Nonlinear Manifold Embedding on Keyword Spotting using t-SNE

George Retsinas[1,2],Nikolaos Stamatopoulos[1], Georgios Louloudis[1], Giorgos Sfikas[1] and Basilis Gatos[1]

[1] Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos", GR-15310 Athens, Greece
Email: {georgeretsi,nstam,louloud,sfikas,bgat}@iit.demokritos.gr

[2] School of Electrical and Computer Engineering, National Technical University of Athens, GR-15773 Athens, Greece

***Abstract*—Nonlinear manifold embedding has attracted considerable attention due to its highly-desired property of efficiently encoding local structure, i.e. intrinsic space properties, into a low-dimensional space. The benefit of such an approach is twofold: it leads to compact representations while addressing the often-encountered curse of dimensionality. The latter plays an important role in retrieval applications, such as keyword spotting, where a sorted list of retrieved objects with respect to a distance metric is required. In this work, we explore the efficiency of the popular manifold embedding method t-distributed Stochastic Neighbor Embedding (t-SNE) on the Query-by-Example keyword spotting task. The main contribution of this work is the extension of t-SNE in order to support out-of-sample (OOS) embedding which is essential for mapping query images to the embedding space. The experimental results demonstrate a significant increase in keyword spotting performance when the word similarity is calculated on the embedding space.**

I. Introduction

Keyword spotting (KWS) is closely related to document indexing and can be defined as the task of locating and retrieving specific words of interest, referred as keywords or queries, in a document collection. In this work, we focus on the segmentation-based Query-by-Example (QbE) keyword spotting category which falls under the content based image retrieval paradigm. Approaches belonging to this category assume as input a query image together with a set of (segmented) word images and return a ranked list of the potentially relevant word images.

Feature extraction is the most crucial step of a QbE KWS method. QbE KWS methods can be categorized, with respect to the extracted features, into (i) methods that extract a feature vector (descriptor) of fixed dimensionality for each word image and (ii) methods that extract a set of features for each word image. Methods of the former category, also called holistic word representations, attract a lot of interest due to their simplicity at the retrieval step. Specifically, such representations require a simple distance/similarity measure for the retrieval step (e.g. Euclidean distance) contrary to techniques belonging to the second category which require more complex matching algorithms (e.g. DTW sequential matching).

The majority of KWS methods that rely on holistic word representations generate high-dimensional descriptors ([1],[2],[3],[4]). However, Euclidean distance on high-dimensional vectors is not a reliable metric for the generation of the retrieval list. This observation leads to the realization that a dimensionality reduction technique is essential in order to fully utilize the descriptive power of holistic representations. Preserving as much of the significant structure of the high-dimensional data as possible in the low-dimensional map is crucial and thus nonlinear dimensionality reduction techniques are required. An interesting property, which has proven to be effective, is to assume that the high-dimensional data lies on a manifold of significant lower intrinsic dimensionality. Thus, the computation of the low-dimensional map is equivalent to learning the underlying manifold. The generated nonlinear mapping is called manifold embedding ([5],[6],[7]).

This work relies on the well-known t-distributed Stochastic Neighbor Embedding (t-SNE) [8] due to its success on the dimensionality reduction task for a large variety of real datasets. The main hindrance for a t-SNE based KWS application is the addition of a new descriptor on the previously learnt embedding, i.e the embedding of the query representation. The majority of manifold learning approaches, including t-SNE, are non-parametric, meaning that no straightforward way exists to add a new descriptor to the embedding. This is referred as the out-of-sample problem. In order to overcome this problem, we propose a novel out-of-sample extension to the t-SNE embedding. This extension enables us to utilize the t-SNE method and explore its efficiency for the KWS task.

The rest of this paper is organized as follows. In Section II related work is highlighted, while in Section III a summarization of t-SNE is presented. Section IV describes in detail the proposed out-of-sample extension. Comparative experimental results are discussed in Section V. Finally, conclusions and future directions are drawn in Section VI.

