OUT-OF-SAMPLE EXTRAPOLATION USING SEMI-SUPERVISED MANIFOLD LEARNING (OSE-SSL): CONTENT-BASED IMAGE RETRIEVAL FOR PROSTATE HISTOLOGY GRADING

Rachel Sparks, Anant Madabhushi

Department of Biomedical Engineering, Rutgers University
Piscataway, New Jersey, 08854

ABSTRACT

In this paper, we present an out-of-sample extrapolation (OSE) scheme in the context of semi-supervised manifold learning (OSE-SSL). Manifold learning (ML) takes samples with high dimensionality and learns a set of low dimensional embeddings. Embeddings generated by ML preserve nonlinear relationships between samples allowing dataset visualization, classification, or evaluation of object similarity. Semi-supervised ML (SSL), a recent ML extension, exploits known class labels to learn embeddings, which may result in greater separation between samples of different classes compared to unsupervised ML schemes. Most ML schemes utilize the eigenvalue decomposition (EVD) to learn embeddings. For instance, Graph Embedding (GE) learns embeddings by EVD on a similarity matrix that models high dimensional feature vector similarity between samples. In datasets where new samples are acquired, such as a content-based image retrieval (CBIR) system, recalculating EVD is infeasible. OSE schemes obtain new embeddings without recalculating EVD. The Nyström method (NM) is an OSE algorithm where new embeddings are estimated as a weighted sum of known embeddings. Known embeddings must describe the embedding space for NM to accurately estimate new embeddings. In this paper, NM and semi-supervised GE (SSGE) are combined to learn embeddings which cluster samples by class and rapidly calculate embeddings for new samples without recalculating EVD. OSE-SSL is compared to (i) NM paired with GE (NM-GE), and (ii) SSGE obtained for the full database, where SSGE results represent ground truth embeddings. OSE-SSL, NM-GE, and SSGE are evaluated in their ability to: (1) cluster samples by label, measured by Silhouette Index (SI); (2) CBIR accuracy, measured by area under the precision-recall curve (AUPRC). In a synthetic Swiss roll dataset of 2000 samples, OSE-SSL requires training on 50% of the dataset to achieve SI and AUPRC similar to SSGE while NM-GE requires 70% of dataset to achieve SI and AUPRC similar to GE. For a prostate histology dataset of 888 glands, a CBIR system was evaluated on its ability to retrieve images according to Gleason Grade. OSE-SSL had AUPRC of 0.6 while NM-GE had AUPRC of 0.3.

Index Terms— Semi-Supervised Manifold Learning, Out-of-Sample Extrapolation, Graph Embedding, Content-Based Image Retrieval, Nonlinear Dimensionality Reduction, Nyström Method

1. INTRODUCTION

Manifold learning (ML) schemes aim to learn a set of embeddings $\underline{\mathbf{y}}$ defined on a manifold $\mathcal{M}$. $\mathcal{M}$ is typically contained in a lower dimensionality compared to the original dataset [1]. $\underline{\mathbf{y}}$ obtained by ML can be used for dataset visualization, classification, or evaluation of object similarity. Most ML schemes rely on the computationally intensive eigenvalue decomposition (EVD) solution to estimate $\mathcal{M}$ [2, 3, 4]. For instance, Graph Embedding (GE) utilizes EVD on a high dimensional similarity matrix to learn $\underline{\mathbf{y}}$ [2]. Semi-supervised ML (SSL), a more recent class of ML schemes, attempts to exploit class labels of selected objects to obtain greater class separation in the reduced dimensional space [5]. Extending ML to account for new samples (outside the existing $\mathcal{M}$) is computationally infeasible because EVD must be recalculated every time new samples are incorporated into the dataset.

More formally let U samples define the dataset $\mathbf{O} = \{o_1, \ldots, o_U\}$ such that $\mathbf{O} \in \mathbb{R}^{U \times U}$. $\mathbf{O}$ has a dimensionality of $U \times U$. The objective of a ML scheme might be to learn a n-dimensional manifold $\mathcal{M} \in \mathbb{R}^{n \times U}$ such that $\mathbf{y} = \{y_1, \ldots, y_U\}$ represent embeddings of $\mathbf{O}$ on $\mathcal{M}$, where $n << U$. Given a new sample o_{U+1}, ML should learn an embedding y_{U+1} which represents the location of o_{U+1} on $\mathcal{M}$. However, recalculating $\mathcal{M}$ and $\mathbf{y}$ is computationally intensive.

