based methods in time series classification and then presents previous works that make use of Bag-of-Words approaches for time series classification.

2.1 Distance-based time series classification

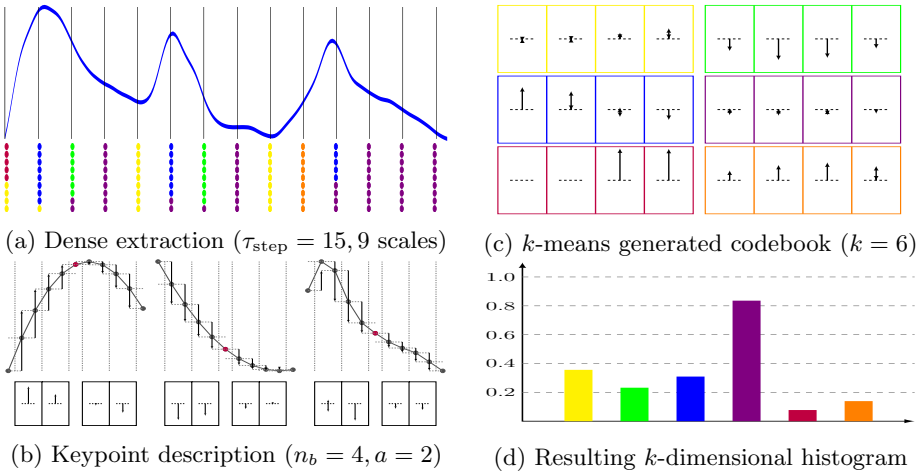
Data mining community has, for long, investigated the field of time series classification. Early works focus on the use of dedicated similarity measures to assess similarity between time series. In [20], Ratanamahatana and Keogh compare Dynamic Time Warping to Euclidean Distance when used with a simple kNN classifier. While the former benefits from its robustness to temporal distortions to achieve high accuracy, ED is known to have much lower computational cost. Cuturi [5] shows that, although DTW is well-suited to retrieval tasks since it focuses on the best possible alignment between time series, it fails at precisely quantifying dissimilarity between non-matching sequences (which is backed by the fact that DTW-derived kernel is not positive definite). Hence, he introduces the Global Alignment Kernel that takes into account all possible alignments in order to produce a reliable similarity measure to be used at the core of standard kernel

methods such as Support Vector Machines (SVM). Lines and Bagnall [15] propose an ensemble classifier based on elastic distance measures (including DTW), named Proportional Elastic Ensemble (PROP). Instead of building classification decision on similarities between time series, Ye and Keogh [26] use a decision tree in which the partitioning of time series is performed with respect to the presence (or absence) of discriminant sub-sequences (named shapelets) in the series. Though accurate, the method is very computational demanding as building the decision tree requires one to check for all candidate shapelets. Douzal and Amblard [6] define a dedicated similarity measure for time series which is then used in a classification tree.

2.2 Bag-of-Words for time series classification

Inspired by text mining, information retrieval and computer vision communities, recent works have investigated the use of Bag-of-Words for time series classification [2, 3, 14, 21, 24]. These works are based on two main operations: converting time series into Bag-of-Words, and building a classifier upon this BoW representation. Usually, standard techniques such as random forests, SVM, neural networks or kNN are used for the classification step. Yet, many different ways of converting time series into Bag-of-Words have been introduced. Among them, Baydogan *et al.* [3] propose a framework to classify time series denoted TSBF where local features such as mean, variance and extremum values are computed on sliding windows. These features are then quantized into words using a codebook learned by a class probability estimate distribution. In [24], discrete wavelet coefficients are extracted on sliding windows and then quantized into words using k-means. In [14, 21], words are constructed using the Symbolic Aggregate approXimation (SAX) representation [13] of time series. SAX symbols are extracted from time series and histograms of n-grams of these symbols are computed to form a Bag-of-Patterns (BoP). In [21], Senin and Malinchik combine SAX with Vector Space Model to form the SAX-VSM method. In [2], Baydogan and Runger design a symbolic representation of multivariate time series (MTS), called SMTS, where MTS are transformed into a feature matrix, whose rows are feature vectors containing a time index, the values and the gradient of time series at this time index (on all dimensions). Random samples of this matrix are given to decision trees whose leaves are seen as words. A histogram of words is output when the different trees are learned.

