A Modified Isomap Approach to Manifold Learning in Word Spotting

Sebastian Sudholt, Gernot A. Fink

{sebastian.sudholt,gernot.fink}@tu-dortmund.de
Technische Universität Dortmund, Germany

Abstract. Word spotting is an effective paradigm for indexing document images with minimal human effort. Here, the use of the Bag-of-Features principle has been shown to achieve competitive results on different benchmarks. Recently, a spatial pyramid approach was used as a word image representation to improve the retrieval results even further. The high dimensionality of the spatial pyramids was attempted to be countered by applying Latent Semantic Analysis. However, this leads to increasingly worse results when reducing to lower dimensions. In this paper, we propose a new approach to reducing the dimensionality of word image descriptors which is based on a modified version of the Isomap Manifold Learning algorithm. This approach is able to not only outperform Latent Semantic Analysis but also to reduce a word image descriptor to up to 0.12 % of its original size without losing retrieval precision. We evaluate our approach on two different datasets.

Keywords: Word Spotting, Manifold Learning, Isomap, Multidimensional Scaling, Bray Curtis Distance, Document Image Analysis

1 Introduction

The automatic transcription of handwritten documents is a challenging task for automated systems. In contrast to machine printed character recognition, it is still considered an unsolved problem and has attracted major interest in the research community. Standard OCR methods perform poorly on these kinds of documents as the variability in chracters is much higher than in a machine printed context. Additionally, a large number of handwritten documents are from ancient times thus exhibiting different kinds of degradation such as fading ink or noise.

In order to overcome the limitations of OCR systems, different approaches have been proposed with *Keyword spotting* or simply *word spotting* being one of the most prominent for automatic document indexing. In *Query-by-Example (QbE)* word spotting the user supplies a query word image to the system and a list of potentially relevant word images is returned from the document collection.

A preliminary version of the presented approach won the ICDAR 2015 Competition on Keyword Spotting for Handwritten Documents

The major advantage here is that only a very small amount of annotated query word images is needed thus reducing manual labeling work.

As QbE word spotting is essentially a form of image retrieval, most word spotting approaches have made use of well established computer vision techniques. Here, the use of local descriptors in a Bag-of-Features approach has been proven to be well suited for this task. As the visual words used here exploit no spatial knowledge, spatial pyramids and Fisher vectors were used to regain a certain amount of spatial information [4, 10]. As the visual vocabulary is generally much bigger in word spotting than in other image retrieval applications, the resulting spatial pyramids and Fisher vectors are very high dimensional [2, 4, 10]. This fact has been accounted for by using *Latent Semantic Analysis (LSA)* to embed the word image descriptors into a lower dimensional space [10]. However, the resulting representations almost always lead to a loss in retrieval precision. Moreover, satisfying results were only achieved when projecting into still high-dimensional spaces (roughly 1500 dimensions).

Based on a metric evaluation to find the dissimilarity measure best suited for comparing spatial pyramid representations of word images, we present a new approach for reducing their dimensionality by modifying the well known Isomap algorithm. This algorithm belongs to the family of manifold learning techniques. It uses a non-linear function to obtain the low-dimensional data thus allowing for more complex projections than LSA. The modified version is able to deal with high-dimensional histograms in a sparsely sampled space. We evaluate the presented method on two different datasets.

2 Manifold Learning

The objective for dimension reduction techniques is to find a low-dimensional representation of the original data. The main assumption in manifold learning is that the original data lies on or close to a manifold which is embedded in a high-dimensional space and has a lower intrinsic dimensionality. When applying dimensionality reduction by manifold learning, the projected data is referred to as the *embedding*.

There exists a vast amount of different unsupervised manifold learning algorithms which can be classified into two classes. Local techniques, such as *Locally Linear Embedding (LLE)* [9] and *Local Tangient Space Alignment (LTSA)* [13], find the embedding by preserving local neighborhood structures of the supplied data. Global techniques, such as *Isomap* [12], aim at keeping global structures of the data thus keeping geometrically close points together while maintaining a bigger distance between geometrically distant data points. In the following, we will concentrate on the Isomap algorithm as it can be exploited in numerous ways in the context of word spotting. It is an unsupervised paradigm thus posing no need for annotated word images. Additionally, there exists an extension for Isomap called *Landmark Isomap* [11] which allows for a computationally efficient approximation of the Isomap embedding when faced with a large amount of data.