II. Related Work

Several manifold embedding methods have been reported in the literature aiming to generate a non-linear mapping which encodes high-dimensional data to a low-dimensional space without significantly affecting the local structure of the initial space. Notable manifold embedding techniques are Isomap [5], which creates an embedding based on geodesic

distances, and Locally Linear Embedding (LLE) [6] as well as Laplacian Eigenmaps [7], which both assume the same local structure (linearity) for both the initial high-dimensional and the resulting low-dimensional space. Such techniques can be viewed as generalized eigenvector problems at adjacency matrices. The aforementioned techniques are sensitive to outliers as well as to the predefined dimensionality of the embedding space and consequently lead to the generation of low quality embeddings for the case of challenging datasets. On the contrary, t-SNE [8] has been extensively used on real datasets, providing embeddings of high quality, even when the embedding dimensionality is lower than the intrinsic dimensionality of the underlying manifold.

The majority of the manifold embedding methods do not support the addition of a new sample to the already learnt embedding. This is referred as the out-of-sample problem for which many approaches have been proposed in the literature. Two main categories can be distinguished: parametric and non-parametric out-of-sample extensions. Parametric approaches assume that the learnt embedding can be modeled by a (non-linear) combination of the initial data along with a set of parameters [9], [10]. By estimating these parameters on the already extracted embedding, the out-of-sample extension is straightforward using the same model and the estimated parameters. The main disadvantage of such approaches is the assumption that the generated mapping can be efficiently represented by a (non-linear) model of ideally few parameters. On the other hand, non-parametric approaches usually exploit the geometric intuition of the local structure and the nature as well as specific characteristics of the selected manifold learning algorithm [11], [12], [13].

To the best of our knowledge, the only approach that utilizes manifold embedding for the task of KWS is the work of Sudholt et al. [14]. The authors of [14] proposed a variation of Isomap embedding for the case of Bag of Visual Words (BoVW) features. Although one can become aware of the efficiency of manifold embedding on the reduction of the descriptor's size without significantly affecting the retrieval performance, the presented system has some notable shortcomings mainly derived from the Isomap embedding, such as its sensitivity to the selection of the embedding dimensionality. Furthermore, the Isomap embedding requires the computation of geodesic distances, even for the out-of-sample scenario, which is a computational overhead for the retrieval step. In addition, although a significant reduction of the memory requirements has been achieved no consistent gain in retrieval performance was reported.

III. t-SNE

The goal of t-SNE is to minimize the divergence between the pairwise similarity distributions of input points and the low-dimensional embedded points. The input points are denoted as $\{x_i\}$ and their corresponding embeddings are denoted as $\{y_i\}$, where $i = 1, \ldots, N$. The joint probability p_{ij} that measures the pairwise similarity between two points x_i and x_j is denoted as follows:

$$p_{j|i} = \frac{exp(-d(x_i, x_j)^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-d(x_i, x_k)^2/2\sigma_i^2)}, \quad p_{i|i} = 0 \tag{1}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{2}$$

For the rest of this work, the distance function $d(\cdot, \cdot)$ is considered to be Euclidean as in [8]. The standard deviation σ_i is computed according to a predefined perplexity which can be considered as the effective number of neighbors for each point x_i.

The pairwise similarities in the embedding space are modeled by a normalized Student's-t distribution with a single degree of freedom. The embedding similarity between two points y_i and y_j is defined as:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{Z}, \quad q_{ii} = 0 \tag{3}$$

$$Z = \sum_k \sum_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1} \tag{4}$$

The choice of the Student's-t kernel prevents the crowding problem, as it is explained in [8], which favors embeddings whose points are gathered in the center of the space. The heavy-tailed Student's-t maps sufficiently well points that are far-apart even if the dimension of the embedding space is lower than the (unknown) intrinsic dimensionality of the existing manifold.

