ACTIVITY RECOGNITION VIA CLASSIFICATION CONSTRAINED DIFFUSION MAPS

By

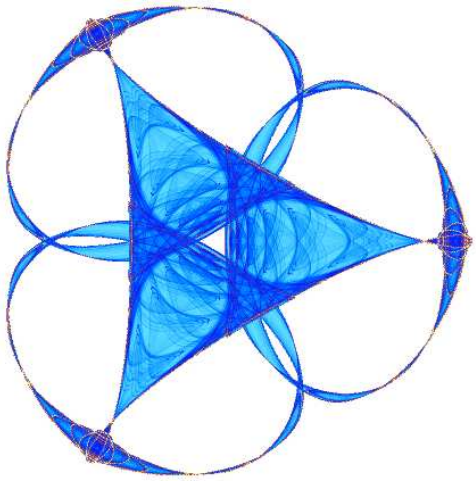
Y. Ma

and

S.B. Damelin

IMA Preprint Series # 2118

(May 2006)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall

207 Church Street S.E.

Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370

URL: http://www.ima.umn.edu

Activity Recognition via Classification Constrained Diffusion Maps

Y. Ma and S.B. Damelin

No Institute Given

Abstract. In this paper, we study the problem of classification of spatial temporal feature actions from video. Our new idea is to use class labels to find optimal hitting times $t \geq 1$ for dimension reduction using diffusion via random walks. We use our methods on real data, and compare our new method with the same method (with $t = 1$) for random walk diffusion and dual root minimal spanning tree diffusion. The results by our former method are shown to be considerably better as the choice of optimal hitting times compensate for sensitive weight width parameters in affinity matrices and choose optimal weighted eigenspaces.

1 Introduction and Background

Recognition of human actions from video streams has recently become an active area of research with numerous applications in video surveillance, which is mostly motivated by the increasing number of video cameras deployed for video surveillance and the current inability of video operators to monitor and analyze large volumes of data. For predefined activities, many rule-based or logic based methods have been proposed. For example, in [16], the authors define a series of rules, e.g. entry violation, escort, theft whereas the results of [17] use a declarative model and a logic based approach to recognize predefined activities. Unfortunately a major drawback of pre-defined activity recognition approaches is that the rules developed for one activity typically may not be applicable for other activities. Indeed, different application domains may be interested in different activities. One of the key challenges in these later systems is the ability to model the activities of interest, as well as develop a methodology that allows automatic recognition of activities. In [10, 2, 7], the authors show that same or similar activity video sequences are clustered close to each other and far from different activity video sequences.

We discuss an automatic activity recognition system which initially has an activity gallery that may be empty or may contain a number of initial simple activities. The system is trained by example, where input video sequences are manually labeled and the system extracts features and automatically learns the new activity.

We suppose that we are given a set of labeled $n \geq 1$ video sequences as training data where each sequence is a high dimensional spatial temporal feature. Also we are given a set of $p \geq 1$ class labels. denoting a fixed number of

activities. Thus each video sequence is represented by a high dimension feature considered isometric to a point in a high dimensional vector space. One may think that high dimension here should be an obstacle for any efficient processing of our data. Indeed, many classical data processing algorithms have a computational complexity that grows exponentially with dimension (the so called "curse of dimension"). All is not lost however, since the information in each feature is in practice of lower dimension. Dimension reduction is a way to find an isometric mapping of each video sequence into a corresponding point in Euclidean space of lower dimension where its description is considered simpler. Spatial temporal features are first associated with the nodes of a graph with a natural metric and then as a classifier, the reduced features are matched using a k-nearest neighbor classifier in the reduced space for some $k \geq 1$. We use class labels to find the optimal hitting time $t \geq 1$ for dimension reduction using diffusion via random walks. More precisely, for each fixed positive integer t, we define a map which takes input sequences to powers of t of reduced points. Then for each t, we compute a cross validation on the reduced points using the given labels and then select the optimal t value which yields the smallest cross validation value. For this optimal t, we perform diffusion dimension reduction on the high dimensional spatial temporal feature space and then use a k nearest neigbor classifier on the reduced space for some $k \geq 1$. For the diffusion process, the different heating times t, produce different reduced dimension features. We use our methods on real data, and compare our new method with the same method (with $t = 1$) random walk diffusion and dual root minimal spanning tree diffusion. The results by our former method are shown to be considerably better as the choice of optimal hitting times compensate for sensitive weight width parameters in affinity matrices and choose optimal weighted eigenspaces.