The backbone of the Isomap algorithm is the use of *Multidimensional Scaling (MDS)*. MDS solves the inverse distance problem: given a set of pairwise distances between unknown points in a d-dimensional space, find the location of the points. Given a matrix $\mathbf{D}$ of pairwise distances between n data samples, MDS starts by double centering the matrix of squared distances $\mathbf{D}^2$:

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}^2\mathbf{H}, \tag{1}$$

$$\mathbf{H} = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T, \tag{2}$$

where I_n is the $n \times n$ identity matrix and $\mathbf{1}_n\mathbf{1}_n^T$ the $n \times n$ matrix of all ones. Essentially, the double centering removes the column and row mean of $\mathbf{D}^2$.

Afterwards, the eigenvalues λ_i and their corresponding eigenvectors $\mathbf{v}_i$ are extracted from $\mathbf{B}$. The eigenvalues are then sorted in descending order. With λ_1 being the biggest eigenvalue and λ_d being the smallest, the embedding $\mathbf{E}$ is then generated as follows:

$$\mathbf{E} = \begin{pmatrix} \sqrt{\lambda_1} \cdot \mathbf{v}_1^T \\ \sqrt{\lambda_2} \cdot \mathbf{v}_2^T \\ \vdots \\ \sqrt{\lambda_d} \cdot \mathbf{v}_d^T \end{pmatrix}. \tag{3}$$

$\mathbf{E}$ is of shape $d \times n$ and each column represents the d-dimensional embedding for a specific data point.

In classical MDS the pairwise dissimilarities are Euclidean distances. In Isomap these distances are replaced by an approximation of the geodesic distances along the manifold: for each data sample the k nearest neighbors are calculated and connected to form a neighborhood graph. The distance between two data samples is now its shortest path distance along the graph.

Data samples that have not been used for the initial embedding computation can easily be projected into the embedding space for MDS as well as Isomap. This process is referred to as *out-of-sample embedding* [5]. Let $\mathbf{d}$ denote the column vector of distances from a new data sample $\mathbf{x}$ to all samples used for embedding (geodesic distances in the case of Isomap) and $\mathbf{m}$ the mean of each column in $\mathbf{D}^2$, then the embedding $\mathbf{e}$ for $\mathbf{x}$ is obtained by computing

$$\mathbf{e} = \frac{1}{2}\mathbf{E}^{\#}\left(\mathbf{m} - \mathbf{d}^2\right), \text{ where} \tag{4}$$

$$\mathbf{E}^{\#} = \begin{pmatrix} \lambda_1^{-\frac{1}{2}} \cdot \mathbf{v}_1^T \\ \lambda_2^{-\frac{1}{2}} \cdot \mathbf{v}_2^T \\ \vdots \\ \lambda_d^{-\frac{1}{2}} \cdot \mathbf{v}_d^T \end{pmatrix}. \tag{5}$$

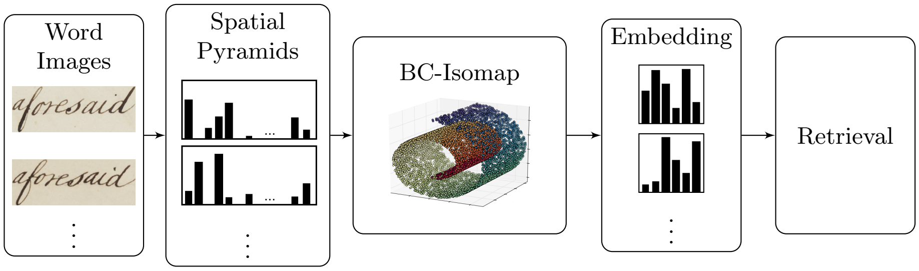


Fig. 1: The figure displays the pipeline of our BC-Isomap method.

3 Method

Using LSA leads to noticable performance drops when applied in a word spotting scenario. We believe the main reason for this to be that the singular value decomposition used in LSA assumes an Euclidean metric on the input data. This distance measure has already been shown to not perform well on histogram representations [6].