Local feature extraction has been investigated for long in the computer vision community. One of the most powerful local feature for image is SIFT [17]. It consists in detecting keypoints as extremum values of the the Difference-of-Gaussians (DoG) function and describing their neighborhoods using histograms of gradients. Xie and Beigi [25] use similar keypoint detection for time series. Keypoints are then described by scale-invariant features that characterize the shapes surrounding the extremum. In [4], extraction and description of time series keypoints in a SIFT-like framework is used to reduce the complexity of DTW: features are used to match anchor points from two different time series and prune the search space when searching for the optimal path for DTW.



(a) Dense extraction ($\tau_{\text{step}} = 15, 9$ scales)

(c) k-means generated codebook ($k = 6$)

(b) Keypoint description ($n_b = 4, a = 2$)

(d) Resulting k-dimensional histogram

Fig. 1: Approach overview: (a) A time series and its dense-extracted keypoints. (b) Keypoint description is based on the time series filtered at the scale at which the keypoint is extracted. Descriptors are quantized into words. (c) Codewords obtained *via* k-means, the color is associated with the dots under each keypoint in (a). (d) Histograms of word occurrences are given to a classifier (linear SVM) that learns boundaries between classes. Best viewed in color.

In this paper, we build upon BoW of SIFT-based descriptors. We propose an adaptation of SIFT to mono-dimensional signals that preserves their robustness to noise and their scale invariance. We then use BoW to gather information from many local features into a single global one.

3 Bag-of-Temporal-SIFT-Words (BoTSW) method

The proposed method is based on three main steps: (i) extraction of keypoints in time series, (ii) description of these keypoints by gradient magnitude at a specific scale and (iii) representation of time series by a BoW, where words correspond to quantized version of the description of keypoints. These steps are depicted in Fig. 1 and detailed below.

3.1 Keypoints extraction in time series

The first step of our method consists in extracting keypoints in time series. Two approaches are described here: the first one is based on scale-space extrema detection (as in [1]) and the second one proposes a dense extraction scheme.

Scale-space extrema detection. Following the SIFT framework, keypoints in time series are detected as local extrema in terms of both scale and (temporal)

location. These scale-space extrema are identified using a DoG function, and form a list of scale-invariant keypoints. Let $L(t, \sigma)$ be the convolution ($*$) of a Gaussian function $G(t, \sigma)$ of width σ with a time series $S(t)$:

$$L(t, \sigma) = G(t, \sigma) * S(t) \tag{1}$$

where $G(t, \sigma)$ is defined as

$$G(t, \sigma) = \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-t^2/2\sigma^2}. \tag{2}$$

Lowe [16] proposes the Difference-of-Gaussians (DoG) function to detect scale-space extrema in images. Adapted to time series, a DoG function is obtained by subtracting two time series filtered at consecutive scales:

$$D(t, \sigma) = L(t, k_{\text{sc}}\sigma) - L(t, \sigma), \tag{3}$$

where k_{sc} is a parameter of the method that controls the scale ratio between two consecutive scales.

Keypoints are then detected at time index t in scale j if they correspond to extrema of $D(t, k_{\text{sc}}^j \sigma_0)$ in both time and scale, where σ_0 is the width of the Gaussian corresponding to the reference scale. At a given scale, each point has two neighbors: one at the previous and one at the following time instant. Points also have neighbors one scale up and one scale down at the previous, same and next time instants, leading to a total of eight neighbors. If a point is higher (or lower) than all of its neighbors, it is considered as an extremum in the scale-space domain and hence a keypoint of S.

Dense extraction. Previous researches have shown that accurate classification could be achieved by using densely extracted local features [10, 23]. In this section, we present the adaptation of this setup to our BoTSW scheme. Keypoints selected with dense extraction no longer correspond to extrema but are rather systematically extracted at all scales every τ_{step} time steps on Gaussian-filtered time series $L(\cdot, k_{\text{sc}}^j \sigma_0)$.

Unlike scale-space extrema detection, regular sampling guarantees a minimal amount of keypoints per time series. This is especially crucial for smooth time series from which very few keypoints are detected when using scale-space extrema detection. In addition, even if the densely extracted keypoints are not scale-space extrema, description of these keypoints (cf. Section 3.2) covers the description of scale-space extrema if τ_{step} is not too large. This usually leads to more robust global descriptors.

A dense extraction scheme is represented in Fig. 1, where we consider a step of $\tau_{\text{step}} = 15$ for the sake of readability. In the following, when dense extraction is performed, we will refer to our method as D-BoTSW (for dense BoTSW).

