## ABSTRACT

Many modern applications require modeling and analysis of functions on large, high dimensional, unstructured data sets. One may assume that the data lies on a low dimensional manifold, but this manifold is not known. We have extended the diffusion geometry paradigm for these problems to study function approximation on data defined manifolds. Our algorithms are applied successfully to recognition of hand written digits, classification and missing data problems, automatic diagnosis of age related macular disease based on multi-spectral images, and prediction of blood glucose levels. The ideas are applied to other problems, such as analysis of terrain data and solutions of differential equations.

## SUBJECT TERMS

Data defined manifolds, kernel based approximation, multiresolution analysis, minimum norm interpolation, quadrature formulas

---

**1. REPORT DATE (DD-MM-YYYY)** 10-06-2015  
**2. REPORT TYPE** Final Report  
**3. DATES COVERED (From - To)** 25-Aug-2009 - 28-Aug-2013

**4. TITLE AND SUBTITLE**  
Final Report: Research Area 3: Mathematical Sciences: 3.4, Discrete Mathematics and Computer Science

**5a. CONTRACT NUMBER** W911NF-09-1-0465  
**5b. GRANT NUMBER**  
**5c. PROGRAM ELEMENT NUMBER** 611102  
**5d. PROJECT NUMBER**  
**5e. TASK NUMBER**  
**5f. WORK UNIT NUMBER**

**6. AUTHORS**  
Hrushikesh N. Mhaskar

**7. PERFORMING ORGANIZATION NAMES AND ADDRESSES**  
California State University - Los Angeles  
5151 State University Drive, GE 314  
Los Angeles, CA 90032-4226

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)**  
U.S. Army Research Office  
P.O. Box 12211  
Research Triangle Park, NC 27709-2211

**10. SPONSOR/MONITOR’S ACRONYM(S)** ARO  
**11. SPONSOR/MONITOR’S REPORT NUMBER(S)** 53313-CS-H.20

**12. DISTRIBUTION AVAILABILITY STATEMENT**  
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**  
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrived as an official Department of the Army position, policy or decision, unless so designated by other documentation.

**14. ABSTRACT**

Many modern applications require modeling and analysis of functions on large, high dimensional, unstructured data sets. One may assume that the data lies on a low dimensional manifold, but this manifold is not known. We have extended the diffusion geometry paradigm for these problems to study function approximation on data defined manifolds. Our algorithms are applied successfully to recognition of hand written digits, classification and missing data problems, automatic diagnosis of age related macular disease based on multi-spectral images, and prediction of blood glucose levels. The ideas are applied to other problems, such as analysis of terrain data and solutions of differential equations.

**15. SUBJECT TERMS**  
Data defined manifolds, kernel based approximation, multiresolution analysis, minimum norm interpolation, quadrature formulas