Given the definitions of pairwise similarity distributions for both the initial and the embedding space, the embedding $\mathbf{Y}$ is calculated by minimizing the Kullback-Leibler divergence:

$$C(\mathbf{Y}) = KL(P||Q(\mathbf{Y})) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{5}$$

The aforementioned minimization problem does not have an analytical solution. To this end, iterative methods are employed in order to find an embedding $\mathbf{Y}$ that (locally) minimizes the divergence. The problem is solved by a gradient descent method, whereas the gradient of the divergence for each point of the embedding space is computed as follows:

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j) \tag{6}$$

IV. Out-of-sample (OOS) Extension of t-SNE

We assume a set of points x_i, which correspond to the descriptors of the word images for the KWS task, and their embeddings y_i as the result of the t-SNE optimization. Given a new point x in the initial space, our goal is to estimate its mapping y to the t-SNE embedding space. We define the following auxiliary functions in accordance to the t-SNE formulation:

$$p(x|x_i) = \frac{exp(-\|x - x_i\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_k - x_i\|^2/2\sigma_i^2)} \tag{7}$$

$$p(x, x_i) = \frac{p(x|x_i) + p(x_i|x)}{2N}, \quad p(x_i, x_i) = 0 \tag{8}$$

$$s(y, y_i) = (1 + \|y - y_i\|^2)^{-1} \tag{9}$$

$$q(y, y_i) = s(y, y_i) / \sum_k \sum_{l \neq k} s(y_k, y_l) \tag{10}$$

A straightforward solution to the out-of-sample problem is to preserve the local structure of the initial space [6] which can be formulated as the minimization of the cost:

$$C_{oos}(y|x) = \sum_i w(x, x_i) \|y - y_i\|^2 \tag{11}$$

All previously learnt embeddings y_i are considered fixed, so we minimize over the sought embedding y. The function $w(x, x_i)$ is a pairwise similarity function (e.g. a Gaussian kernel) and in correspondence to t-SNE, the previously defined function $p(x, x_i)$ can be used. The above minimization has a closed form solution:

$$y^\star = \frac{\sum_i p(x, x_i) y_i}{\sum_i p(x, x_i)} \tag{12}$$

A drawback of this solution, as well as of the majority of the existing OOS methods, is that it provides a general approach for the OOS problem (i.e. locality preservation of the initial space) while ignoring crucial aspects of t-SNE success, namely the Student's-t distribution and the locality of the embedding space. Contrary to existing approaches, in order to address the out-of-sample problem, we examine the initial equations of t-SNE.

Proposed Gradient Descent Approach: The estimation of the new embedding y is computed iteratively by minimizing the t-SNE cost according to a gradient descent procedure:

$$y^{t+1} = y^t - \alpha \frac{\partial C(y^t)}{\partial y^t} \tag{13}$$

$$\frac{\partial C(y)}{\partial y} = 4 \sum_i [p(x, x_i) - q(y, y_i)] s(y, y_i)(y - y_i) \tag{14}$$

The main shortcoming of a gradient descent estimation is its convergence rate. If a fixed step size a is predefined, the convergence may be extremely slow. In order to avoid a slow convergence, we propose the use of the following adaptive step size:

$$\alpha(y^t) = \big[4 \sum_i p(x, x_i) s(y^t, y_i)\big]^{-1} \geq 0 \tag{15}$$

Therefore, the update equation for iteratively estimating the embedding y is:

$$y^{t+1} = y^t - \frac{\sum_i [p(x, x_i) - q(y^t, y_i)] s(y^t, y_i)(y^t - y_i)}{\sum_i p(x, x_i) s(y^t, y_i)} \tag{16}$$

It should be noted that the update equation (Eq. 16) can be derived from the solution of $\|\frac{\partial C(y)}{\partial y}\| = 0$ and thus it is equivalent to a fixed point iteration approach.