3.2 Description of the extracted keypoints

Next step in our process is the description of keypoints. A keypoint at time index t and scale j is described by gradient magnitudes of $L(\cdot, k_{\text{sc}}^j \sigma_0)$ around t. To do

so, n_b blocks of size a are selected around the keypoint. Gradients are computed at each point of each block and weighted using a Gaussian window of standard deviation $\frac{a \times n_b}{2}$ so that points that are farther in time from the detected keypoint have lower influence. Then, each block is described by two values: the sum of positive gradients and the sum of negative gradients. Resulting feature vector is hence of dimension $2 \times n_b$.

3.3 Bag-of-Temporal-SIFT-Words for time series classification

The set of all training features is used to learn a codebook of k words using k-means clustering. Words represent different local behaviors in time series. Then, for a given time series, each feature vector is assigned the closest word in the codebook. The number of occurrences of each word in a time series is computed. (D-)BoTSW representation of a time series is the ℓ_2-normalized histogram (*i.e.* frequency vector) of word occurrences.

Bag-of-Words normalization. Dense sampling on multiple Gaussian-filtered time series provides considerable information to process. It also tends to generate words with little informative power, as stop words do in text mining applications. In order to reduce the impact of those words, we compare two normalization schemes for BoW: Signed Square Root normalization (SSR) and Inverse Document Frequency normalization (IDF). These normalizations are commonly used in image retrieval and classification based on histograms [8, 9, 19, 22].

Jégou *et al.* [9] and Perronin *et al.* [19] show that reducing the influence of frequent codewords before ℓ_2 normalization could be profitable. They apply a power $\alpha \in [0, 1]$ on their global representation. SSR normalization corresponds to the case where $\alpha = 0.5$, which leads to near-optimal results [9, 19].

IDF normalization also tends to lower the influence of frequent codewords. To do so, document frequency of words is computed as the number of training time series in which the word occurs. BoW are then updated by diving each component by its associated document frequency.

SSR and IDF normalizations both reduce the influence of frequent codewords in the codebook, and are applied before ℓ_2 normalization. We show in the experimental part of this paper that using BoW normalization improves the accuracy of our method.

Normalized histograms are finally given to a classifier that learns how to discriminate classes from this D-BoTSW representation.

4 Experiments and results

In this section, we investigate the impact of both dense extraction of the keypoints and normalization of the Bag-of-Words on classification performance. We then compare our results to the ones obtained with standard time series classification techniques.

For the sake of reproducibility, C++ source code used for (D-)BoTSW in these experiments is made available for download[1]. To provide illustrative timings for our methods, we ran it on a personal computer, for a given set of parameters, using dataset *Cricket_X* [11] that is made of 390 training time series and 390 test ones. Each time series in the dataset is of length 300. Extraction and description of dense keypoints takes around 1 second for all time series in the dataset. Then, 35 seconds are necessary to learn a k-means and fit a linear SVM classifier using training data only. Finally, classification of all D-BoTSW corresponding to test time series takes less than 1 second.

4.1 Experimental setup

Experiments are conducted on the 86 currently available datasets from the UCR repository [11], the largest online database for time series classification. It includes a wide variety of problems, such as sensor reading (*ECG*), image outline (*ArrowHead*), human motion (*GunPoint*), as well as simulated problems (*TwoPatterns*). All datasets are split into a training and a test set, whose size varies between less than 20 and more than 8000 time series. For a given dataset, all time series have the same length, ranging from 24 to more than 2500 points.

Parameters a, n_b, k and C_{SVM} of (D-)BoTSW are learned, while we set $\sigma_0 = 1.6$ and $k_{\text{sc}} = 2^{1/3}$, as these values have shown to produce stable results [17]. Parameters a, n_b, k and C_{SVM} vary inside the following sets: $\{4, 8\}$, $\{4, 8, 12, 16, 20\}$, $\left\{2^i, \forall i \in \{5..10\}\right\}$ and $\{1, 10, 100\}$ respectively. Codebooks are obtained *via* k-means quantization and a linear SVM is used to classify time series represented as (D-)BoTSW. For our approach, the best sets (in terms of accuracy) of (a, n_b, k, C_{SVM}) parameters are selected by performing cross-validation on the training set. Due to the heterogeneity of the datasets, leave-one-out cross-validation is performed on datasets where the training set contains less than 300 time series, and 10-fold cross-validation is used otherwise. These best sets of parameters are then used to build the classifier on the training set and evaluate it on the test set. For datasets with little training data, it is likely that several sets of parameters yield best performance during the cross-validation process. For example, when using *DiatomSizeReduction* dataset, BoTSW has 150 out of 180 parameter sets yielding best performance, while there are 42 such sets for D-BoTSW with SSR normalization. In both cases, the number of *best* parameter sets is too high to allow a fair parameter selection. When this happens, we keep all parameter sets with best performance at training and perform a majority voting between their outputs at test time.

