Unsupervised Metric Learning by Self-Smoothing Operator

Jiayan Jiang
UCLA
jetjiang@cs.ucla.edu

Bo Wang
University of Toronto
wangbo.yunze@gmail.com

Zhuowen Tu
UCLA and Microsoft Research Asia
ztu@loni.ucla.edu

Abstract

In this paper, we propose a diffusion-based approach to improving an input similarity metric. The diffusion process propagates similarity mass along the intrinsic manifold of data points. Our approach results in a global similarity metric which differs from the query-specific one for ranking produced by label propagation [25]. Unlike diffusion maps [7], our approach directly improves a given similarity metric without introducing any extra distance notions. We call our approach Self-Smoothing Operator (SSO). To demonstrate its wide applicability, experiments are reported on image retrieval, clustering, classification, and segmentation tasks. In most cases, using SSO results in significant performance gains over the original similarity metrics, with also very evident advantage over diffusion maps.

1. Introduction

In many vision/learning applications, data samples are modeled as high dimensional points in an ambient Euclidean space, and the distance between two samples is measured by Euclidean (or Mahalanobis) distances. It has been shown that data samples often live in a much lower-dimensional intrinsic space (i.e. the Riemannian manifold), where Euclidean assumption is valid only locally [16, 21, 15]. How to capture and utilize the intrinsic manifold structure therefore becomes a central problem in the vision and learning community.

There exists a large body of literature on manifold learning [21, 15, 7]. The idea is to explicitly construct a new embedding space with a corresponding metric which is more faithful to the manifold structure and hence induces a better distance/similarity measure. These algorithms have been applied to clustering and image segmentation [18, 12]. The same idea can be extended to semi-supervised cases, where a limited portion of data labels are given. For example, label propagation [25] uses a diffusion process to propagate the limited label information to the unlabeled data samples along the manifold. Other examples include Markov random walks on manifolds [20] and Ranking on Data Manifolds (RDM) [24]. In particular, RDM uses a similarity-induced diffusion kernel to improve the ranking result with respect to a single query. Although the above methods have been shown to be effective in classification/ranking tasks, they lack the notion of a global similarity metric, which is crucial in many applications.

From a different point of view, diffusion maps [7] define diffusion distances between data samples. An input similarity matrix is then improved through a diffusion process. The advantages of diffusion maps over the previous approaches are: (1) there is an explicit notion of a global distance/similarity metric; and (2) the diffusion step parameter t provides a natural way of doing multi-scale data analysis. Isomap [21] applies shortest-path algorithm to compute pairwise geodesic distances, which essentially defines a *min* operator on the data manifold. On the other hand, diffusion based algorithms assembles and accumulates all the paths between two samples, which defines an *average* operator on the manifold. It has been observed that the average operator might be more robust than the min operator used by Isomap [24].

Our work is closely related to diffusion maps [7]. However, instead of using the notion of diffusion distances between data samples, we work on the similarity matrix directly using the self-induced smoothing kernel. Therefore we name our approach Self-Smoothing Operator (SSO). Compared with diffusion maps, the structure of the input similarity metric is better respected and preserved. Compared with label propagation [25] or RDM [24], SSO gives a global similarity metric rather then with respect to a single query. It is also natural and intuitive to understand our approach as a smoothing process: a smoothing kernel is constructed from the input metric and the smoothing process, like diffusion, is to propagate the similarity through weighted connections. After a few steps, the metric gets "smoothed" through manifold geometry but without the need to explicitly construct the manifold, which is often a time-consuming and difficult task.

2. Related Work

Given a graph $G = (\Omega, W)$ where $\Omega = \{x_i, i = 1, \cdots, n\}$ is the space for finite nodes representing data

samples and W is a similarity matrix with each entry $W(i,j) \in [0,1]$ being the similarity between sample x_i and x_j. The higher $W(i,j)$ is, the more similar x_i is to x_j. In practice, W is often obtained from applying Gaussian kernel to a distance matrix:

$$W(i,j) = \exp\left\{-d^2(i,j)/(k\sigma^2)\right\} \tag{1}$$

where $d(i,j)$ denotes the distance between x_i and x_j, and k and σ control the width of kernel. A stochastic diffusion process on G allows local similarities to be propagated along data manifold, without explicitly constructing the manifold geometry. In the following, we discuss several related approaches which can be broadly categorized into equilibrium-based and dynamics-based approaches.

A most well-known equilibrium-based approach is the PageRank algorithm [14] which exploits the global hyperlink structure of the web. The transition kernel is given by row-wise-normalizing the similarity matrix W which encodes the outbound link information:

$$P = D^{-1}W \tag{2}$$

where D is a diagonal matrix with $D(i,i) = \sum_{k=1}^{n} W(i,k)$. Let us assume that P has eigenvalues $|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}|$ with associated left and right eigen-vectors: $\phi_i^T P = \lambda_i \phi_i^T$ and $P\psi_i = \lambda_i \psi_i$ ($i = 0, \cdots, n-1$). If the kernel is ergodic and aperiodic, then only $|\lambda_0| = 1$ and the diffusion process

$$f_{t+1} = P^T f_t \tag{3}$$

converges to the first left eigen-vector $f_{t\to\infty} = \phi_0$, as long as the initial f_0 is not orthogonal to ϕ_0. PageRank then uses ϕ_0 to rank all the pages on the web.