Aiming to further promote the simplicity of the update equation and the speed convergence, we choose to omit the terms of Eq. 16 referring to $q(y^t, y_i)$. The term $\sum_i q(y^t, y_i) s(y^t, y_i)(y - y_i)$ corresponds to the derivative of the normalizing term $Z = \sum_k \sum_{l \neq k} s(y_k, y_l)$ and it is responsible for keeping the new embedding y sufficiently apart from the embeddings y_i, as a repulsive force. Concerning retrieval applications, only the relative distances between the new embedding and the already embedded points are of interest and thus this repulsion property is not important. Consequently, for the rest of this work, the proposed out-of-sample embedding is approximated by minimizing the cost $C_r(y|x) = \sum_i p(x, x_i) log\big(p(x, x_i)/s(y, y_i)\big)$ which is performed by the following update equation:

$$y^{t+1} = \frac{\sum_i p(x, x_i) s(y^t, y_i) y_i}{\sum_i p(x, x_i) s(y^t, y_i)} \tag{17}$$

The adaptive step size $a(y)$ for the latest update equation can be easily proven to concede with the optimum step size for the line search strategy over the gradient descent algorithm. This means that the acquired step size of Eq. 15 at each iteration is the solution to the minimization problem:

$$\alpha(y^t) = \arg\min_{\alpha > 0} C_r\big(y^t - \alpha \frac{\partial C_r(y^t)}{\partial y^t}\big) \tag{18}$$

The aforementioned observation ensures significantly faster convergence compared to setting a predefined step size, which was empirically verified through experimentation.

Implementation Issues: The computation of $q(y, y_i)$ is straightforward at each iteration. However, $p(x, x_i)$ is calculated only once, before the iteration process, and involves summations over all the pairwise Gaussian functions. To overcome this problem, we store the standard deviations σ_i and the partial sums $S_i = \sum_{k \neq i} exp(-\|x_k - x_i\|^2 / 2\sigma_i^2)$ as auxiliary variables generated during the t-SNE embedding. Having estimated the standard deviation σ for the unseen point x (using the predefined perplexity), we redefine the equations of $p(x, x_i)$ as follows:

$$p(x_i|x) = \frac{exp(-\|x - x_i\|^2 / 2\sigma^2)}{\sum_{k=1}^{N} exp(-\|x - x_i\|^2 / 2\sigma^2)} \tag{19}$$

$$p(x|x_i) = \frac{exp(-\|x - x_i\|^2 / 2\sigma_i^2)}{S_i + exp(-\|x - x_i\|^2 / 2\sigma_i^2)} \tag{20}$$

$$p(x, x_i) = \frac{p(x|x_i) + p(x_i|x)}{2N}, \quad p(x_i, x_i) = 0 \tag{21}$$

The above formulation requires only the distances of the new point x from the existing points x_i, i.e. O(N) computations.

Complexity: Given a set of N points $\{x_i\}$ of d_x dimensions and their embeddings $\{y_i\}$ of $d_y \ll d_x$ dimensions, the complexity of computing the embedding y of an out-of-sample point x is estimated as follows:

- $O(N d_x)$ for computing the $p(x, x_i)$ pairwise similarities.
- $O(N d_y)$ for updating y in each iteration.

Assuming k as the total number of iterations for convergence, the overall complexity is $O(N(d_x + k d_y))$. For small embedding dimension d_y and number of iterations ($d_y = 3 \,\&\, k = 10$), the computation of the pairwise similarities in the initial space, when d_x is large enough which is usually the case, governs the computation time ($d_x \gg k d_y$). This observation hints that the proposed OOS extension is only sligthly slower than the closed form solution approach of Eq. 12. It should be noted that the OOS embedding procedure should be fast

for KWS applications, because it is computed during query (retrieval) time. Memory requirements correspond to storing the initial data points and their embeddings, as well as the standard deviations σ_i and sums S_i, i.e. $O(N(d_x + d_y + 2))$, which requires only $N \times (d_y + 2)$ more memory space compared to storing only the initial data points.