---

**16. SECURITY CLASSIFICATION OF:**  
a. REPORT UU  
b. ABSTRACT UU  
c. THIS PAGE UU

**17. LIMITATION OF ABSTRACT**

**15. NUMBER OF PAGES**

**19a. NAME OF RESPONSIBLE PERSON** Hrushikesh Mhaskar  
**19b. TELEPHONE NUMBER** 626-914-7022
ABSTRACT

Many modern applications require modeling and analysis of functions on large, high dimensional, unstructured data sets. One may assume that the data lies on a low dimensional manifold, but this manifold is not known. We have extended the diffusion geometry paradigm for these problems to study function approximation on data defined manifolds. Our algorithms are applied successfully to recognition of handwritten digits, classification and missing data problems, automatic diagnosis of age related macular disease based on multi-spectral images, and prediction of blood glucose levels. The ideas are applied to other problems, such as analysis of terrain data and solutions of partial differential equations. The scientific barriers include the development of kernel based methods so as to avoid computation of eigenvalues and eigenvectors of large matrices, and quadrature formulas which are guaranteed to work better than the straightforward Monte Carlo integration method.
Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>06/07/2015 12.00</td>
<td>Hrushikesh Mhaskar, Fran Narcowich, Juergen Prestin, Joseph Ward. Lp BERNSTEIN ESTIMATES AND APPROXIMATION BY SPHERICAL BASIS FUNCTIONS, Mathematics of Computation, (07 2010): 1647. doi:</td>
</tr>
<tr>
<td>06/07/2015 19.00</td>
<td>Charles Chui, Hrushikesh Mhaskar. Smooth function extension based on high dimensional unstructured data, Mathematics of Computation, (11 2014): 2865. doi:</td>
</tr>
<tr>
<td>06/07/2015 14.00</td>
<td>Charles Chui, Hrushikesh Mhaskar. MRA contextual-recovery extension of smooth functions on manifolds, Applied and Computational Harmonic Analysis, (01 2010): 104. doi:</td>
</tr>
<tr>
<td>09/10/2012 6.00</td>
<td>F. Filbir, H. N. Mhaskar, J. Prestin. On the problem of parameter estimation in exponential sums, Constructive Approximation, (06 2012): 323. doi:</td>
</tr>
<tr>
<td>09/10/2012 10.00</td>
<td>Martin Ehler, Frank Filbir, Hrushikesh Mhaskar. Locally learning biomedical data using diffusion frames, Journal of Computational Biology, (07 2012): 0. doi:</td>
</tr>
</tbody>
</table>

TOTAL: 14

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received  Paper

TOTAL:
Number of Papers published in non peer-reviewed journals:

(c) Presentations

Conference on Complexity theory, Dagstuhl, Germany, September 2009
Conference on Optimal configurations of points on the sphere and other manifolds, Nashville, TN, May 2010.
Conference on Mathematics of Search Engines, Lake Arrowhead, June, 2010
Joszef Marcinkiewicz Centenary Conference, Poznan, Poland, June, 2010
Colloquium talk: University of Warsaw, Warsaw, Poland, July, 2010.
International conference on approximation theory, Nashville, May, 2011
International conference on approximation theory, Ubeda, Spain, June, 2011 (Plenary lecture)
Katholische Universität, Eichstätt, Germany, July, 2011
International congress on industrial and applied mathematics, Vancouver, Canada, July, 2011
International conference on approximation theory and harmonic analysis, Barcelona, Spain, Dec. 2012 (Plenary speaker)
Annual meeting western states mathematical physics meeting, February, 2012
Invited lecture, University of Texas--Pan American, March 2012
Invited lecture, Ohio State University, April, 2012
Oberwolfach meeting on machine learning and approximation theory, June, 2012.
Oberwolfach meeting on spherical designs, August, 2012.
International conference on applied inverse problems, Daejeon, S. Korea, July, 2013
Invited lecture, HelmholtzZentrum, Munich, July, 2013
Invited lecture, RICAM, Linz, Austria, July 2013

Number of Presentations:  20.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
</table>

TOTAL:
Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>06/07/2015 18.00</td>
<td>S. Chandrasekaran, K. R. Jayaraman, H. N. Mhaskar, S. Pauli. Minimum Sobolev Norm schemes and applications in image processing, IS&amp;T/SPIE Electronic Imaging. 16-JAN-10,</td>
</tr>
<tr>
<td>08/11/2011 3.00</td>
<td>S. Chandrasekaran, K. R. Jayaraman, M. Gu, H. N. Mhaskar, J. Moffitt. Higher order numerical discretization methods with Sobolev norm minimization, 11th International Conference on Computational Sciences. 01-JUN-11,</td>
</tr>
<tr>
<td>09/10/2012 8.00</td>
<td>Frank Filbir, Hrushikesh Mhaskar, Martin Ehler. Learning Biomedical Data Locally using Diffusion Geometry Techniques, Imaging and Signal Processing in Health Care and Technology / 772: Human--Computer Interaction / 773: Communication, Internet and Information Technology. 14-MAY-12,</td>
</tr>
</tbody>
</table>

TOTAL: 3

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
</table>

TOTAL:

Number of Manuscripts:

Books

<table>
<thead>
<tr>
<th>Received</th>
<th>Book</th>
</tr>
</thead>
</table>

TOTAL:
Patents Submitted

Patents Awarded

Awards
John von Neumann distinguished professor, Technical University, Munich, Summer, 2011.