Another example is the RDM algorithm [24] to improve ranking results with respect to a query. Instead of constructing a transition kernel P, RDM constructs a symmetrically normalized matrix $S = D^{-1/2}WD^{-1/2}$. Although S is not a transition matrix because none of its rows or columns are normalized, it encodes the same global information as P does. The query information is injected by a vector:

$$y = [y_1, \cdots, y_n]^T \text{ where } y_i = \begin{cases} 1 & \text{if } x_i \text{ is a query} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Then the diffusion process is simply:

$$f_{t+1} = \alpha S f_t + (1-\alpha) y \tag{5}$$

where α is a parameter in $[0,1)$. Note the first term of r.h.s. in (5) defines a global diffusion process similar to (3), and the specific query information is injected by the second term in each diffusion step. It can be shown that the equilibrium point has a closed form $f_{t\to\infty} = (I - \alpha S)^{-1} y$.

One example of dynamics-based approaches is label propagation [25], which is widely used to solve the semi-supervised learning problem. It propagates each labeled sample's label information to its neighboring samples. Here we give a formulation in the context of retrieval [5]. Label propagation uses the same row-wise normalized transition kernel (2). With an initial $f_0 = y$, it iterates between the following two steps: (1). $f_{t+1} = P f_t$; (2). $f_{t+1}(i) = 1$ if $y_i = 1$. Note the second clamping step injects query information into the diffusion process, and the first step is very similar to but different from (3). Actually, because P is row-wise normalized, we have $\mathbf{1} = P\mathbf{1}$ where $\mathbf{1}$ denotes a constant vector. Therefore the iterations have to be stopped before the equilibrium is achieved, and the step parameter t provides a natural way of doing multi-scale data analysis.

Although RDM and label propagation improves the ranking/retrieval results through a diffusion process, they lack a global notion of similarity/distance metrics.

3. Self-Smoothing Operator

Diffusion maps introduce a global distance metric (i.e. diffusion distances) over data samples. Given the transition kernel P in (2), the diffusion distance between x_i and x_j at step t is defined as:

$$\begin{aligned} d_t^2(i,j) &= ||p_t(i,\cdot) - p_t(j,\cdot)||^2_{1/\phi_0} \\ &= \sum_{k=1}^{n} \frac{1}{\phi_0(k)} (p_t(i,k) - p_t(j,k))^2 \end{aligned} \tag{6}$$

where $p_t(i,\cdot)$ is the i-th row of the t-step transition matrix $P_t = P^t$, and ϕ_0 is the equilibrium distribution. It can be shown that the diffusion distance can be directly computed from the diffusion map: $\Psi_t : x_i \to \left[\lambda_1^t \psi_1(i), \cdots, \lambda_q^t \psi_q(i)\right]^T$, and $d_t^2(i,j) \approx \|\Psi_t(i) - \Psi_t(j)\|^2 = \sum_{k=1}^{q} \lambda_k^{2t} (\psi_k(i) - \psi_k(j))^2$. Here ψ_k's are the right eigen-vectors of P, and $q \leq n-1$ captures the leading nontrivial eigenvalues. Since $p_{t\to\infty}(i,\cdot) = \phi_0^T$, at the equilibrium point any pairwise diffusion distance is 0, and what matters is the dynamics of the diffusion process.

It is unclear, however, how close this distance notion is related to the input similarity. For example, let us assume that the initial W (and its induced P) is close to identity. This is not uncommon because in order to reveal the underlying manifold structure of the data, only a small portion of similarity mass is distributed among the closest neighbors at the beginning. Otherwise the mixing rate of the diffusion process will be drastic and the chance of discovering the intrinsic structure becomes slim. In this case, the diffusion distances at the beginning would be almost same for any pair of data samples due to the near-orthogonal condition. At the other extreme, when the process approaches equilibrium, the diffusion distances would be exactly 0 for any pair

of data samples, for the convergence condition analyzed in the above text. The given similarity W therefore has little, if any, impact on the two extreme cases. In those intermediate steps, how W influences the diffusion distances is also largely obscure.

3.1. The method

Instead of using the notion of diffusion distances, we introduce a new smoothing operator which better respects and preserves the structure of W. We call this operator Self-Smoothing Operator (SSO). In analogy to smoothing in image processing, SSO propagates the mass of similarity to the nearest data samples. Here the smoothing kernel is not hard-coded, but induced from the input metric itself. After a few steps, the metric gets "smoothed" through manifold geometry but without the need to explicitly construct the manifold, which is often a time-consuming and difficult task. The algorithm of SSO is as follows:

1. Computing the smoothing kernel: $P = D^{-1}W$
   where D is a diagonal matrix with $D(i,i) = \sum_{k=1}^{n} W(i,k)$.
2. Performing smoothing for t steps: $W_t = WP^t$
3. Self-normalization: $W^* = \Delta^{-1}W_t$ where Δ is a diagonal matrix with $\Delta(i,i) = W_t(i,i)$.
4. PSD Projection: $\widehat{W}^* = \text{proj\_to\_psd}(W^*)$ Function $\text{proj\_to\_psd}(A) = V\text{diag}(\max(\lambda, 0))V^T$, where λ and V are the eigenvalues and eigen-vectors of $\frac{1}{2}(A + A^T)$ respectively.

Figure 1. Algorithm of Self-Smoothing Operator (SSO). The last step is optional and is only used when a positive semi-definite similarity matrix is required for subsequent algorithms (e.g. input metrics for kernel k-means).