Out-of-sample extrapolation (OSE) estimates a new embedding location, defined by $\tilde{y}_{U+1}$, without having to recompute $\mathcal{M}$ or $\mathbf{y}$. The Nyström method (NM) is an OSE algorithm which estimates $\tilde{y}_{U+1}$ as a weighted linear combination of known $y \in \mathbf{y}$ [6, 7]. NM relies on a large initial training set that accurately models $\mathcal{M}$. Small training sets may be unable to model $\mathcal{M}$ and thus give erroneous $\tilde{y}$.

In this paper, we present a new OSE method OSE-SSL, which combines OSE and SSL. OSE-SSL allows for the construction of embeddings where distances between samples belonging to different classes are large and more importantly where new samples can be rapidly introduced into the existing embedding without the high computational overhead associated with EVD.

OSE-SSL is specifically evaluated in the context of a content-based image retrieval (CBIR) system. CBIR systems in the context of medical image analysis allow physicians to retrieve previously archived images which have similar diagnostic or prognostic attributes as a query image. Physicians can then relate the knowledge from previously seen cases to the current query case. Recently, CBIR schemes [8] have been proposed where the database of images is embedded into a reduced space and a new query image is compared to the most similar images in the embedding space. However, the query sample needs to be extrapolated into the embedding space of the image database in order to evaluate image similarity. Having to recompute the entire embedding (using the database images and

This work was made possible via grants from the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, R01CA14077201, and R03CA143991-01), and The Cancer Institute of New Jersey. We would like to thank Drs. J. E. Tomasezwski and M. D. Feldman from the Hospital of the University of Pennsylvania for providing prostate histology imagery and annotations.

978-1-4244-4128-0/11/$25.00 ©2011 IEEE

Evaluation Measure	Description
Approximation Error	$\frac{1}{\lvert\hat{W}\rvert}\lvert\lvert W(o_i, o_i) - B'\hat{W}^{-1}B\rvert\rvert_F$ where $B = W(o_i, o_j)$ and $\{o_i \in \mathbf{O}_E, o_j \in \mathbf{O}_T\}$
Silhouette Index (SI)	$\eta^{SI} = \sum_{i=1}^{N} \frac{G(i) - C(i)}{\max\left[C(i), G(i)\right]}$ where $C(i) = \sum_{j, l_j = l_i} \lvert\lvert\tilde{y}_i - \tilde{y}_j\rvert\rvert_2$ and $G(i) = \sum_{j, l_j \neq l_i} \lvert\lvert\tilde{y}_i - \tilde{y}_j\rvert\rvert_2$
Area Under the Precision Recall Curve (AUPRC)	Area generated by plotting $p(\alpha)$ versus $r(\alpha)$ where $p(\alpha) = \frac{\Phi(\alpha)}{\alpha}$ and $r(\alpha) = \frac{\Phi(\alpha)}{\Phi(N)}$. $\Phi(\alpha)$ denotes the number of relevant objects in the closest α points.

Table 1. Evaluation measures to compare ML schemes.

each new query image) is infeasible. Therefore, OSE-SSL can be utilized to obtain an embedding location for the query image. Accurately extrapolating the embedding location is essential to accurately retrieve images most similar to the query image.

In this work we compare OSE-SSL against (1) NM paired with GE (NM-GE) and (2) semi-supervised GE (SSGE). NM-GE is used as an unsupervised comparative strategy to evaluate how incorporating label information in OSE-SSL improves CBIR performance. SSGE represents the ideal result of OSE-SSL since extrapolation should ideally estimate $\tilde{y}_{U+1}$ as y_{U+1}. OSE-SSL is evaluated in terms of its ability to: (i) estimate the similarity graph, (ii) cluster samples by class as measured by Silhouette Index (SI), and (iii) CBIR accuracy as measured by area under the precision-recall curve. These measures are evaluated in the context of (a) a synthetic Swiss roll dataset where class labels are artificially introduced to divide the $2D$ manifold of the Swiss roll into two distinct classes and (b) a prostate histology dataset with three grades of cancer present: benign, Gleason grade 3 (less aggressive), and Gleason grade 4 (more aggressive). The Swiss roll is a synthetic dataset showcases the concept of OSE-SSL and the prostate dataset is one specific clinical application of OSE-SSL.