V. Experimental Results

A. Experimental Setup

The proposed OOS extension of t-SNE embedding method is applied on QbE keyword spotting as a post-processing step after the extraction of fixed-sized descriptors. In order to highlight the efficiency of the manifold assumption and the capability of the proposed method, its evaluation is performed on three state-of-the-art descriptors. The performance of the KWS task was recorded in terms of the Precision at Top 5 Retrieved words (P@5) as well as the Mean Average Precision (MAP).

The workflow for the application of t-SNE embedding on KWS includes the following steps: **1)** Extract the descriptors for each word image of the dataset. **2)** Perform Principal Component Analysis (PCA) on the dataset descriptors. This is suggested before using t-SNE because it preserves the global structure of the points/descriptors and reduces noise. For this work, the feature vector dimension after PCA is set to $d_{pca} = 400$ regardless the initial dimension of the descriptors (selected descriptors have a dimensionality over 400). **3)** Perform t-SNE embedding on the descriptors. The accelerated version of t-SNE is selected, which uses tree-based structures [15]. Due to the fact that the t-SNE approach generates an embedding which corresponds to a local minimum of the t-SNE cost optimization problem, the process was repeated multiple (five) times with different (random) initialization. The embedding with the lowest cost was selected as the final embedding. **4)** Compute the auxiliary values σ_i and S_i, which are required in the proposed out-of-sample extension. **5)** Given a query image, compute the initial descriptor and perform PCA. The resulting descriptor is used as input (along with descriptors of the dataset, their corresponding embeddings and the auxiliary variables) to the out-of-sample estimation method. **6)** Generate the retrieval list using Euclidean distance on the embedding space.

B. Preprocessing and Descriptors

Before we proceed with feature extraction, we apply a preprocessing step which consists of contrast and main-zone normalization. Contrast normalization is performed by replacing Sauvola's binarization hard assignment with a soft one. Main-zone normalization is based on detecting the main-zone in a way similar to [1]. After the detection of the main-zone, skew correction is performed using the slope of the detected main-zone, as well as a vertical normalization of the image by moving the main zone at the center of the generated normalized image. An example of the effect of the aforementioned preprocessing step is depicted in Figure 1.

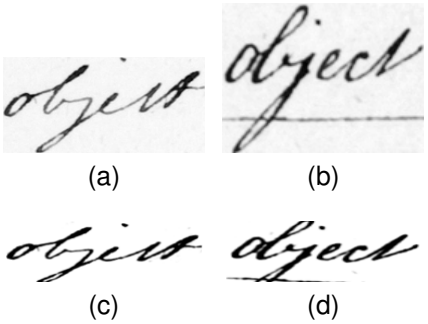

Fig. 1. Different instances of the same word before (a),(b) and after (c),(d) the preprocessing.

Three state-of-the-art holistic descriptors are selected, which are briefly described below:

BoVW: A Bag of Visual Words (BoVW) approach organized in Spatial Pyramids was implemented due to the established efficiency of such methods in keyword spotting [3]. Dense SIFT features at multiple scales were chosen as local descriptors and a codebook of 1024 entries for the histogram encoding. Spatial pyramids are employed to encode indirectly the spatial information as in [3].

POG: An image is segmented in three (overlapping) parts and each segment is encoded using the Projections of Oriented Gradients (POG) descriptors, which have shown to perform well in keyword spotting [1]. In this work, a slight modification of POG descriptor is used in order to be applied to gray-scale images.

ZAH: Zoning Aggregating Hypercolumns (ZAH) features are based on a pre-trained Deep Convolutional Network (DCN) [2]. The features are extracted from the output of the convolutional layers of a DCN, which was trained on an independent set of typewritten characters. The final descriptor is produced by the concatenation of the aggregated convolution responses over (six) image segments.