Parameters a and n_b both influence the descriptions of the keypoints; their optimal values vary between sets so that the description of keypoints can fit the shape of the data. If the data contains sharp peaks, the size of the neighborhood on which features are computed (equal to $a \times n_b$) should be small. On the contrary, if it contains smooth peaks, descriptions should take more points into account. Parameter k of the k-means needs to be large enough to precisely

[1] http://people.irisa.fr/Adeline.Bailly/code.html

represent the different features. However, it needs to be small enough in order to avoid overfitting. We consequently allow a large range of values for k.

In the following, BoTSW denotes the approach where keypoints are selected as scale-space extrema and BoW histograms are ℓ_2-normalized. For all experiments with dense extraction, we set $\tau_{\text{step}} = 10$, and we extract keypoints at all scales. Using such a value for τ_{step} enables one to have a sufficient number of keypoints even for small time series, and guarantees that keypoint neighborhoods overlap so that all subparts of the time series are described.

4.2 Experiments on dense extraction

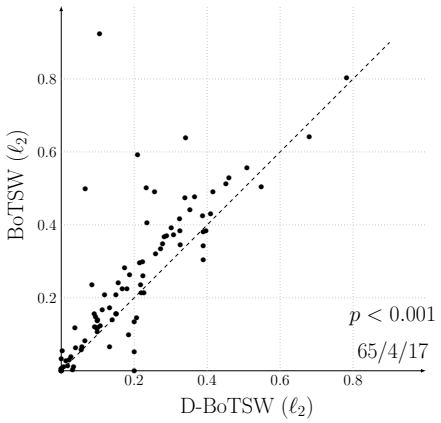


Fig. 2: Error rates of BoTSW compared to D-BoTSW.

Fig. 2 shows a pairwise comparison of error rates between BoTSW and its dense counterpart D-BoTSW for all datasets in the UCR repository. A point on the diagonal means that obtained error rates are equals. A point above the diagonal illustrates a case where D-BoTSW has a smaller error rate than BoTSW. Wilcoxon signed rank test's p-value and Win/Tie/Lose scores are given in the bottom-right corner of the figure. Win/Tie/Lose scores indicate that D-BoTSW reaches better performance than BoTSW on 61 datasets, equivalent performance on 4 datasets and worse on 21 datasets. Wilcoxon test shows that this difference is significant (in the following, we will use a significance level of 10% for all statistical tests).

D-BoTSW improves classification on a large majority of the datasets. However, most points are close to the diagonal, which means that the improvement is of little magnitude. In the following, we show how to further improve these results thanks to D-BoTSW normalization.

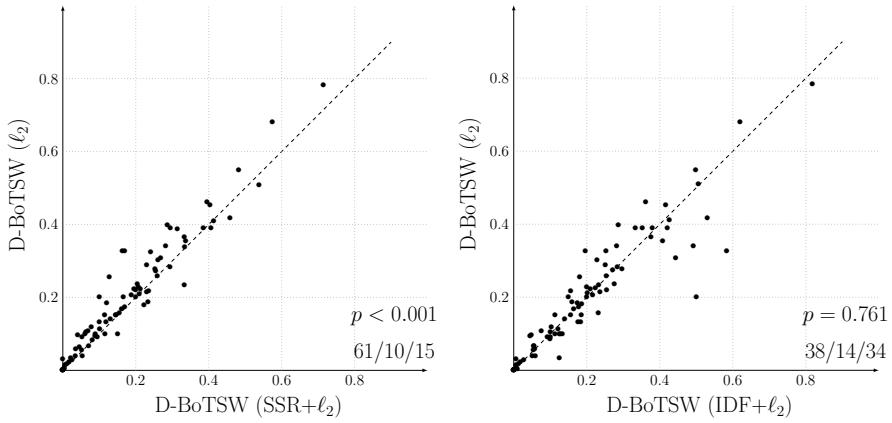


Fig. 3: Error rates of D-BoTSW with and without normalization.