2. METHODS

A dataset of N samples is defined by $\mathbf{O} = \{o_1, \ldots, o_N\}$. Each sample has a corresponding label defined by $\mathbf{L} = \{l_1, \ldots, l_N\}$. Dissimilarity between samples is defined by the function $A(o_i, o_j)$ where $o_i : i \in \{1, \ldots, N\}$ and $o_j : j \in \{1 \ldots, N\}$. Evaluated over $\mathbf{O}$ the matrix A exists in a high dimensional space such that $A \in \mathbb{R}^{N \times N}$.

2.1. Manifold Learning (ML) via Graph Embedding (GE)

GE is a ML scheme which seeks to find a set of embeddings $\mathbf{y} \in \mathbb{R}^{n \times N}$ that exist in a lower dimensional space such that $n << N$ via the EVD,

$$W\mathbf{y}' = \lambda D\mathbf{y}', \tag{1}$$

where $W(a, b) = e^{-A(a,b)/\sigma}$, σ is an empirically determined scaling term, and D is the diagonal matrix $D(a, a) = \sum_b W(a, b)$. The n eigenvectors corresponding to the top n eigenvalues in λ define the n dimensional embeddings $\mathbf{y}$.

2.2. Out-of-Sample Extrapolation (OSE)

A training set is defined as $\mathbf{O}_T \subset \mathbf{O}$ where $\mathbf{O}_T = \{o_1, \ldots, o_M\}$ and M is the number of samples in the training set such that $M < N$. The dissimilarity matrix of $\mathbf{O}_T$ is defined as $\hat{A}(o_i, o_j) : o_i \in \mathbf{O}_T, o_j \in \mathbf{O}_T$. NM finds a set of approximate embeddings $\tilde{\mathbf{y}}$ by:

Step 1: Compute $\hat{W} = e^{-\hat{A}(a,b)}$.

Step 2: Apply Equation 1 to $\hat{W}$ to obtain the training set embeddings, $\hat{\mathbf{y}}$, and eigenvalues, $\hat{\lambda}$.

Step 3: Apply NM [6] to $o_i \notin \mathbf{O}_T$ to obtain the approximate embeddings $\tilde{y}_i$ by,

$$\tilde{y}_{i,k} = \frac{1}{\hat{\lambda}_k} \sum_{j=1}^{M} \hat{y}_{j,k} W(o_j, o_i), \tag{2}$$

where $k \in \{1, \ldots, n\}$ is the kth embedding dimension corresponding to the kth largest eigenvalue $\hat{\lambda}_k$. Intuitively, NM estimates $\tilde{y}_i$ as a weighted sum of the training embeddings $\hat{y}_j : j \in \{1, \ldots, M\}$ where weights are based on the object similarity $W(o_j, o_i)$.

2.3. Semi-Supervised Manifold Learning (SSL)

A set of known labels is defined $\mathbf{L}_R \subset \mathbf{L}$ where $\mathbf{L}_R = \{l_1, \ldots, l_S\}$ and S is the number of known labels such that $S < N$.

Step 1: The similarity matrix W_R is derived from A and $\mathbf{L}_R$ by,

$$W_R(a, b) = \begin{cases} \gamma(1 + \gamma) & \text{if } l_a = l_b, \\ \gamma(1 - \gamma) & \text{if } l_a \neq l_b, \\ \gamma & \text{otherwise,} \end{cases} \tag{3}$$

where $\gamma = e^{-A(a,b)/\sigma}$. If there are no labels present ($S = 0$) the similarity matrix is defined as $W(a, b) = \gamma$, giving the unsupervised GE weighting term.