Based on this observation, we propose the use of a dimensionality reduction technique that does not assume an Euclidean metric on the input data. While the use of a manifold learning approach appears to be a well suited solution here, we will show that it performs poorly on this task as well. The main reason for this is that the standard manifold algorithms expect real valued data. We will show that treating the histogram representations as residing in $\mathbb{R}^n$ leads to an insufficient approximation of the geodesic distances and subsequently to bad embeddings. Thus, we propose to combine Isomap and a local metric which is suitable for spatial pyramid representations.

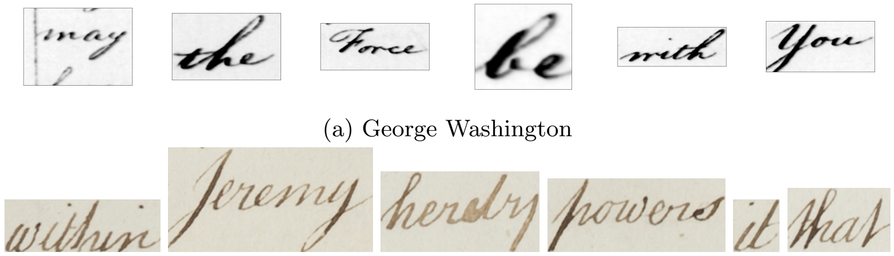
The standard metric for histogram comparison in word spotting has been the Cosine distance [2,3,10]. For other image retrieval tasks, such as [6], the L1 and L2 norms are used. Other discrete distributions, i.e. *Local Binary Pattern (LBP) histograms* [1], are often times compared by the χ^2 distance. Given two histograms $\mathbf{a}$ and $\mathbf{b}$ the χ^2 distance is obtained by

$$\chi^2(\mathbf{a}, \mathbf{b}) = \sum_i \frac{(\mathbf{a}_i - \mathbf{b}_i)^2}{\mathbf{a}_i + \mathbf{b}_i} \tag{6}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the i-th elements of the respective histograms.

Though the χ^2 distance leads to good results, this metric is not well suited for spatial pyramid comparison in a word spotting scenario. Opposed to LBP histograms, spatial pyramids are very sparse quite frequently which leads to multiple zero-divisions when applying the χ^2 distance metric. This problem is accounted for by the Bray Curtis distance:

$$BC(\mathbf{a}, \mathbf{b}) = \frac{\sum\limits_i |\mathbf{a}_i - \mathbf{b}_i|}{\sum\limits_i \mathbf{a}_i + \mathbf{b}_i}. \tag{7}$$



(a) George Washington

(b) Bentham

Fig. 2: Sample word images for the a) George Washginton dataset and b) Bentham validation dataset.

Here, no zero-division occurs when assuming that one of the histograms compared contains at least one non-zero entry. To the best of our knowledge, the BC distance has not been used in a computer vision context before.

As will be shown in the following section, the Bray Curtis distance emerges as most suitable metric for spatial pyramid comparisons on the tested benchmarks. Thus, we use this metric instead of the Euclidean distance to compute nearest neighbors and their approximate geodesic distance. Subsequently, we will term our approach *Bray Curtis Isomap (BC-Isomap)*.

The pipeline for our method is outlined in figure 1. First, a spatial pyramid is extracted for each word image. Afterwards, a nearest neighbor graph is extracted from the spatial pyramids where the nearest neighbor distance is calculated with the Bray Curtis distance metric. MDS is used on the geodesic distances computed from the graph to find an embedding that preserves these distances. The embedded representations are then used to perform word spotting. Please note that after embedding the word image representations reside in an Euclidean space. Thus the Euclidean distance has to be used in order to perform word spotting.