C. Out-of-Sample Approaches

Concerning the efficiency of the proposed OOS method, the following OOS embedding methods have been considered for comparison:

CFS: Out-of-sample extension using Eq. 12. This approach assumes that the local structure of the embedding space is defined only by pairwise similarities of the initial points [6].

Parametric: A parametrization between the initial data and the produced embeddings is introduced according to [9], where the parametric form $y(x) = f_a(x)$ is assumed and a are the sought parameters. Non-linearity is introduced by Gaussian kernels of the form $k(x, x_i) = exp(-\|x - x_i\|^2/2\sigma_i^2)$. Thus, the parametrization is defined as $y(x) = \sum_i a_i k(x, x_i) / \sum_l k(x, x_l)$. The parameters are estimated in a least square manner: $A = K^+ Y$, where $[K]_{ij} = k(x_i, x_j) / \sum_l k(x_i, x_l)$. The parametric approach of [9] suggests using only a set of landmark points, i.e. a subset of the initial points, which alleviates the computation overhead

TABLE I
MAP AND P@5 EVALUATION ON ALL DATASETS FOR $d_y = 3$

Descriptor	OOS approach	GW20		Bentham14		Modern14	
		MAP	P@5	MAP	P@5	MAP	P@5
BoVW	No Embedding	72.30	91.59	55.29	74.50	28.93	50.60
	CFS	80.10	82.22	48.42	49.50	19.67	23.00
	Parametric-90	82.59	88.58	35.93	32.25	12.81	13.00
	Parametric-100	85.13	92.30	38.13	35.12	13.62	15.33
	Proposed	**85.18**	**92.42**	**63.57**	**78.00**	**33.93**	**53.87**
POG	No Embedding	62.49	**85.85**	66.01	**82.25**	36.83	61.80
	CFS	68.63	74.51	56.09	53.69	29.81	28.60
	Parametric-90	70.38	79.82	55.67	57.12	35.21	47.73
	Parametric-100	74.15	85.04	60.77	64.31	43.64	57.13
	Proposed	**74.19**	85.14	**70.43**	80.94	**48.21**	**65.80**
ZAH	No Embedding	61.19	86.40	**53.49**	75.69	33.29	56.33
	CFS	70.69	76.45	37.41	44.06	24.65	30.07
	Parametric-90	74.10	84.44	28.51	33.56	30.30	43.33
	Parametric-100	76.77	89.13	34.95	42.62	33.08	47.87
	Proposed	**76.82**	**89.23**	53.17	**76.75**	**38.84**	**59.93**

of inverting a matrix of size $N \times N$. However, in practice, selecting a subset of the initial space leads to poor performance. To highlight this behavior, 90% of the points are randomly selected in order to estimate the parameter matrix A. It should be stressed that 90% is a very high percentage of points kept, which yields no significant computational acceleration. As a result, we distinguish two variations, **parametric-100** and **parametric-90**, where 100% (all) and 90% of the points are used as landmark points, respectively.

Proposed: Out-of-sample extension using gradient descent (Eq. 17) based on the initial t-SNE cost function. The proposed step size is adaptive and optimal according to the line search strategy, which guarantees fast convergence. The maximum number of iterations is set to $N_{max} = 15$, since the majority of the out-of-sample experiments achieve convergence under 10 iterations.

D. Datasets

The evaluation is performed on the widely used George Washington Dataset [16] as well as on the more challenging datasets of ICFHR 2014 KWS Competition [17]. The datasets are summarized below:

GW20: This dataset is the well-known collection of writings of George Washington, consisting of 20 pages segmented into 4860 words. Words with ten or more instances and three or more characters are selected as queries as in [3], resulting in 1844 image queries.

Bentham14: This dataset was part of the ICFHR 2014 H-KWS competition and includes manuscripts in English written by Jeremy Bentham himself as well as by Bentham's secretarial staff. It consists of 10370 segmented word images from 50 document images and 320 image queries.

Modern14: This dataset was also part of the ICFHR 2014 H-KWS competition and includes handwritten documents written in four languages (English, French, German and Greek). It consists of 14754 segmented word images from 100 document images (25 for each language) and 300 image queries.