In Step 1, a smoothing kernel P is induced from an input similarity matrix W; P is then used as a smoothing kernel in a diffusion process for t steps in Step 2; Step 3 guarantees that the diagonal entries of the smoothed similarity matrix are always 1, which reflects the identity of indiscernible for a distance metric (any entry in W^* with value greater than one will also be clamped to 1). Note that in general W_t and W^* are neither symmetric nor positive semi-definite. If such property is required for the subsequent data analysis algorithms (e.g. kernel kmeans), an optional projection step can be done in the last step.

As in other manifold learning algorithms, SSO is based on the assumption that long-range similarities can be approximated by an accumulation of local similarities. SSO is not expected to work well when the assumption fails or the local similarities can not be reliably estimated.

The difference between SSO and label propagation, RDM, or other query-specific methods [5] is that SSO induces a global similarity metric and hence is non-query-specific. Unlike diffusion maps, the similarity metric given by SSO is a more direct improvement over the input similarity W, without introducing any additional distance notions.

The improved similarity W^* can be used in numerous applications. For example, given W^* retrieval can be done on a per-row basis [5]. In other words, for each row (a query), the similarity scores are sorted in descending order and the first K (retrieval window) items are returned. When label information is available for the retrieved items, classification can be done by majority voting inside the K-NN window. For clustering, kernel kmeans [17] can be employed to deal with the similarity matrix directly. Because positive-semidefinite-ness is required for the convergence of the algorithm, we use the projected version $\widehat{W}^*$ as the input to kernel kmeans.

Given an input similarity matrix W, the only parameter in SSO is the step t. As in other dynamics-based approaches, t controls the scale at which the data are analyzed. Return to our analogy to image de-noising, performing smoothing on a noisy image naturally increases the signal-to-noise ratio. However, if too much smoothing is done, the image becomes over-smoothed and too much information is lost. In this regard, t in SSO has the same effect in improving the similarity. When necessary, t can be a fractional number in Step 2. In this case, a complex W_t can emerge and only its real part is used subsequently.

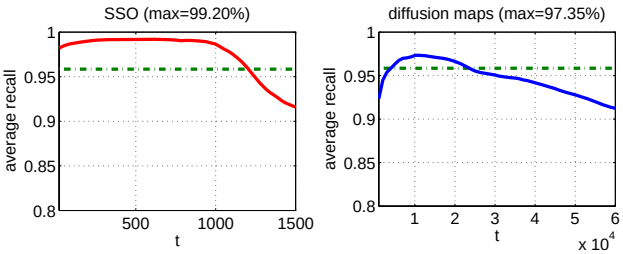


Figure 2. A running example. Bull's eye score (retrieval rate with window size $K = 40$) over t on the MPEG7 dataset. Dash-dot lines are the baseline from the input similarity.

Fig. (2.a) shows a running example where the result is evolving during the SSO process in a retrieval task. The performance consistently improves and is stable over a wide range of t, and then it slowly drops down. Eventually it will converge to a degenerated (over-smoothed) point where the retrieval result stays constant over all queries, because

$$W_{t\to\infty} = W \times (\mathbf{1}\phi_0^T) \propto \phi_0\phi_0^T, \tag{7}$$

where ϕ_0 is the equilibrium distribution of the smoothing kernel P. For a direct comparison, we also plot out the result of diffusion maps [7] in Fig. (2.b). We have the following observations: (1) SSO has the similar asymptotic behavior as diffusion maps as both are dynamics-based approaches; (2) the parameter t in SSO is not very sensitive

and similar performance gain can be observed over a wide range of t (a typical choice is $t = 500 \sim 1,000$); (3) SSO improves over the input metric immediately whereas diffusion maps has to pick up the performance from the point where it is even worse than the original input; and (4) SSO delivers results significantly better than diffusion maps. In the next section, we report more experimental results by SSO on a wide variety of applications, which further confirms the above observations.

3.2. Justification

In this section, we provide some justification for the intuition and motivation of SSO. Consider the following 3×3 toy example: Suppose the ground-truth $W = [1, 0.9, 0.9; 0.9, 1, 0.9; 0.9, 0.9, 1]$ is corrupted to $\hat{W}$, whose entries $\hat{W}(1,3) = \hat{W}(3,1) = 0$. The Frobenius-norm of the difference is 1.27. Based on $\hat{W}$ and its induced kernel P, one iteration of smoothing rectifies $\widetilde{W}_1(1,3) = \sum_{i=1}^{3} \hat{W}(1,i) \times P(i,3) = 0.29$, and the F-norm of difference is decreased to 0.35 after 5 iterations of SSO.

As in the unsupervised manifold learning cases [21, 15], local distances are mostly accurate and long distances are problematic. Assume we are given an input similarity matrix W with one entry being inaccurate (corrupted) $W_{13} \rightarrow \hat{W}_{13}$ ($\hat{W}_{13} = \hat{W}_{31}$).

$$W = \begin{pmatrix} W_{11} & W_{12} & \hat{W}_{13} \\ W_{21} & W_{22} & W_{23} \\ \hat{W}_{31} & W_{32} & W_{33} \end{pmatrix}, and$$

$$P = \begin{pmatrix} W_{11}/D_{11} & W_{12}/D_{11} & \hat{W}_{13}/D_{11} \\ W_{21}/D_{22} & W_{22}/D_{22} & W_{23}/D_{22} \\ \hat{W}_{31}/D_{33} & W_{32}/D_{33} & W_{33}/D_{33} \end{pmatrix},$$

where $D_{ii} = \sum_{j=1}^{3} W_{ij}$. For simplicity of the illustration, we let the distance measure $d_{ij} = 1 - W_{ij}$. The corruption of entry $W_{13} \rightarrow \hat{W}_{13}$ results in $d_{12} + d_{23} \ll \hat{d}_{13}$ (equivalently $W_{12} + W_{23} \gg \hat{W}_{13}$), which violates the triangular inequality.