4 Experiments

4.1 Datasets and Implementation Details

For the following experiments we are going to use two datasets. The first is the *George Washington dataset (GW)* [7]. It consists of a 20 page excerpt from a bigger collection of letters by George Washington and his associates. The corresponding ground truth contains 4860 words. As the writing style does not exhibit large variations, it is widely considered a single writer scenario [4, 8]. Sample word images from the George Washington database can be seen in figure 2a. We follow the evaluation protocol used in [3] and [4] with minor modifications: each segmented word image is used once as a query to retrieve a ranked list of the remaining word images. Words which appear only once in the dataset are not used as queries. In order to generate a spatial pyramid representation for each word, SIFT descriptors are extracted in a dense grid with a step size of 5 pixels and a descriptor size of 40 × 40 pixels. The descriptors are then clustered into a visual vocabulary of size 4096. This descriptor and quantization parametrization has already been shown to produce competitive results [8, 10]. A two level spatial pyramid is then constructed from the quantized descriptors with a global Bag-of-Features histogram in the first level and a left and right partition in the second level as is done in [10]. While in [10] each partition is weighted by the amount of partitions on the corresponding level, we found that weighting by the square of partitions gives slightly better results. This way, the spatial pyramid's bins with finer resolution are weighted higher than those with a coarser resolution.

The second dataset is the validation subset of the Bentham benchmark used in the 2015 Keyword Spotting for Handwritten Documents competition which was conducted as part of the 2015 International Conference on Document Analysis and Recognition[1]. It consists of 95 dedicated query word images and 3234 test word images. A subsample of the test words can be seen in figure 2b. Just as with the GW dataset, we densely extract SIFT descriptors at a single scale and pool them into spatial pyramids. In a preliminary experiment we found descriptor sizes of 24 × 24 at a step size of 2 pixels to work well. Additionally, smaller visual vocabularies generally performed better than larger ones. Here, we found codebooks of size 1024 to work the best. The spatial pyramid itself has two levels with the first level being split into a 2 × 3 grid and the second level into a 2 × 9 grid.

As a baseline, we extract the spatial pyramids from the word images of each dataset, perform a tf-idf transform and reduce the dimensionality of the resulting representation with LSA. The resulting lower dimensional representations are then compared using the Cosine distance metric. For each query q the *Average Precision (AP)* is calculated by

$$AP(q) = \frac{\sum\limits_{i=1}^{s} P(q,i) \cdot rel(q,i)}{\sum\limits_{i=1}^{s} rel(q,i)}, \tag{8}$$

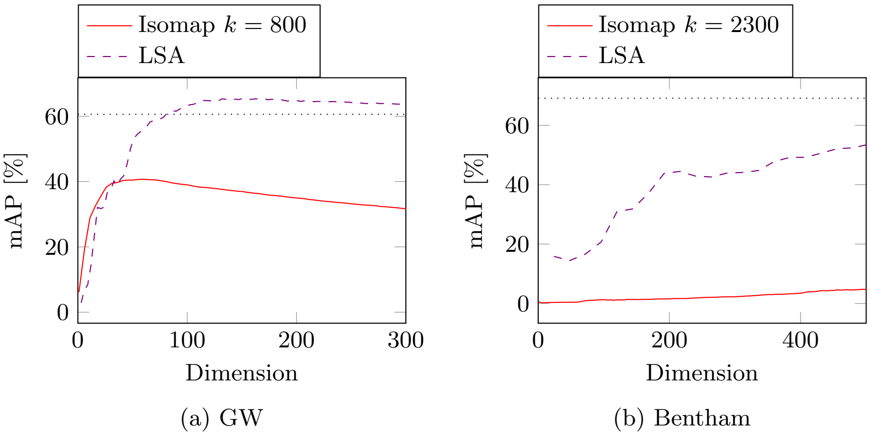
where $rel(q,i)$ is an indicator function that evaluates to 1 if the element at i is relevant w.r.t. q and 0 otherwise, $P(q,i)$ represents the precision of the retrieval list for query q when cut off at i elements and s is the length of the retrieval list. Please note that the retrieval list is not cut off at any point which leads to a recall of 100 %.

The *mean Average Precision (mAP)* then evaluates to the mean of all queries.