E. Performance Evaluation

To verify the efficiency of the proposed OOS extension, we apply the aforementioned OOS methods on all the descriptors and datasets for the case of $d_y = 3$ (embedding dimension). The results, in terms of MAP and P@5, are presented in Table I. The *No Embedding* case corresponds to the absence of a manifold embedding step, i.e. the PCA generated descriptors are used. The main observations are summarized below:

- The proposed OOS extension performs significantly better compared to the other OOS methods, especially in the challenging Betham14 and Modern14 datasets.
- The parametric method shows similar performance only on the GW20 dataset, which is smaller and less challenging, while in the other two datasets its performance deteriorates significantly. This leads to the conclusion that the parametric approach of [9] cannot model the t-SNE embedding sufficiently well. Specifically, the *parametric-90* variation reports a considerable drop in performance, even though 90% of all points are used. This observation hints that the use of landmark points yields unreliable parameters for parametric OOS extension.
- Another important observation is that the performance may drop after the use of manifold embedding compared to the case of using the initial descriptors (No Embedding case). This drop in performance is mainly credited to t-SNE embedding of the word descriptors, rather than the OOS methods. It is possible that the selected embedding dimensionality is much lower than the intrinsic dimensionality of the underlying manifold and thus the generated embedding is not suitable. It should be noted that the intrinsic dimensionality depends on the descriptor and the dataset selection, since both define the initial space.

Furthermore, we investigate the importance of the embedding dimension d_y. Tables II, III and IV summarize the results for different embedding dimensions concerning the GW20, Bentham14 and Modern14 datasets, respectively. The proposed OOS extension is used to obtain the low-dimensional

embedding of the query image. In addition, the performance of state-of-the-art KWS methods are provided for comparison. The main observations are summarized below:

- A significant gain in performance, more than 10% in some cases, is observed for the majority of the datasets and the descriptors, when using the t-SNE embedding.
- It can be observed that only in few cases the overall gain is small. In addition, the efficiency increases along with the embedding dimension. This behavior hints towards a higher intrinsic manifold dimensionality.
- The presented KWS approach provides results that outperform the majority of state-of-the-art techniques without any fine-tuning (Aldavert et al. [3] performed fine-tuning on the GW20 dataset). A noteworthy observation is that these results have been reported using a very low embedding dimensionality ($d_y = 2, 3, 4, 5$), which highlights the efficiency of the t-SNE method.

TABLE II
MAP EVALUATION ON GW20 DATASET

Method	No Embedding	$\mathbf{d_y}$ 2	3	4	5
Kovalchuk [4]	66.30	-	-	-	-
Aldavert [3]	76.50	-	-	-	-
BoVW	72.30	84.12	85.35	85.62	86.23
POG	62.49	70.76	74.01	74.79	74.96
ZAH	61.19	74.85	78.51	78.43	79.54

TABLE III
MAP EVALUATION ON BENTHAM14 DATASET

Method	No Embedding	$\mathbf{d_y}$ 2	3	4	5
Kovalchuk [17]	52.40	-	-	-	-
Almazan [17]	51.30	-	-	-	-
Howe [17]	46.20	-	-	-	-
fPOG [1]	57.70	-	-	-	-
Aldavert [3]	46.50	-	-	-	-
BoVW	55.29	61.02	62.38	64.82	64.81
POG	66.01	67.95	70.68	70.66	71.54
ZAH	53.49	53.30	54.46	53.83	54.07

TABLE IV
MAP EVALUATION ON MODERN14 DATASET

Method	No Embedding	$\mathbf{d_y}$ 2	3	4	5
Kovalchuk [17]	33.80	-	-	-	-
Almazan [17]	52.30	-	-	-	-
Howe [17]	27.80	-	-	-	-
fPOG [1]	35.50	-	-	-	-
Aldavert [3]	38.90	-	-	-	-
BoVW	28.93	34.09	34.56	35.64	36.29
POG	36.83	48.82	50.51	49.39	51.61
ZAH	33.29	39.30	40.02	39.50	40.65