Observation: Assume $(1,2)$ and $(2,3)$ are neighbors, then process of SSO, $W^{(t+1)} = W^{(t)}P$, improves the neighborhood situation between $(1,3)$ with corrupted entry $\hat{W}_{13}$.

Proof: Our neighborhood assumption implies $W_{12} > 0.5$; $\hat{W}_{13} < 0.5$ due to the corruption. Without loss of generability, we further assume $W_{12} = W_{23}$. One round of smoothing gives:

$$\begin{aligned} W_{12}^{(1)} &= \frac{W_{11} \cdot W_{12}}{D_{11}} + \frac{W_{12} \cdot W_{22}}{D_{22}} + \frac{\hat{W}_{13} \cdot W_{32}}{D_{33}} \\ W_{23}^{(1)} &= \frac{W_{21} \cdot \hat{W}_{13}}{D_{11}} + \frac{W_{22} \cdot W_{23}}{D_{22}} + \frac{W_{23} \cdot W_{33}}{D_{33}} \\ W_{13}^{(1)} &= \frac{W_{11} \cdot \hat{W}_{13}}{D_{11}} + \frac{W_{12} \cdot W_{23}}{D_{22}} + \frac{\hat{W}_{13} \cdot W_{33}}{D_{33}} \end{aligned} \quad (8)$$

Thus,

$$\begin{aligned} & (d_{12}^{(1)} + d_{23}^{(1)} - d_{13}^{(1)}) - (d_{12} + d_{23} - \hat{d}_{13}) \\ = & \frac{2W_{12}^2}{1 + W_{12} + \hat{W}_{13}} + \frac{2\hat{W}_{13}}{1 + W_{12} + \hat{W}_{13}} + \frac{W_{12}^2}{1 + 2W_{12}} \\ & - \frac{2W_{12}}{1 + 2W_{12}} - \hat{W}_{13}, with\ W_{12} > \hat{W}_{13} \\ > & \frac{3W_{12}^2 + 2\hat{W}_{13} - 2W_{12}}{1 + 2W_{12}} - \hat{W}_{13} \\ \geq & 0, \ when \ W_{12} \geq \frac{\hat{W}_{13} + 1 + \sqrt{\hat{W}_{13}^2 - \hat{W}_{13} + 1}}{3} \end{aligned} \quad (9)$$

$W_{12} \geq \frac{\hat{W}_{13}+1+\sqrt{\hat{W}_{13}^2-\hat{W}_{13}+1}}{3}$ is not a hard condition to meet, e.g. when $\hat{W}_{13} = 0.4$ then it requires $W_{12} \geq 0.76$.

4. Experiments

To demonstrate the general applicability of SSO, experiments on image retrieval, clustering, classification, and segmentation are reported in this section. In most cases, using SSO leads to significant improvement over the input similarity matrix and the advantage over diffusion maps [7] is also very evident.

4.1. Image Retrieval

Three datasets are used in image retrieval. The first one is the MPEG7 dataset [10], which consists of $1,400$ silhouette shape images. These images are evenly distributed over 70 classes with 20 shapes in each class. For a given similarity matrix, retrieval is benchmarked using the average recall rate at a window size $K = 40$ for each query shape, also known as the bull's eye score.

SSO is applied to a linearly combined similarity matrices $W = \frac{1}{3}(W_{SC} + W_{IDSC} + W_{DDGM})$, where W_{SC}, W_{IDSC}, and W_{DDGM} are obtained by applying eqn. (1) to Shape Context distances [6], IDSC distances [11], and Data-Driven Generative Models [22] computed over the dataset respectively. In computing eqn. (1), we set $k = 0.1$ and σ^2 is estimated by the average distance from each shape to its 30 nearest neighbors. The baseline bull's eye score of the input W on SC+IDSC+DDGM is 95.84% (92.77% on SC+IDSC).

The dynamics of SSO and diffusion maps are shown in Fig. (2.a) and (2.b) respectively. We observe that SSO improves the input similarities over a wide range of t. The improvement peaks around 500 steps, and thereafter the performance drops slowly. In particular, a maximum score of 99.20% is obtained by SSO, which translates to 80.77% relative reduction in error from W. Table 1 lists some reported bull's eye scores on the dataset. Note that the highest score reported so far was 97.72% in [4]. The focus of [4] is however the fusion of two (limited to two) input similarity mea-

SC [6]	84.87%
IDSC [11]	86.81%
DDGM [22]	80.86%
[5] on IDSC	91.61%
[23] on IDSC	93.32%
(SC+IDSC)	92.77%
[5] on (SC+IDSC)	92.92%
DM on (SC+IDSC)	92.07%
SSO on (SC+IDSC)	97.64%
(SC+IDSC+DDGM)	95.84%
DM on (SC+IDSC+DDGM)	97.35%
SSO on (SC+IDSC+DDGM)	**99.20%**

Table 1. Some reported bull's eye scores on MPEG7. SSO achieves 99.2% based on a direct linear combination of three input similarity measures. On SC+IDSC, the diffusion maps method does not improve over the input similarity measures at all and it has worse results than SSO overall.

sures whereas we here study a metric learning algorithm, which is more general (applicable in retrieval, clustering, segmentation, classification, etc.) than the transduction-based framework in [5, 4]. With a much simpler linear combination, SSO has achieved a 99.2% bull's eye score on MPEG7.