4.2 Standard Isomap

The first experiment evaluates the practicability of the standard Isomap to reduce the dimensionality of the spatial pyramids. Figure 3 shows the results for

[1] http://transcriptorium.eu/~icdar15kws/data.html



(a) GW

(b) Bentham

Fig. 3: The figure displays the different mAP values when applying standard Isomap and LSA to the two datasets. The dotted black line indicates the mAP without any dimension reduction.

this approach with an exemplary parametrization compared to reducing the dimensionality with LSA. As already hinted at in section 3, this manifold learning approach performs poorly compared to LSA which holds true for all parametrizations tested (please refer to the supplemental material for a complete evaluation). The major reason for this is the nature of the data: using a 12 288 dimensional spatial pyramid for the GW dataset and a 24 756 spatial pyramid for the Bentham dataset, both input spaces are sparsely sampled. The path lengths along the nearest neighbor graphs appear not to be a good approximation of the geodesic distance as the underlying manifold is not sampled densely enough.

4.3 Distance Metric Evaluation

In the second experiment, we will provide evidence for our claim that the BC distance is the metric best suited for word spotting on our benchmarks.

Figure 4 shows the mAP for the two datasets when applying no dimension reduction and sorting the retrieval list according to the individual metrics. As expected, the L1 and L2 norm fall short of the results obtained by the Cosine distance on both datsets. However, the BC distance is able to outperform all other metrics which were evaluated.

4.4 Bray Curtis Isomap

In the third experiment, we apply the proposed BC-Isomap to the spatial pyramid representations and conduct word spotting on the embedding representations. Additionally, we use the BC distance metric in combination with MDS to embed the word image descriptors. We will dub this combination *Bray Curtis*

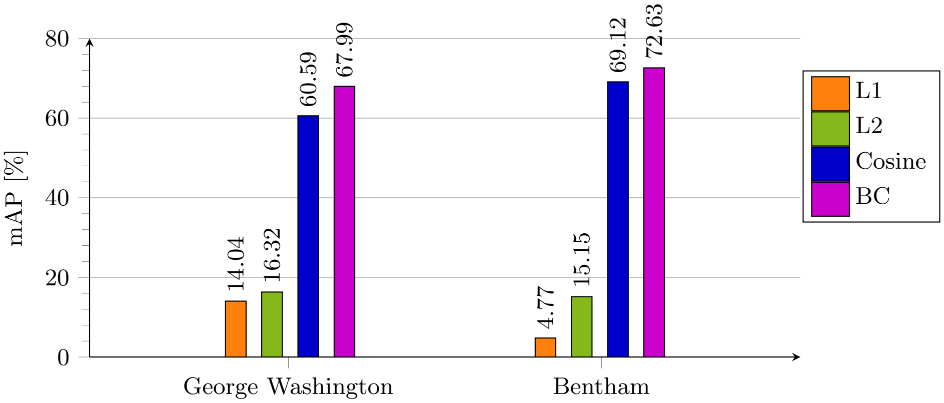


Fig. 4: The figure displays the mAP values when sorting the retrieval lists by the specified metrics for the two datasets (no dimension reduction is applied).

MDS (BC-MDS). In order to give a fair comparison between the baseline and the proposed method, we will compare the representations obtained with LSA with the BC distance metric as well.

Figure 5 compares the retrieval results of the low-dimensional representations obtained from LSA, BC-MDS and BC-Isomap. As can be seen in the figure, the LSA results are the worst on both datasets for smaller dimensions. LSA is only able to outperform the BC-Isomap results on the GW dataset when the dimensionality gets higher. For the Bentham dataset it can only outperform the manifold learning approach when the parameter k is set to very small values. LSA is not able to achieve better results on either dataset when compared to BC-MDS for any embedding dimension. On both datasets BC-Isomap is able to obtain higher mAP values when the dimensionality is low but gets outperformed by BC-MDS with a rising number of dimensions. Please note that the plots for BC-Isomap in figure 5b stop at dimension 450. This is due to the eigenvalue decomposition yielding negative results after the first 450 eigenvalues (see equation 3). For a complete comparison of all BC-Isomap parametrizations with BC-MDS and LSA please refer to the supplemental material.

Table 1 lists the mAP results for LSA, BC-MDS and different BC-Isomap parametrizations when setting the dimensionality of the embedding to 0.4 % of the original spatial pyramid dimension. For the George Washington benchmark, the mAP is improved by an absolute value of 18.29 % when comparing BC-Isomap to LSA and still 4.86 % compared to no dimension reduction. While for the Bentham validation dataset the retrieval precision of the standard spatial pyramid could not be surpassed, the LSA results were improved by 41.29 %.