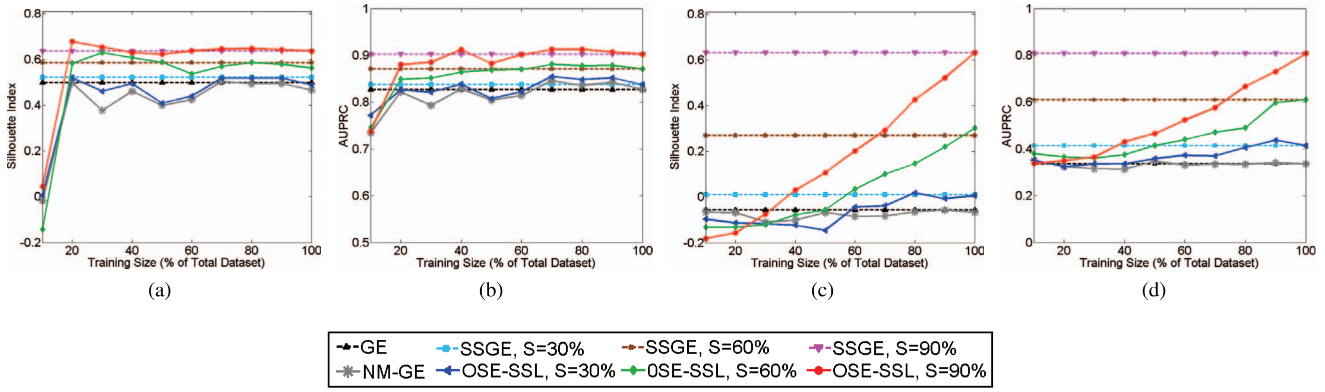
Step 2: Apply Equation 1 to W_R to obtain semi-supervised GE (SSGE) embeddings $\mathbf{y}_R$.

3. OUT-OF SAMPLE EXTRAPOLATION FOR SEMI-SUPERVISED MANIFOLD LEARNING (OSE-SSL)

OSE-SSL assumes a $\mathbf{O}$ with N samples, $\mathbf{O}_T$ with M samples, and $\mathbf{L}_R$ with S labels such that $S < M < N$. From the given information the aim is to find a set of embeddings $\tilde{\mathbf{y}}_R$ which approximates the full embedding $\mathbf{y}_R$. OSE-SSL is trained by:

```
Algorithm TrainOSE-SSL
Input: $A$, $\hat{A}$, $\mathbf{L}_R$
Output: $\hat{\lambda}$ , $\hat{\mathbf{y}}$
begin
    1. Find $W_R$, $\hat{W}_R$ by Equation 3.
    2. Find $\hat{\lambda}_R$ , $\hat{\mathbf{y}}_R$ by Equation 1.
end
```

Once training eigenvalues, $\hat{\lambda}_R$, and embeddings, $\hat{\mathbf{y}}_R$, are known a new embedding $\tilde{y}_{R,i}$ for the newly acquired object $o_i \notin \mathbf{O}_T$ is estimated by Equation 2 leveraging $\hat{\lambda}_R$ and $\hat{\mathbf{y}}_R$.



(a)

(b)

(c)

(d)

Fig. 1. Embeddings for the (a), (b) Swiss roll or (c), (d) prostate histology dataset were generated by OSE-SSL, NM-GE, or SSGE algorithms. Embeddings were evaluated on (a),(c) clustering samples belonging the the same class measured by SI and (b), (d) accurately returning samples from the same class as the query sample using AUPRC. OSE-SSL outperforms NM-GE and asymptotically increases to performance which is similar to SSGE. Increasing S and M results in a higher SI and AUPRC.

4. EXPERIMENTAL DESIGN

OSE-SSL embeddings, $\tilde{\mathbf{y}}_R$, were compared to: (1) NM-GE embeddings $\tilde{\mathbf{y}}$ described in Section 2.2 and (2) SSGE embeddings $\mathbf{y}$ described in Section 2.3. Embeddings were also evaluated for varying size of the training set (M) and known labels (S): (1) $10\% \leq M \leq 100\%$ of the dataset, and (2)$10\% \leq S \leq 100\%$ of M.