On the other hand, diffusion maps do not perform as well as SSO in this task. The first diffusion step actually worsens the retrieval results. On SC+IDSC, the diffusion maps method does not improve the input similarity measures at all. This is because the notion of diffusion distance does not have a direct connection to the input similarity, especially in the initial phase of diffusion as we discussed in Section 3. Although afterwards the scores rise and peak around $10,000$ diffusion steps, they are significantly inferior to the results by SSO overall.

Fig. (3) shows some retrieved shapes by the input W and W^* after 500 SSO diffusion steps. It can be seen that for W there are many false positives in the first 10 retrieved shapes for some queries. For example, six of the top 10 results of query `spoon` are actually `jar`. Those false positives are eliminated by W^*. Furthermore, for the correct retrievals, the ranking makes more sense for W^* than for W: The more similar a intra-class shape is to the query, the higher position it has in the retrieved list.

The second dataset we used in retrieval is Tari1000 [3], which is another commonly used shape dataset. Its configuration is mostly the same as the MPEG7 dataset, except that there are 50 classes, each one covering 20 different shapes. Because the bull's eye score (retrieval window size $K = 40$) saturates in this dataset, a more strict $K = 20$ is used in this experiment. The input $W = \frac{1}{2}(W_{SC} + W_{IDSC})$. Other experimental settings are the same as the above. Fig. (4) shows the dynamics of SSO and diffusion maps. In this case, although diffusion maps (98.02%) improve over the baseline (95.49%), it is outperformed by SSO again, which attains a maximum score of 98.92%.

The last dataset we tested in retrieval is the N-S image dataset [19]. It is a large-scale natural image dataset, consisting of $2,550$ objects/scenes, each of which is imaged from 4 different viewpoints. Hence there are $10,200$ images in total. The initial similarity matrix is obtained by applying eqn. (1) to a distance matrix computed from [9]. Here the intention is not to provide the best results on this dataset, but rather to demonstrate the general effectiveness of SSO on improving the retrieval results. With a window size $K = 4$, the initial similarity leads to an average recall rate of 79.72%, and is improved to 81.04% by SSO, as shown in Fig. (5). The improvement is smaller compared to those on the other two datasets, because local affinities are not reliably estimated in the dataset, thus violating the assumption of SSO: A nearest-neighbor classification leads to an accuracy of 85.86%, compared to 99.43% for the MPEG7 dataset and 99.80% for the Tari1000 dataset. Nevertheless, SSO still improves over this "noisy" initial condition, and diffusion maps seem not work at all in this case.

4.2. Image Clustering

We also used the previous MPEG7 and Tari1000 datasets, along with the ORL face image dataset for the illustration of clustering. ORL has 40 subjects with 10 gray-scale images per subject. Slight variations of pose, illumination, and expression are present.Each image is first down sampled to size 16×16 and then normalized to 0-mean and 1-variance. The input similarity matrix W is computed by applying eqn. (1) to the pairwise Euclidean distances of the

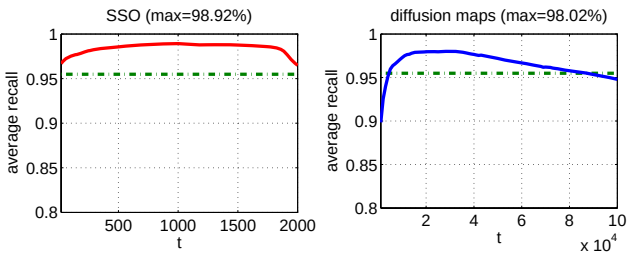


Figure 4. Retrieval rate at a window size $K = 20$ over t on the Tari1000 dataset.

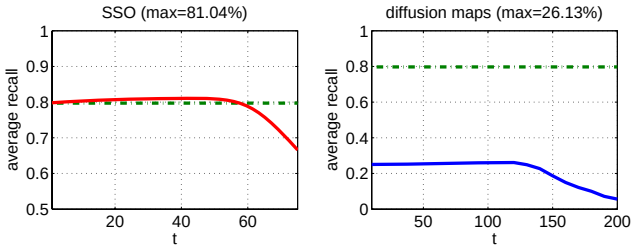


Figure 5. Retrieval rate at a window size $K = 4$ over t on the N-S dataset.