4.3 Experiments on BoW normalization

In image retrieval and classification, Bag-of-Words normalizations have been shown to improve classification rates with dense extracted keypoints. We investigate here the impact of SSR and IDF normalizations on D-BoTSW for time series classification.

As it can be seen in Fig. 3, both SSR and IDF normalizations improve classification performance (though the improvement of using IDF is not statistically significant). Lowering the influence of largely-represented codewords hence leads to more accurate classification with D-BoTSW.

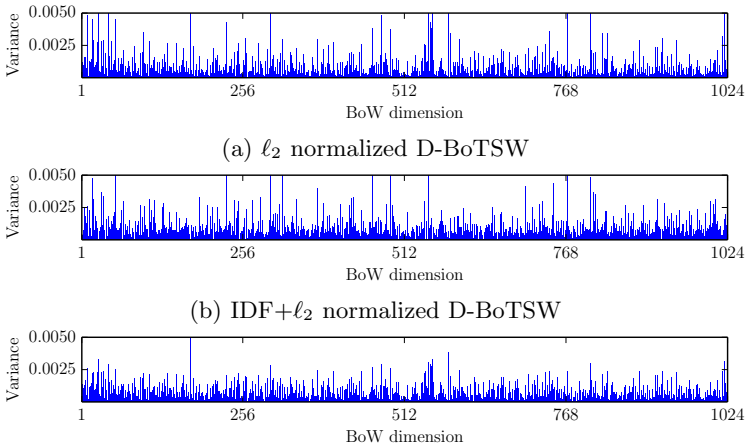
IDF normalization only leads to a small improvement in classification accuracy: Win/Tie/Lose score against non-normalized D-BoSTW is 38/14/34. On the contrary, SSR normalization significantly improves the classification accuracy, with a Win/Tie/Lose score of 61/10/15 over non-normalized D-BoSTW.

This is backed by Fig. 4, in which one can see that when using SSR normalization, variance (*i.e.* energy) is spread across all dimensions of the BoW, leading to a more balanced representation than with other two normalization schemes.

4.4 Comparison with state-of-the-art methods

In the following, we will refer to dense SSR-normalized BoTSW as D-BoTSW, since this setup is the one providing the best classification performance. We now compare D-BoTSW to the most popular state-of-the-art methods for time series classification. The UCR repository provides error rates for the 86 datasets with Euclidean distance 1NN (EDNN) and Dynamic Time Warping 1NN (DTWNN) [20]. We use published error rates for TSBF (45 datasets) [3], SAX-VSM (51 datasets) [21], SMTS (45 datasets) [2], PROP (46 datasets) [15] and BoP (20 datasets).

As BoP [14] only provides classification performance for 20 datasets, we decided not to plot pairwise comparison of error rates between D-BoTSW and BoP.



(a) ℓ_2 normalized D-BoTSW

(b) IDF+ℓ_2 normalized D-BoTSW

(c) SSR+ℓ_2 normalized D-BoTSW

Fig. 4: Per-dimension energy of D-BoTSW vectors extracted from dataset *ShapesAll*. The same codebook is used for all normalization schemes so that dimensions are comparable across all three sub-figures.

Note however that the Win/Tie/Lose score is 17/1/2 in favor of D-BoTSW and this difference is statistically significant ($p < 0.001$). BoP has smaller error rate than D-BoTSW on *wafer* (0.003 *vs.* 0.004) and *Olive Oil* (0.133 *vs.* 0.167) data sets.

Fig. 5 shows that D-BoTSW performs better than 1NN combined with ED (EDNN) or DTW (DTWNN), TSBF, SAX-VSM and SMTS. Though relying on a single similarity measure that has linear time complexity in the length of time series, D-BoTSW slightly outperforms PROP, which relies on outputs from several similarity measures with quadratic time complexity. In Fig. 5, it is striking to realize that D-BoTSW not only improves the classification, but might improve it considerably. Error rate on *Shapelet Sim* dataset drops from 0.461 (EDNN) and 0.35 (DTWNN) to 0 (D-BoTSW), for example. Pairwise comparisons of methods show that all observed differences between D-BoTSW and state-of-the-art methods are statistically significant, except for PROP. Error rates (ER) obtained with D-BoTSW are reported in Table 1, together with baseline scores publicly available at [11].

This set of experiments, conducted on a wide variety of time series datasets, shows that D-BoTSW significantly outperforms most state-of-the-art methods.