4.1. Evaluation Measures

Embeddings were evaluated by the measures described in Table 1. Norms are denoted by: $||\cdot||_F$ the Frobenius norm, $||\cdot||_2$ the L2 norm, and $|\cdot|$ the L1 norm. Approximation error describes the differences between true similarity W and estimated similarity $\tilde{W}$, defined by NM [7]. SI is a measure of how well samples cluster by class label [10] with 1 corresponding to well separated class clusters and -1 corresponding to mixing of classes. AUPRC measures CBIR accuracy with 0 corresponding to inability to retrieve samples accurately and 1 corresponding to perfect retrieval of samples.

4.2. Dataset Description

Swiss Roll: The Swiss roll is characterized by a $2D$ planar manifold structure which exists in a $3D$ coordinate system space [4]. We generated a Swiss roll containing 2000 samples. We assigned labels to points on the roll be dividing the $2D$ manifold structure into two domains, where each domain corresponds to a different label.

The dissimilarity matrix A is defined as,

$$A(o_i,o_j)\begin{cases}||o_i-o_j||_2 & \text{if } ||o_i-o_j||_2<\mathcal{N},\\ 0 & \text{otherwise.}\end{cases} \qquad (4)$$

$\mathcal{N}$ is a neighborhood parameter determined empirically.

Prostate Histopathology: A set of 58 prostate tissue biopsy cores were stained with Hemotoxylin and Eosin (H & E) and digitized using a ScanScope CS$^{\text{TM}}$ whole-slide scanning system at 40× optical magnification. An expert pathologist selected and classified 120 region into three classes: benign (24 regions), Gleason grade 3 (67 regions), or Gleason grade 4 (11 regions). Gleason grades correspond to prostate cancer aggressiveness, with higher grades corresponding to more aggressive forms of cancer. Every gland contained within each region was segmented by a human expert to obtain 888 glands distributed across three classes: benign ($N = 93$), Gleason grade 3 ($N = 748$), and Gleason grade 4 ($N = 47$).

The dissimilarity matrix A is defined by comparison of medial axis shape models (MASMs) using a diffeomorphic based similarity (DBS) measure as described in [9]. DBS has been shown to be able to describe the subtle morphologic differences between gland appearance corresponding to different Gleason grades making this an appropriate measure for a prostate histology CBIR system.

5. RESULTS AND DISCUSSION

5.1. Swiss Roll

Figure 1 shows results for select values of known labels (S) and training sets (M). The X-axis corresponds to varying M. Different lines represent varying S and methods (NM-GE, SSGE, OSE-SSL). Flat lines correspond to SSGE embeddings as $M = 100\%$ to obtain the true embedding for all images.

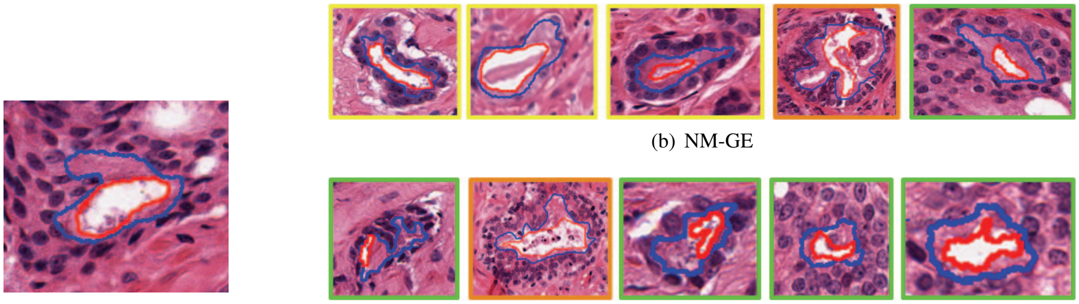
Approximation Error: For all S and M, error was on the order of 10^{-4} demonstrating NM-GE and OSE-SSL accurately approximate the similarity matrix.

SI: Increasing M or S allows for better clustering. For increasing M NM-GE and OSE-SSL increase toward SSGE, which represents the best case scenario where the EVD solution is obtained for the full dataset. For $M = 70\%$ for NM-GE approximates the SSGE embeddings. For OSE-SSL the value of S determine the size of the training set needed to approximate SSGE embeddings. For $S = 90\%$ only $M = 20\%$ is required.