TOTAL:

Graduate Students

<table>
<thead>
<tr>
<th>NAME</th>
<th>PERCENT_SUPPORTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTE Equivalent:</td>
<td></td>
</tr>
<tr>
<td>Total Number:</td>
<td></td>
</tr>
</tbody>
</table>

Names of Post Doctorates

<table>
<thead>
<tr>
<th>NAME</th>
<th>PERCENT_SUPPORTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTE Equivalent:</td>
<td></td>
</tr>
<tr>
<td>Total Number:</td>
<td></td>
</tr>
</tbody>
</table>

Names of Faculty Supported

<table>
<thead>
<tr>
<th>NAME</th>
<th>PERCENT_SUPPORTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTE Equivalent:</td>
<td></td>
</tr>
<tr>
<td>Total Number:</td>
<td></td>
</tr>
</tbody>
</table>

Names of Under Graduate students supported

<table>
<thead>
<tr>
<th>NAME</th>
<th>PERCENT_SUPPORTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTE Equivalent:</td>
<td></td>
</tr>
<tr>
<td>Total Number:</td>
<td></td>
</tr>
</tbody>
</table>
**Student Metrics**

This section only applies to graduating undergraduates supported by this agreement in this reporting period.

The number of undergraduates funded by this agreement who graduated during this period: ..... 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields: ..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields: ..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale): ..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering: ..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense: ..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: ..... 0.00

---

### Names of Personnel receiving masters degrees

<table>
<thead>
<tr>
<th>NAME</th>
<th>Total Number:</th>
</tr>
</thead>
</table>

### Names of personnel receiving PHDs

<table>
<thead>
<tr>
<th>NAME</th>
<th>Total Number:</th>
</tr>
</thead>
</table>

### Names of other research staff

<table>
<thead>
<tr>
<th>NAME</th>
<th>PERCENT_SUPPORTED</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>FTE Equivalent:</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Total Number:</th>
</tr>
</thead>
</table>

---

### Sub Contractors (DD882)

| Inventions (DD882) |

| Scientific Progress |

| Technology Transfer |

See attachment
1 Forward

Many modern applications require modeling and analysis of functions on large, high dimensional, unstructured data sets. One may assume that the data lies on a low dimensional manifold, but this manifold is not known. We have extended the diffusion geometry paradigm for these problems to study function approximation on data defined manifolds. Our algorithms are applied successfully to recognition of hand written digits, classification and missing data problems, automatic diagnosis of age related macular disease based on multi–spectral images, and prediction of blood glucose levels. The ideas are applied to other problems, such as analysis of terrain data and solutions of partial differential equations. The scientific barriers include the development of kernel based methods so as to avoid computation of eigenvalues and eigenvectors of large matrices, and quadrature formulas which are guaranteed to work better than the straightforward Monte Carlo integration method.

2 Statement of Problems Studied

The grant was a continuation of our research on function approximation on the Euclidean sphere, supported by the ARO. During this project, we studied a variety of extensions of this theory to the context of data defined manifold, bringing the theory on such manifolds to the same level of completion as that on the sphere.

The main underlying problem is the following. One starts with a data structure called point cloud, which is a finite subset \( \mathcal{P} = \{x_i\}_{i=1}^N \) of some high dimensional ambient Euclidean space, together with an affinity relation \( W \), where one interpretes \( W_{i,j} = W(x_i, x_j) \), indicating how “close” \( x_i \) is to \( x_j \). The matrix \( W \) clearly defines an undirected graph, which can be embedded into a low dimensional manifold using the diffusion geometry paradigm: i.e., one considers the limit of the graph Laplacian as the Laplace–Beltrami operator on a manifold, which has eigenvalues \( -\lambda_k^2 \) and the corresponding eigenfunctions \( \phi_k \). While most of the existing theory focuses on understanding the data geometry and data visualization, applications to semi–supervised learning can be cast as problems of function approximation. Thus, in classification problems, one knows the class labels on a small training data \( \mathcal{C} \subset \mathcal{P} \), which may be viewed as a function \( f : \mathcal{C} \rightarrow \mathbb{R} \). Then the problem of semi–supervised learning is to extend this function to \( \mathcal{P} \); i.e., to learn the class labels of every point in \( \mathcal{P} \). The main questions of interest to us were the following:

1. Study the connection between the smoothness properties of the target function \( f \) as defined by the geometry of the unknown manifold, and the intrinsic approximation error that can be expected in approximating \( f \) by a linear combination of finitely many eigenfunctions \( \phi_k \).