5 Conclusion

In this paper, we presented the D-BoTSW technique, which transforms time series into histograms of quantized local features. The association of SIFT keypoints and Bag-of-Words has been widely used and is considered as a standard

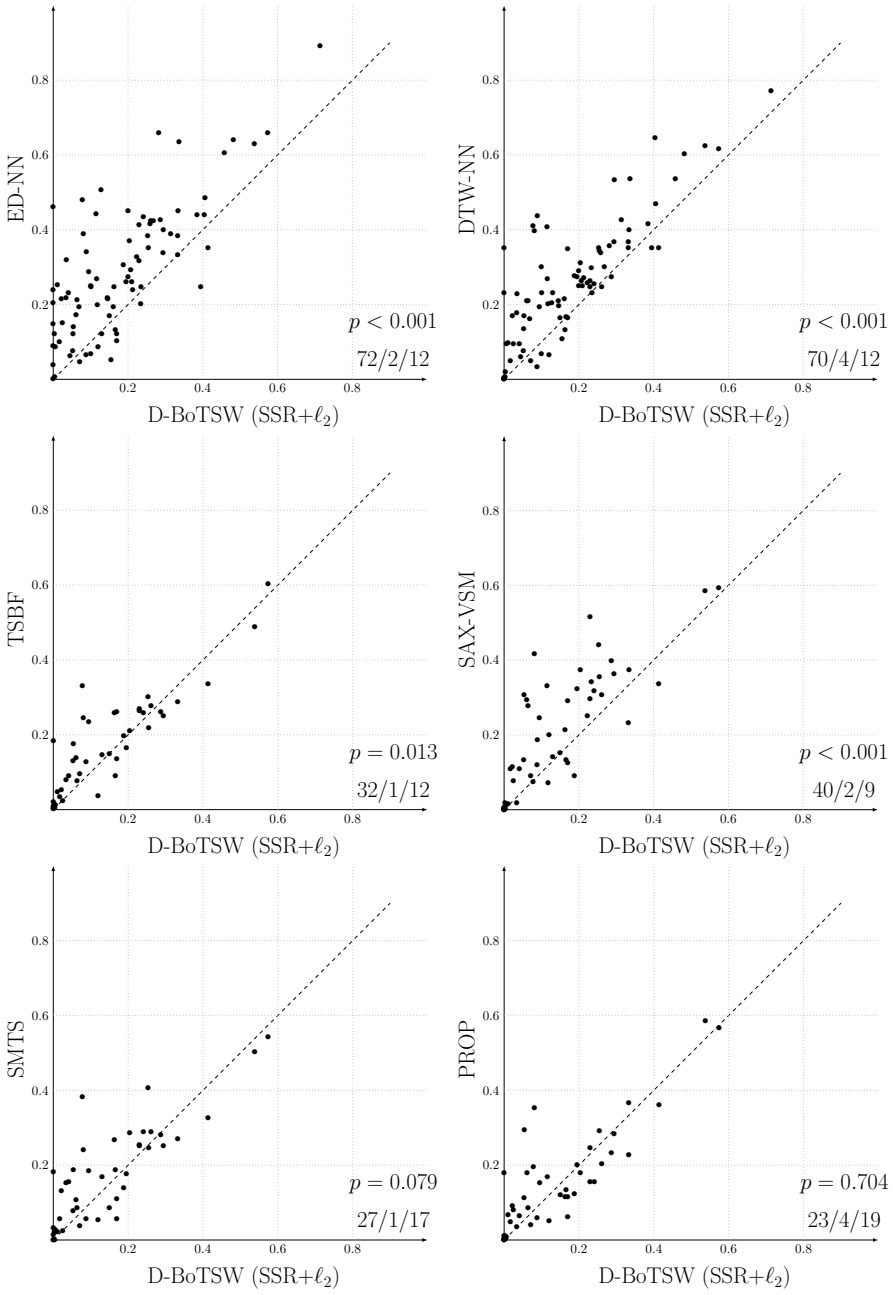


Fig. 5: Error rates for D-BoTSW with SSR normalization versus baselines (ED-NN, DTW-NN, TSBF, SAX-VSM, SMTS, PROP).