AUPRC: For OSE-SSL at $M > 40\%$ an equivalent AUPRC for OSE-SSL and SSGE is achieved. This demonstrates OSE-SSL is able to approximate the SSGE embeddings. A higher S results in a higher AUPRC and a smaller M necessary to obtain the ideal AUPRC, represented by SSGE.

5.2. Prostate Gland Morphology

Figure 1 shows results for select values of known labels (S) and training sets (M). The X-axis corresponds to varying M and different lines correspond to varying S and methods (NM-GE, SSGE,



(a) Query

(b) NM-GE

(c) OSE-SSL

Fig. 2. The first 5 images retrieved for the (a) query image belonging to Gleason grade 4 by (b) NM-GE and (c) OSE-SSL. Retrieved images belonging to the same class as the query image are outlined in green while those belonging to Gleason grade 3 are in yellow, and benign glands are outlined in orange. OSE-SSL outperforms NM-GE by making use of known labels.

OSE-SSL). For $M > 30\%$ of the dataset, SI > 0 and AUPRC > 0.4. However unlike the synthetic case, for large S only $M > 90\%$ was able to have performance measures similar as those calculated on the full dataset.

Figure 2 gives a query image and the first 5 retrieved images for a CBIR system based on (b) NM-GE or (c) OSE-SSL with $M = 80\%$ and $S = 60\%$. As modeled by the AUPRC rates, OSE-SSL is better able to retrieve images of the same class as the query image (Gleason grade 4) than NM-GE.

Approximation Error: For all S and M, error was on the order of 10^{-4} demonstrating NM-GE and OSE-SSL accurately approximate the similarity matrix.

SI: Increasing M or S increases SI. For $S = 50\%$, SI $= 0.5$, showing a tendency to cluster glands according to class. Increasing S made a larger M necessary for OSE-SSL to approximate SSGE.

AUPRC: Increasing M or S increases AUPRC. For OSE-SSL a small asymptote can be seen around $M = 90\%$ for $S = 30\%$ and $S = 60\%$. This indicates for this dataset $M = 100\%$ is necessary to learn the underlying embedding space. For larger S OSE-SSL is unable to approximate the characteristics of SSGE indicating a need to expand our M to accurately model the embedding space.

6. CONCLUDING REMARKS

In this paper, we presented out-of-sample extrapolation for semi-supervised manifold learning (OSE-SSL) in the context of CBIR for a prostate histology dataset. OSE-SSL is able to accurately extrapolate embedding locations for samples not present in the original dataset. OSE-SSL represents a powerful way to extend ML to incorporate samples not present in the original dataset with a computationally efficient scheme. Furthermore our results demonstrate the SSL component of our algorithm affects the ability of OSE to extrapolate new sample embeddings.

OSE-SSL allows extrapolation of new embedding locations enabling our CBIR system to rapidly retrieve images most similar to a newly acquired query image. By combining SSL with OSE to learn the embedding space our CBIR system is better able to distinguish prostate histology images belonging to the same Gleason grade as a query image, as is evident by higher AUPRC compared to the unsupervised NM-GE scheme.

7. REFERENCES

[1] G. Lee, C. Rodriguez, and A. Madabhushi, "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE TCBB*, vol. 5, no. 3, pp. 368–384, 2008.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[3] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[4] J Tenenbaum, V. de Silvia, and J. Langford, "A global framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[5] H. Zhao, "Combining labeled and unlabeled data with graph embedding," *Neurocomputing*, vol. 69, no. 16-18, pp. 2385–2389, 2006.

[6] 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," in *NIPS*, 2003, pp. 177–184.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE TPAMI*, vol. 28, no. 2, pp. 214–225, 2004.

[8] J. He, M. Li, H.J. Zhang, H. Tong, and C. Zhang, "Generalize manifold-ranking-based image retrieval," *IEEE TMI*, vol. 15, pp. 3170–3177, 2006.

[9] R. Sparks and A. Madabhushi, "Novel morphometric based classification via diffeomorphic based shape representation using manifold learning," in *MICCAI*, 2010, vol. 13, pp. 658–665.

[10] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational Applied Mathematics*, vol. 20, pp. 53–65, 1987.