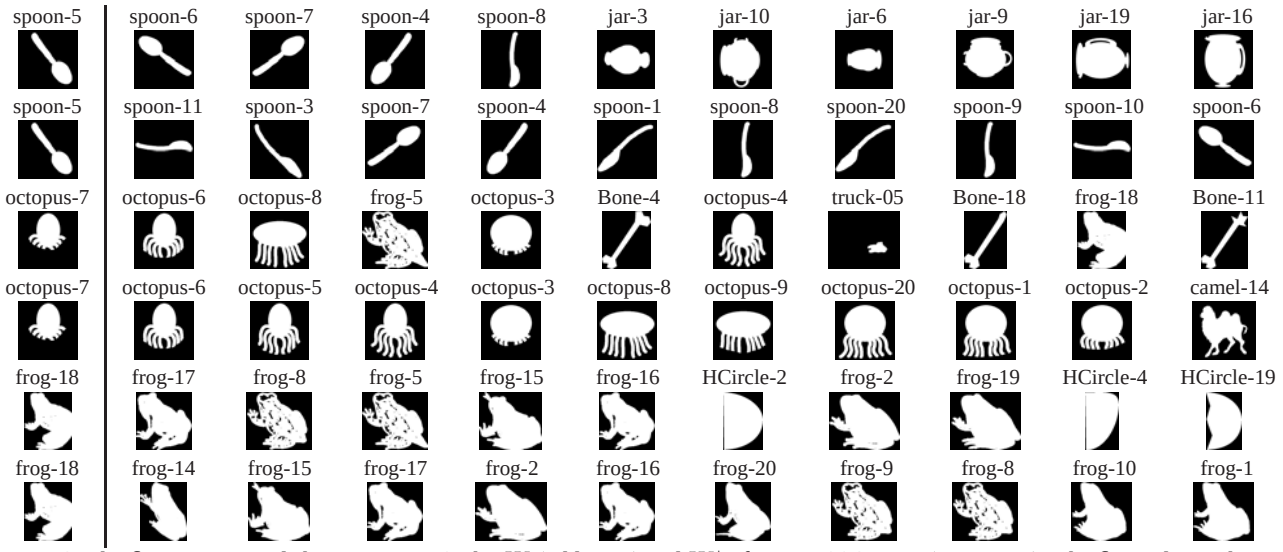


Figure 3. The first 10 retrieved shapes in MPEG7 by W (odd rows) and W^* after 500 SSO steps (even rows). The first column shows the query shape.

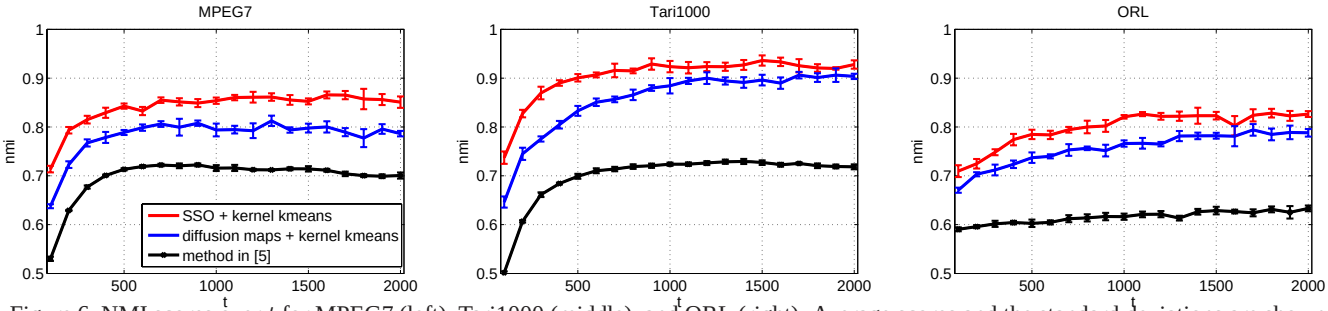


Figure 6. NMI scores over t for MPEG7 (left), Tari1000 (middle), and ORL (right). Average scores and the standard deviations are shown. Note that the method in [5] uses affinity propagation [8] to handle the non-global similarity metric it produces, but the performance is significantly inferior.

normalized images. In eqn. (1), we used $k = 0.1$ and σ^2 was estimated by the average distance from each image to its 15 nearest neighbors.

Normalized mutual information (NMI) is used for a quantitative measure of the clustering results. The ground-truth class partition Γ and the returned cluster partition Δ define a confusion matrix with each entry $n_i^{(j)}$ being the number of data samples in cluster i and class j, and n is the total number of samples. Then NMI is computed as follows:

$$\frac{2\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_i^{(j)}}{n}\log\frac{n_i^{(j)}n}{\sum_{k=1}^{I}n_k^{(j)}\sum_{k=1}^{J}n_i^{(k)}}}{H(\Gamma)+H(\Delta)}$$

where I is the number of clusters and J is the number of classes. $H(\Gamma) = -\sum_{i=1}^{I}\frac{n_i}{n}\log\frac{n_i}{n}$ and $H(\Delta) = -\sum_{j=1}^{J}\frac{n^{(j)}}{n}\log\frac{n^{(j)}}{n}$ are the entropies of partition Γ and Δ, respectively. A high value of NMI indicates that Γ and Δ match well. Kernel kmeans is used in clustering given a similarity matrix, and we set $I = J$ in the experiments. Five trials were conducted to account for stochastic variations.

The NMI scores of SSO and diffusion maps over t are plotted in Fig. (6). Note that SSO leads to a clear advantage over diffusion maps. For completeness, we also tested with another popular clustering algorithm affinity propagation [8]. Affinity propagation does not require a positive semi-definite similarity metric and was used in [5] to cluster images given a similarity matrix constructed by independently applying label propagation to each query image. As is discussed in Section 3, this procedure does not lead to a global similarity metric. The result of this scheme (SSO without the optional projection step, followed by affinity propagation) is also included in Fig. (6), which is significantly inferior to either SSO or diffusion maps. Furthermore, it is difficult to specify the number of clusters in affinity propagation, and it tends to over-cluster in most cases. For example, using the default settings, affinity propagation leads to about 100 clusters on the ORL dataset. This justifies our statement for SSO being more general than transduction-based approaches such as [5, 4].