2. Develop algorithms to compute efficiently a linear approximation process that yields a near best approximation.

The ideas developed during this work found applications in some other areas as well, such as solutions of partial differential equations, image processing, and prediction of blood glucose levels in diabetes patients.

2.1 Scientific barriers

- The target function is defined on a manifold, or even a graph, with no known structure. The only information available is the unstructured data.
- The eigenfunctions of the heat kernel do not have any such special function properties as the Funk–Hecke formula, addition formula, etc. familiar in the classical theory.
- Since the heat kernel is the only object that can be constructed approximately, the conditions must be formulated in terms of this kernel.
- The existence of quadrature formulas is not clear, since the underlying manifold and its Riemannian measure are both unknown. The classical tools like the Bernstein inequality are not available in this setting, and must be developed new.
- The data may be nominally high dimensional, posing formidable numerical problems.
2.2 Our approach

- Consider the data to be a sample from an unknown manifold.

- Approximate the heat kernel on this manifold using a graph Laplacian, constructed from the data. The eigenfunctions of this kernel form the class of approximants.

- Develop a filter which yields a highly localized mollifier, and which is efficient to apply. This yields local approximation given Fourier information.

- Develop quadrature formulas exact for high complexity approximants, based on scattered data, and use these to convert the filter approximation to a kernel based approximation using the available data.

3 Summary of results

The research resulted in 17 publications and a number of colloquium and conference presentations.

The paper [19] is an invited survey paper, where the basic ideas behind the research are illustrated in the context of multivariate trigonometric polynomials.

The papers [9, 10] deal with the question of developing quadrature formulas that enable us to discretize various integral operators in the theory of diffusion geometry while keeping track of the errors, so that the net effect on the accuracy of function approximation is not affected significantly. An important, but special case, of this theory was developed in the case of quadrature formulas for spherical triangles in [1].

A further extension of this work, so that the function approximation is achieved so as to preserve the known values of the target function at certain landmarks on the manifold, is studied in [6].

The papers [15, 16] deal with the question of function approximation without computing the eigenvalues and eigenfunctions explicitly, even though the smoothness of the target function is related intimately with the spaces spanned by these eigenfunctions. In particular, the paper [15] generalizes the results in [17] for radial basis function approximation on the sphere.

The paper [20] is an offshoot of the ideas in this research in the context of expansions of functions in terms of Jacobi polynomials.

In the remaining publications, we focused on applications to various areas. The paper [5] deals with image completion problems. We study the question of contextual recovery of missing data while preserving a certain number of normal derivatives at the boundary. The classical approach to this problem is to solve a differential equation. We have demonstrated both the applicability and limitations of this approach in its full generality.

The paper [3] deals with the question of function extension from a set of points on the torus to the whole torus. Unlike the rest of our research, the points are not dense on the torus, but the extension is to minimize a Sobolev norm. In particular, we are able to overcome Runge’s example.


In [4], we use the ideas in [3] for image segmentation problems. Figure 1 illustrates some of the results.

![Image](image-url)

**Figure 1:** The figures from left to right: The original cameraman image, the segmentation of the cameraman image, the original mandrill image, the segmentation of the mandrill image.
In [11], we consider the problem of signal separation in stationary signals; i.e., given the samples

\[ x(k) = \sum_{j=1}^{K} a_j \exp(-i\omega_j k) + \text{noise}, \quad k = -N, \cdots, N, \]

for sufficiently large value of \( N \), we wish to find \( \omega_j, j = 1, \cdots, K \). Once we find \( \omega_j \) accurately, it is simple solution of a linear system of equations to find \( a_j \). In [11], we used the theory of orthogonal polynomials on the unit circle together with our earlier theory of trigonometric polynomial frames to solve this problem very accurately. We gave theoretical bounds on the effect of noise, without assuming any special distribution of the noise.