Dataset	EDNN	DTWNN	D-BoTSW
50words	0.369	0.31	**0.204**
Adiac	0.389	0.396	**0.082**
ArrowHead	**0.2**	0.297	0.234
Beef	**0.333**	0.367	**0.333**
BeetleFly	0.25	0.3	**0.1**
BirdChicken	0.45	0.25	**0.2**
Car	0.267	0.267	**0.117**
CBF	0.148	0.003	**0**
Chlorine Concentration	**0.35**	0.352	0.414
CinC_ECG _torso	**0.103**	0.349	0.17
Coffee	**0**	**0**	**0**
Computers	0.424	0.3	**0.268**
Cricket_X	0.423	**0.246**	0.262
Cricket_Y	0.433	0.256	**0.241**
Cricket_Z	0.413	0.246	**0.231**
DiatomSize Reduction	0.065	**0.033**	0.088
DistalPhalanx OutlineAgeGroup	0.218	0.208	**0.145**
DistalPhalanx OutlineCorrect	0.248	**0.232**	0.235
DistalPhalanxTW	0.273	0.29	**0.2**
Earthquakes	0.326	0.258	**0.224**
ECG200	**0.12**	0.23	0.13
ECG5000	0.075	0.076	**0.052**
ECGFiveDays	0.203	0.232	**0**
ElectricDevices	0.45	0.399	**0.334**
FaceAll	0.286	0.192	**0.095**
FaceFour	0.216	0.17	**0.023**
FacesUCR	0.231	0.095	**0.041**
FISH	0.217	0.177	**0.034**
FordA	0.341	0.438	**0.089**
FordB	0.442	0.406	**0.116**
Gun_Point	0.087	0.093	**0.007**
Ham	0.4	0.533	**0.295**
HandOutlines	0.199	0.202	**0.119**
Haptics	0.630	0.623	**0.539**
Herring	0.484	0.469	**0.406**
InlineSkate	0.658	0.616	**0.575**
Insect WingbeatSound	0.438	0.645	**0.405**
ItalyPowerDemand	**0.045**	0.05	0.072
LargeKitchen Appliances	0.507	0.205	**0.128**
Lightning2	0.246	**0.131**	0.164
Lightning7	0.425	**0.274**	0.288
MALLAT	0.086	**0.066**	0.12
Meat	**0.067**	**0.067**	0.1
MedicalImages	0.316	0.263	**0.23**
MiddlePhalanx OutlineAgeGroup	0.26	0.25	**0.21**
MiddlePhalanx OutlineCorrect	**0.247**	0.352	0.395
MiddlePhalanxTW	0.439	0.416	**0.386**
MoteStrain	**0.121**	0.165	0.17
NonInvasiveFetal ECG_Thorax1	0.171	0.209	**0.061**
NonInvasiveFetal ECG_Thorax2	0.12	0.135	**0.053**
OliveOil	**0.133**	0.167	0.167
OSULeaf	0.479	0.409	**0.079**
PhalangesOutlines Correct	0.239	0.272	**0.213**
Phoneme	0.891	0.772	**0.714**
Plane	0.038	**0**	**0**
ProximalPhalanx OutlineAgeGroup	0.215	0.195	**0.146**
ProximalPhalanx OutlineCorrect	0.192	0.216	**0.162**
ProximalPhalanxTW	0.292	0.263	**0.208**
RefrigerationDevices	0.605	0.536	**0.459**
ScreenType	0.64	0.603	**0.483**
ShapeletSim	0.461	0.35	**0**
ShapesAll	0.248	0.232	**0.102**
SmallKitchen Appliances	0.659	0.357	**0.283**
SonyAIBORobot Surface	0.141	0.169	**0.055**
SonyAIBORobot SurfaceII	0.305	0.275	**0.189**
StarLightCurves	0.151	0.093	**0.026**
Strawberry	0.062	0.06	**0.046**
SwedishLeaf	0.211	0.208	**0.064**
Symbols	0.1	0.05	**0.017**
synthetic_control	0.12	0.007	**0.003**
ToeSegmentation1	0.320	0.228	**0.035**
ToeSegmentation2	0.192	0.162	**0.069**
Trace	0.24	**0**	**0**
Two_Patterns	0.09	**0**	**0**
TwoLeadECG	0.253	0.096	**0.011**
uWaveGesture Library_X	0.261	0.273	**0.195**
uWaveGesture Library_Y	0.338	0.366	**0.294**
uWaveGesture Library_Z	0.35	0.342	**0.255**
uWaveGesture LibraryAll	**0.052**	0.108	0.156
wafer	0.005	0.02	**0.004**
Wine	0.389	0.426	**0.315**
WordsSynonyms	0.382	0.351	**0.254**
WordSynonyms	0.382	0.351	**0.334**
Worms	0.635	0.536	**0.337**
WormsTwoClass	0.414	0.337	**0.26**
yoga	0.170	0.164	**0.15**

Table 1: Classification error rates for D-BoTSW with SSR normalization (for each dataset, best performance is written as bold text).

technique in image domain, however it has never been investigated for time series classification. We carried out extensive experiments and showed that dense keypoint extraction and SSR normalization of Bag-of-Words lead to the best performance for our method. We compared the results with standard techniques for time series classification: D-BoTSW has comparable performance to PROP with lower time complexity and significantly outperforms all other techniques.