4.5 Discussion

The results presented in the previous section show that both BC-MDS and BC-Isomap are superior to LSA when applied to spatial pyramids in a word spotting scenario. For the George Washington dataset, the modified Isomap algorithm is able to achieve the same mAP values compared to no dimension reduction at an embedding dimensionality of 16. This is 0.12 % of the original representation size. The manifold learning approach is also fairly robust with respect to its parameters (figure 5a, please refer to the supplemental material for a complete evaluation of $k = 300$ to $k = 2300$).

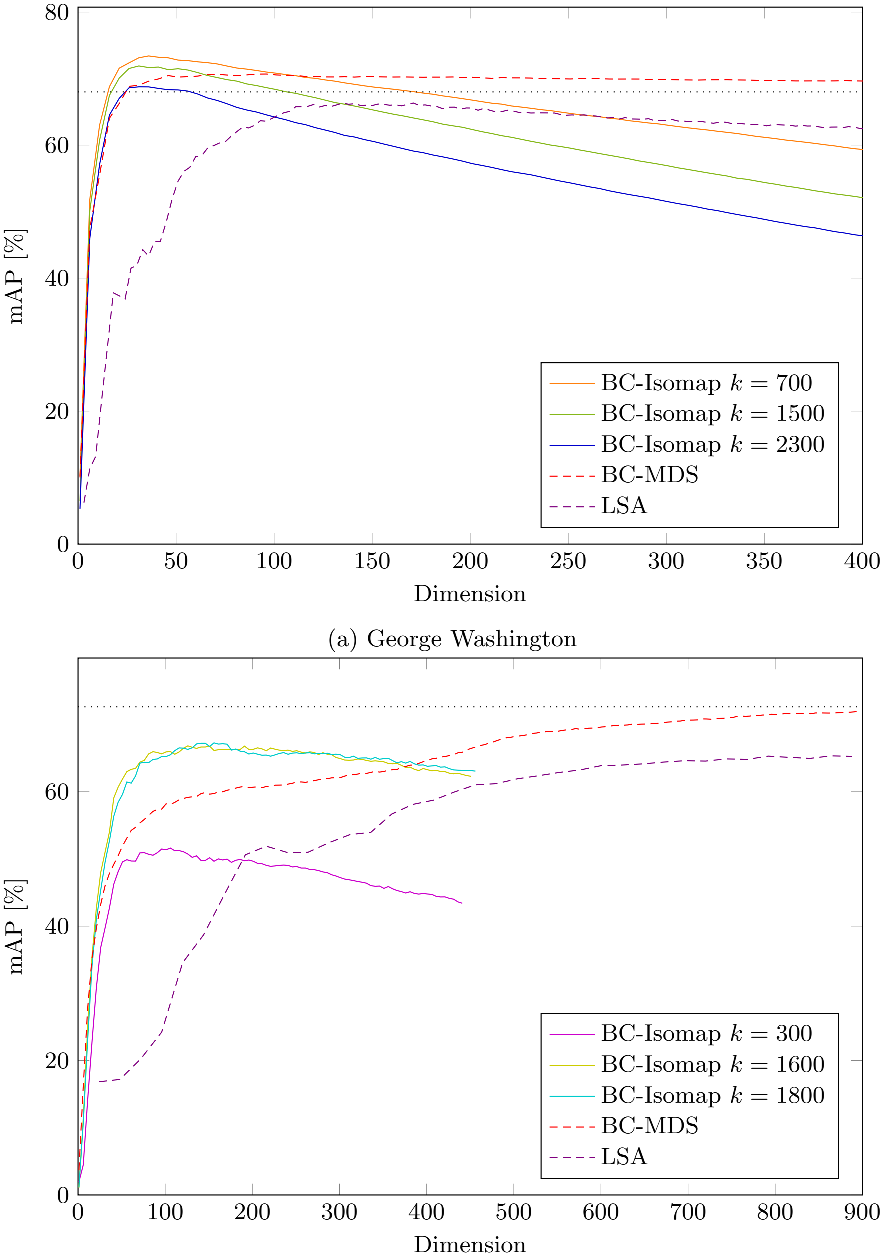
For the Bentham dataset, the retrieval precision of BC-MDS converges to the mAP value of the plain spatial pyramids with increasing dimensionality (figure 5b). When reducing to smaller dimensions, BC-Isomap outperforms the other two approaches. As with the George Washington benchmark, the parameters are fairly stable to even a medium amount of change (table 1, figure 5b). While neither dimension reduction technique is able to achieve the same mAP value compared to using no dimension reduction, it should be noted that using a 24 576-dimensional word image representation to obtain the best mAP possible is more of an academic than a practically applicable solution. The Bentham validation set contains merely 3234 segmented word images and is only a small subset of the overall Bentham collection which contains 60 000 manuscripts and an estimated 30 000 000 words. Performing word spotting with the standard spatial pyramid would become virtually impossible on this task.