Fig. (7) visualizes W and $\widehat{W}^*$ after 200 SSO steps on the

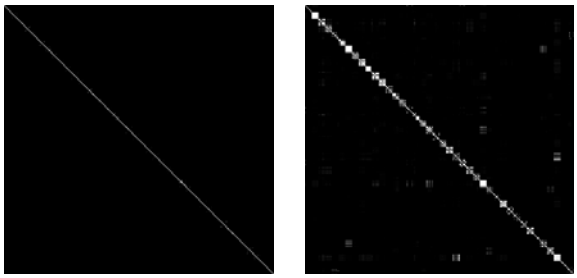


Figure 7. The input similarity W (left) and the new $\widehat{W}^*$ (right) after 200 SSO steps on the ORL dataset. The block-diagonal structure of 40 clusters begins to emerge in $\widehat{W}^*$.

ORL dataset. As is discussed in Section 3, the near-identity characteristic of W is crucial for the revelation of the underlying manifold structure in the process of diffusion. It can be observed that the block-diagonal structure of 40 clusters is discernible in $\widehat{W}^*$.

4.3. Image Segmentation

The Berkeley image dataset [13] was used to test SSO on the image segmentation task. We only used the `test` subset in this experiment, which contains 100 color natural images. These images are first converted to gray scale for the ease of processing. We used the code in [2] to construct the input similarity matrix W across pixels inside an image. Then SSO is applied to obtain a new $\widehat{W}^*$.

Three segments are obtained for each image by clustering based on the similarity matrix. It is noted that our intention in this experiment is not to propose the best image segmentation algorithm; rather we show the general applicability of SSO and demonstrate that using improved similarity measures by SSO enhances a state-of-the-art image segmentation algorithm, Normalized Cuts [2].

To obtain a quantitative summary of the segmentation results, we used two measures proposed in [13]. In particular, we want to measure the regional consistency between two segmentations Γ (human annotation) and Δ (algorithm output). For each pixel p_i, the local refinement error is:

$$E(\Gamma, \Delta, p_i) = |R(\Gamma, p_i) \backslash R(\Delta, p_i)| \,/\, |R(\Gamma, p_i)|$$

where $R(\Gamma, p_i)$ is the region p_i belongs to in Γ, and similarly $R(\Delta, p_i)$ is the region p_i belongs to in Δ. Note that E is asymmetric, i.e. $E(\Gamma, \Delta, p_i) \neq E(\Delta, \Gamma, p_i)$. Hence Global Consistency Error (GCE) and Local Consistency Error (LCE) are defined as:

$$GCE(\Gamma, \Delta) = \frac{1}{n} \min \left\{ \sum_i E(\Gamma, \Delta, p_i), \sum_i E(\Delta, \Gamma, p_i) \right\}$$

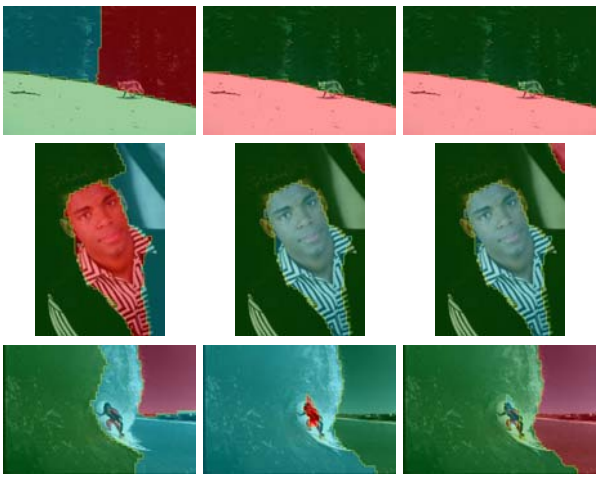
$$LCE(\Gamma, \Delta) = \frac{1}{n} \sum_i \min \left\{ E(\Gamma, \Delta, p_i), E(\Delta, \Gamma, p_i) \right\}$$

where n is the number of pixels inside an image. Naturally, we have $LCE \leq GCE$. In the dataset, there are some images with multiple annotations Γ. In these cases, we compute the error to each annotation and simply take the lowest one. Finally, the average GCE/LCE scores are reported over the whole dataset in Table 2.

The improvement of SSO to NCut is also significant (more than 25% relative reduction in both GCE and LCE). This result is encouraging considering that NCut can also be formulated as a diffusion-based approach to explore the data manifold [12], and SSO still finds room to improve the results. Fig. (8) provides more visual examples, which confirms the improvement brought by SSO.

	SSO GCE / LCE	DM GCE / LCE
$t = 0$	0.1568 / 0.1269	
$t = 5$	0.1483 / 0.1178	0.1301 / 0.1067
$t = 10$	0.1443 / 0.1134	0.1304 / 0.1077
$t = 20$	0.1420 / 0.1116	0.1385 / 0.1169
$t = 50$	0.1310 / 0.1034	0.1426 / 0.1177
$t = 100$	**0.1118 / 0.0881**	0.1482 / 0.1224

Table 2. GCE and LCE scores on the Berkeley segmentation dataset. NCut is used in clustering (3 segments) given a similarity matrix.