We believe that classification performance could be further improved by taking more time information into account, as well as reducing the impact of quantization losses in our representation. Indeed, only local temporal information is embedded in our model and the global structure of time series is ignored. Moreover, more detailed global representations for sets of features than the standard BoW have been proposed in the computer vision community [9, 18], and such global features could be used in our framework.

Acknowledgments

This work has been partly funded by ANR project ASTERIX (ANR-13-JS02-0005-01), Région Bretagne and CNES-TOSCA project VEGIDAR.

References

1. Adeline Bailly, Simon Malinowski, Romain Tavenard, Thomas Guyet, and Laetitia Chapel. Bag-of-Temporal-SIFT-Words for Time Series Classification. *ECML-PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, 2015.
2. Mustafa G. Baydogan and George Runger. Learning a symbolic representation for multivariate time series classification. *Data Mining and Knowledge Discovery*, 29(2):400–422, 2015.
3. Mustafa G. Baydogan, George Runger, and Eugene Tuv. A Bag-of-Features Framework to Classify Time Series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2796–2802, 2013.
4. Kasim S. Candan, Rosaria Rossini, and Maria L. Sapino. sDTW: Computing DTW Distances using Locally Relevant Constraints based on Salient Feature Alignments. *Proceedings of the International Conference on Very Large DataBases*, 5(11):1519–1530, 2012.
5. Marco Cuturi. Fast global alignment kernels. In *Proceedings of the International Conference on Machine Learning*, pages 929–936, 2011.
6. Ahlame Douzal-Chouakria and Cécile Amblard. Classification trees for time series. *Elsevier Pattern Recognition*, 45(3):1076–1091, 2012.
7. Pauline Dusseux, Thomas Corpetti, and Laurence Hubert-Moy. Temporal kernels for the identification of grassland management using time series of high spatial resolution satellite images. In *Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3258–3260, 2013.
8. Hervé Jégou and Ondrej Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *European Conference on Computer Vision*, 2012.
9. Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010.

10. Frederic Jurie and Bill Triggs. Creating Efficient Codebooks for Visual Recognition. *International Conference on Computer Vision*, 2005.
11. Eamonn Keogh, Qiang Zhu, Bing Hu, Yuan Hao, Xiaopeng Xi, Li Wei, and Chotirat A. Ratanamahatana. The UCR Time Series Classification/Clustering Homepage, 2011. `www.cs.ucr.edu/~eamonn/time_series_data/`.
12. Yann Le Cun and Yoshua Bengio. Convolutional networks for images, speech, and time series. pages 255–258, 1995.
13. Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11, 2003.
14. Jessica Lin, Rohan Khade, and Yuan Li. Rotation-invariant similarity in time series using bag-of-patterns representation. *International Journal of Information Systems*, 39:287–315, 2012.
15. Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 2014.
16. David G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, 1999.
17. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
18. Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
19. Florent Perronnin, Yan Liu, Jorge Sanchez, and Hervé Poirier. Large-Scale Image Retrieval with Compressed Fisher Vectors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, 2010.
20. Chotirat A. Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Proceedings of the Workshop on Mining Temporal and Sequential Data*, pages 22–25, 2004.
21. Pavel Senin and Sergey Malinchik. SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. *Proceedings of the IEEE International Conference on Data Mining*, pages 1175–1180, 2013.
22. Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, pages 1470–1477, 2003.
23. Heng Wang, Muhammad M. Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference*, 2009.
24. Jim Wang, Ping Liu, Mary F.H. She, Saeid Nahavandi, and Addas Kouzani. Bag-of-Words Representation for Biomedical Time Series Classification. *Biomedical Signal Processing and Control*, 8(6):634–644, 2013.
25. Jierui Xie and Mandis Beigi. A Scale-Invariant Local Descriptor for Event Recognition in 1D Sensor Signals. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1226–1229, 2009.
26. Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 947–956, 2009.