For example, we consider

\[ x(k) = 34 + 300 \cos(k\pi/4) + \cos(k\pi/2) + \epsilon(k), \quad k = -1024, \cdots, 1024, \]

where \( \epsilon(k) \) is a random variable uniformly distributed in the range \([-3, 3]\). Thus, in addition to the large differences in the magnitudes at different frequencies, also the noise is three times the strength of the weakest signal at \( \pi/2 \). As an average over 500 trials, the frequencies were recovered as

\[ (-3.992361996552509e - 16, \pm 0.785398165178676, \pm 1.570153903610014). \]

In particular, the weakest frequency \( \pi/2 \) was detected with an accuracy of \( 6.4242 \epsilon - 4 \) in spite of the noise being 3 times the strength of the corresponding signal.

In [7, 8], we apply our theory for biomedical applications. We analyzed the Cleveland heart disease data set to determine the stage of the heart disease of a patient, based on 13 attributes, and also, a variant of the Wisconsin breast cancer data set where, instead of the classification problem, we omitted one of the independent variables, and treated the problem as that of missing data recovery. In each of these examples, we outperformed the art Support Vector Machine algorithms. A novelty of these papers is the classification of drusen in retina of patients with age related macular disease. The classification was based on a first of its kind data set obtained by the National Institute of Health. Each image in this data set was a multi-spectral image with 24 different frequencies. Our methods gave an automatic prognosis, as illustrated in Figure 2.

![Figure 2: The left two images are at different frequencies for the retina of a patient with advanced AMD, the right-most image shows our classification of drusen based on 24 such images at different frequencies.](image)

In [18], we applied the ideas in our research for the prediction of blood glucose level and rate of change of this level in a 15 minute prediction horizon, based on continuous glucose monitoring (CGM) device readings during the past half hour. To quantify the clinical accuracy of the considered predictors, we use the Prediction Error-Grid Analysis (PRED-EGA) [21], which has been designed especially for the blood glucose predictors. This assessment methodology records reference glucose estimates paired with the estimates predicted for the same moments. As a result, the PRED-EGA reports the numbers (in percent) of Accurate (Acc.), Benign (Benign) and Erroneous (Error) predictions in hypoglycemic (0-70 mg/dL), euglycemic (70-180 mg/dL) and hyperglycemic (180-450 mg/dL) ranges. This stratification is of great importance because consequences caused by a prediction error in the hypoglycemic range are very different from ones in the euglycemic range. In [18] the assessment is done with respect to the references given as simulated noise-free BG-readings. On a data set of 10 virtual patients obtained from Padova/University of Virginia simulator [13], we obtained the following representative results shown in Table 1, where we compare our results with those obtained by using the state of the art Modified Savitzky-Golay Filtering (MSGF) [12, 14]. Remarkably, in the hyperglycemic region, this research has achieved 100% accuracy, while the previously known academic record is 91.43%. Also, in the euglycemic region, the percentage of benign and dangerous errors is reduced 50%.
Table 1: Percentage of the accurate, benign, and erroneous predictions in hypoglycemic (columns 2-4), euglycemic (columns 5-7), and hyperglycemic (columns 8-9) regions respectively.

<table>
<thead>
<tr>
<th></th>
<th>Acc.</th>
<th>Benign</th>
<th>Error</th>
<th>Acc.</th>
<th>Benign</th>
<th>Error</th>
<th>Acc.</th>
<th>Benign</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSGF</td>
<td>99.78</td>
<td>0.22</td>
<td>–</td>
<td>98.98</td>
<td>0.93</td>
<td>0.091</td>
<td>91.43</td>
<td>8.57</td>
</tr>
<tr>
<td>Our method</td>
<td>99.74</td>
<td>0.26</td>
<td>–</td>
<td>99.40</td>
<td>0.55</td>
<td>0.05</td>
<td>100</td>
<td>–</td>
</tr>
</tbody>
</table>

References