VI. CONCLUSIONS

A novel out-of-sample extension of t-SNE has been proposed, which displays superior performance compared to other out-of-sample extension methods. The proposed extension is applied on the keyword spotting task, where word descriptors are embedded using t-SNE and query retrieval corresponds to an out-of-sample problem. The experimental results demonstrate a significant gain in KWS retrieval performance while using Euclidean distance on the embedding space. As a future direction, the estimation of the intrinsic manifold dimensionality as well as an efficient way of generating higher dimensional t-SNE embeddings could be further explored.

ACKNOWLEDGMENT

This work has been supported by the EU project READ (Horizon-2020 programme, grant Ref. 674943).

REFERENCES

[1] G. Retsinas, G. Louloudis, N. Stamatopoulos and B. Gatos, "Keyword Spotting in Handwritten Documents using Projections of Oriented Gradients", International Workshop on Document Analysis Systems, pp. 411-416, Greece, 2016.

[2] G.Sfikas, G.Retsinas and B.Gatos, "Zoning Aggregated Hypercolumns for Keyword Spotting", International Conference on Frontiers in Handwriting Recognition, pp. 283-288, China, 2016.

[3] D. Aldavert, M. Rusiñol, R. Toledo, and J. Lladós, "A study of Bag-of-Visual-Words representations for handwritten keyword spotting", International Journal on Document Analysis and Recognition, vol. 18, no. 3, pp. 223-234, 2015.

[4] A. Kovalchuk, L. Wolf, and N. Dershowitz, "A simple and fast keyword spotting method", International Conference on Frontiers in Handwriting Recognition, pp. 3-8, Greece, 2014.

[5] J.B. Tenenbaum, "Mapping a manifold of perceptual observations", Advances in Neural Information Processing Systems, vol. 10, pp. 682-688, 1998.

[6] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by Locally Linear Embedding", Science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[7] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and spectral techniques for embedding and clustering", Advances in Neural Information Processing Systems, vol. 14, pp. 585-591, 2002.

[8] L.J.P. van der Maaten and G.E. Hinton, "Visualizing High-Dimensional Data Using t-SNE", Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

[9] A. Gisbrecht, A. Schulz and B. Hammer, "Parametric nonlinear dimensionality reduction using kernel t-SNE", Neurocomputing, vol. 147, pp. 71-82, 2015.

[10] L.J.P. van der Maaten, "Learning a Parametric Embedding by Preserving Local Structure", International Conference on Artificial Intelligence & Statistics, pp. 384-391, 2009.

[11] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering", Advances in Neural Information Processing Systems, vol. 16, pp. 177-184, 2004.

[12] H. Strange and R. Zwiggelaar, "A generalised solution to the out-of-sample extension problem in manifold learning", American Association for Artificial Intelligence Conference, 2011.

[13] M. Carreira-Perpinan and Z. Lu, "The Laplacian Eigenmaps Latent Variable Model", International Conference on Artificial Intelligence & Statistics, pp. 59-66, 2007.

[14] S. Sudholt and G.A. Fink, "A Modified Isomap Approach to Manifold Learning in Word Spotting", German Conference on Pattern Recognition, pp. 529-539, 2015.

[15] L.J.P. van der Maaten, "Accelerating t-SNE using Tree-Based Algorithms", Journal of Machine Learning Research, vol. 15, pp. 3221-3245, 2014.

[16] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents", Workshop on Document Image Analysis for Libraries, pp. 278-287, 2004.

[17] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis and N. Stamatopoulos, "ICFHR 2014 Competition on Handwritten KeyWord Spotting (H-KWS 2014)", International Conference on Frontiers in Handwriting Recognition, pp. 814-819, Greece, 2014.