5 Conclusion

In this paper, we presented Bray Curtis Isomap which is an extension of the Isomap manifold learning algorithm. This extension is able to deal will high-dimensional histogram representations in a sparsely sampled input space such as spatial pyramids. These representations occur quite frequently in a word spotting context. The resulting low-dimensional embedding is able to outperform the

Table 1: mAP values when reducing to 0.4 % of the original size

Method	GW mAP @ dim. 50	Bentham mAP @ dim. 100
No Dim. Reduction	67.99	72.63
LSA	54.56	24.23
BC-MDS	70.22	58.19
BC-Isomap $k = 500$	**72.85**	61.07
BC-Isomap $k = 900$	72.64	57.50
BC-Isomap $k = 1300$	71.92	61.82
BC-Isomap $k = 1700$	70.97	**65.52**



(a) George Washington

(b) Bentham

Fig. 5: The figure displays the different mAP values for different neighborhood sizes k when reducing to a certain dimension for a) the George Washington dataset and b) the Bentham validation dataset. The dotted black line indicates the mAP without any dimension reduction. Please note that the BC-Isomap plots in b) stop at 450 dimensions as this was the maximum dimension for embedding (the eigenvalue decomposition yielded negative eigenvalues for larger dimensions).

commonly used Latent Semantic Analysis on the George Washington and Bentham datasets. We contribute this improvement to the use of the Bray Curtis distance metric. Opposed to the Euclidean distance metric used in LSA, BC-Isomap bases its embedding on the BC distance which is a metric specifically designed for histogram representations. Additionally, the non-linear projection is able to uncover more complex structures than its linear counterpart.

References

1. Ahonen, T., Hadid, A., Pietik, M., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: European Conference on Computer Vision. pp. 469–481 (2004)
2. Aldavert, D., Rusinol, M., Toledo, R., Llados, J.: Integrating Visual and Textual Cues for Query-by-String Word Spotting. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. pp. 511–515 (2013)
3. Almazan, J., Fornes, A., Valveny, E.: Deformable HOG-Based Shape Descriptor. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. pp. 1022–1026 (2013)
4. Almazan, J. and Gordo, A. and Fornes, A. and Valveny, E.: Word Spotting and Recognition with Embedded Attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(12), 2552–2566 (2014)
5. Bengio, Y., Paiement, J.F., Vincent, P., Delalllaux, O., Le Roux, N., Ouimet, M.: Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering. In: Advances in Neural Information Processing Systems 16. pp. 177–184 (2004)
6. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: Computer Vision - ECCV 2010, vol. 6314, pp. 143–156 (2010)
7. Rath, T.M., Manmatha, R.: Word Spotting for Historical Documents. International Journal on Document Analysis and Recognition 9, 139–152 (2007)
8. Rothacker, L., Rusinol, M., Fink, G.A.: Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. pp. 1305–1309 (2013)
9. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290, 2323–2326 (2000)
10. Rusiñol, M., Aldavert, D., Toledo, R., Lladós, J.: Efficient segmentation-free keyword spotting in historical document collections. Pattern Recognition 48(2), 545–555 (2015)
11. Silva, V.D., Tenenbaum, J.B.: Global Versus Local Methods in Nonlinear Dimensionality Reduction. Advances in Neural Information Processing Systems 15, 705–712 (2003)
12. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
13. Zhang, Z.Y., Zha, H.Y.: Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. SIAM Journal on Scientific Computing 26(1), 313–338 (2005)