$t = 0$ $t = 10$ $t = 100$

Figure 8. Segmentation results of NCut with SSO. We apply SSO on the input similarity matrix and give it to the Normalized Cuts method. The number of desired segments is fixed to be 3. The figures show the segmentation results of NCut using the enhanced similarity matrix by SSO at different step t.

4.4. Medical Image Classification

Finally, SSO was tested on a MRI brain image classification task. The dataset consists of 120 brain MRI scans, which are randomly sampled from the same cross section (12 months after the start) of the publicly available ADNI dataset [1]. These scans belong to three classes:

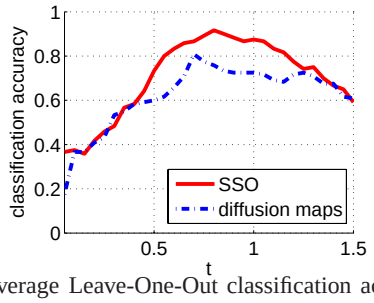


Figure 9. Average Leave-One-Out classification accuracy over t on the MRI dataset. SSO reaches a maximum accuracy of 91.67% over the original 38.33% accuracy and is significantly better than diffusion maps (80.83%).

Alzheimer's Disease (AD), its preclinical stage Mild Cognitive Impairment (MCI), and Normal, and each class covers 40 scans. All the scans were skull-stripped and aligned first, after which the cortical and subcortical structures were extracted. Six different similarity matrices were constructed out of two kinds of measures: overlap-based measure and registration-based measures. For the limit of spaces, we choose not to elaborate on all the details. Again, the purpose here is to show the general applicability of SSO.

Here we took a simple fusion scheme to combine the information in the six similarities: Each similarity matrix is diffused separately and for each step t, the average similarity matrix is used in classification. A Leave-One-Out classification setting is taken based on majority voting for each query scan with a $K = 15$ window size . We found on this dataset using a fractional diffusion step t leads to more refined classification results, as shown in Fig. (9). Again, the effectiveness of SSO (with a maximum accuracy of 91.67% over the original 38.33% accuracy) is significant (86.5% relative reduction in error), and its advantage over diffusion maps (80.83% maximum accuracy) is evident.

5. Conclusion

We have presented Self-Smoothing Operator (SSO) in this paper. The smoothing kernel is induced from an input similarity matrix, which will be directly improved through a smoothing/diffusion process along the data manifold. Our approach produces a direct global metric, which differentiates it from other diffusion-based methods, such as label propagation, RDM, or diffusion maps. The algorithm of SSO is simple and intuitive. Its effectiveness has been demonstrated on tasks of image retrieval, clustering, segmentation, and classification, across which a consistent improvement is observed. The advantage of SSO over diffusion maps is also very evident, and in many cases overwhelming. Future research includes automatic estimation of the parameter t and other metric fusion schemes along the smoothing process.

Acknowledgment: This work is supported by Office of Naval Research Award N000140910099 and NSF CAREER award IIS- 0844566.

References

[1] http://adni.loni.ucla.edu/. 7

[2] http://www.cis.upenn.edu/ jshi/software/. 7

[3] C. Aslan, A. Erdem, E. Erdem, and S. Tari. Disconnected skeleton: Shape at its absolute scale. *IEEE PAMI*, 30:2188–2201, 2008. 5

[4] X. Bai, B. Wang, X. Wang, W. Liu, and Z. Tu. Co-transduction for shape retrieval. In *Proc. of ECCV*, 2010. 4, 5, 6

[5] X. Bai, X. Yang, L. Latecki, W. Liu, and Z. Tu. Learning context sensitive shape similarity by graph transduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010. 2, 3, 5, 6

[6] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:705–522, 2002. 4, 5

[7] R. Coifman and S. Lafon. Diffusion maps. *Applied and Comp. Harmonic Ana.*, 2006. 1, 3, 4

[8] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007. 6

[9] H. Jegou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *IEEE Trans. PAMI*, 32(1):2–11, 2010. 5

[10] L. Latecki, R. Lakámper, and U. Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *Proc. of CVPR*, pages 424–429, 2000. 4

[11] H. Ling and D. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):286–299, 2007. 4, 5

[12] M. Maila and J. Shi. Random walk view of segmentation, and learning spectral graph partitioning: Learning segmentation with random walk. In *Proc. NIPS*, 2001. 1, 7

[13] D. R. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms. In *ICCV*, 2001. 7

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford Digital Libraries Working Paper*, 1998. 2

[15] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000. 1, 4

[16] H. S. Seung and D. D. Lee. The manifold ways of perception. *Science*, 290(5500):2268–2269, 2000. 1

[17] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. 3

[18] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000. 1

[19] H. Stewénius and D. Nistér. Object recognition benchmark. In *http://vis.uky.edu/%7Estewe/ukbench/*. 5

[20] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *NIPS*, pages 945–952, 2001. 1

[21] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2322, 2000. 1, 4

[22] Z. Tu and A. L. Yuille. Shape matching and recognition - using generative models and informative features. In *Proc. ECCV*, pages 195–209, 2004. 4, 5

[23] X. Yang, S. Koknar-Tezel, and L. Latecki. Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval. In *Proc. of CVPR*, 2009. 5

[24] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *Proc. NIPS*, 2004. 1, 2

[25] X. Zhu. Semi-supervised learning with graphs. In *Doctoral Dissertation, CMU-LTI-05-192*, 2005. 1, 2