The results of [18, 14] use dimensionality reduction by eigenmap methods. For example, in [18] the authors calculate the co-occurrence matrix between features, and solve for the smallest eigenvectors to find an embedding space for dimensionality reduction. The results of [14] use eigenvector decomposition of feature similarity matrices for abnormal vehicle trajectories detection.

The remainder of this paper is organized as follows. In Section 2, we first describe spatial temporal features and then present recent random walk and dual root minimal spanning tree diffusion map methods of Lavon et al and Grischat et al (see [3, 6] for dimensionality reduction on which our methods are based. Section 3 describes our new classification constrained diffusion map method as well as an existing classification method which we use using random walk and dual root minimal spanning tree diffusion maps. In Section 4, we present experimental results and finally in Section 5, we present an overview summary.

2 Spatial temporal feature and existing diffusion map methods

In this section, we describe how given any action, we are able to extract features from an image using IIR filters. The idea, following [5], is to represent motion by

its recency. That is, recent motion is represented as brighter than older motion. This technique is called recursive filtering.

The extracted features will represent the properties of the interested objects. In the experiments of this paper, we use image-based motion features that are directly computed from the image sequence as explained below. We use a similar approach described in [11] which, unlike [5], an action is represented by several feature images rather than just one image. Actions may be complex and repetitive making it difficult to capture motion details in one feature image. In this method, a weighted average at time $i \geq 1$, M_i is computed as

$$M_i = \alpha I_{i-1} + (1 - \alpha) M_{i-1}$$

where I_i is the image at time i and $0 \leq \alpha \leq 1$ is a fixed scalar. The feature image at time i, which we denote by F_i is computed as $F_i = |I_i - M_i|$. Figure 3 below shows an example image and feature image for a running person with $\alpha = 0.3$. Note that it is the contrast of the gray level of the moving object which determines the magnitude of the feature image not the actual gray level value.

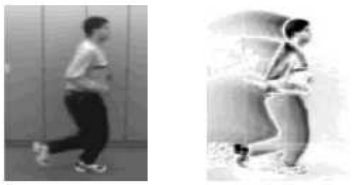


Fig. 1. Feature

Other features can also be used. For example Blob features typically includes centroid based and skeleton based features. The centroid based features (e.g. for far-field object) include not only instantaneous information of the spatial features of objects, such as width, height, and aspect ratio, but also temporal information about changes in the objects sizes as well as motion features, such as direction of movement and speed.

Fig. 2. An original size and a reduced size feature image from a an action of marching soldiers

If we use 12 frames to represent a video sequence, each spatial temporal feature is isometric to a point in $\mathbb{R}^{25\times 31\times 12}$, where spatial resolution can be reduced to 25x31 pixels [11], as shown in Figure 3.

2.1 Dimension Reduction

Here and in what follows, we denote by X, the space of all spatial temporal features viewed as points in $\mathbb{R}^d$ for some large but fixed positive integer d. Let $n \geq 1$ denote the cardinality of X .First introduced in the context of manifold learning, eigenmap techniques, (see [1, 3, 6, 19, 4] and the references cited therein) are methods to isometrically embedd points of X into a lower dimensional Euclidean space. Currently, these techniques fall into two categories. The first are classical linear global dimension reduction techniques such as principle component analysis which assume the elements of X lie on a manifold. On the other hand, spectral methods,take into account local distortion of data points in X. In this paper, we make use of recent random walk and dual root minimal spanning tree diffusion map methods of Lavon et al and Grischat et al (see [3, 6]. In what follows, we now describe these later methods. We construct a graph on X where each point is considered a node and every two nodes are connected by an edge via a non negative, symmetric, positive definite kernel $w : X \times X \to \mathbb{R}$. In this paper, we consider the heat kernel given by

$$w_\sigma(\mathbf{x}_i, \mathbf{x}_j) := \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right),\ \mathbf{x}_i, \mathbf{x}_j \in X,\ i, j = 1, .., n$$

where σ is a real kernel width parameter fixed in advance. The weight w reflects the degree of similarity or interaction between the points $\mathbf{x}_i, \mathbf{x}_j \in X$ and depends only on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in X. Here, $||.||$ is the Euclidean norm in $\mathbb{R}^d$. We remark that everything we propose from this point, goes through trivially for any non negative, symmetric, positive definite kernel w.

2.2 Diffusion via Random Walks

A Markov chain is defined on X as follows. Given a node $\mathbf{x}_i \in X$, we define the degree of $\mathbf{x}_i$ by $d(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in X} w_\sigma(\mathbf{x}_i, \mathbf{x}_j)$. We then form a $n \times n$ affinity matrix P with entries $p(\mathbf{x}_i, \mathbf{x}_j) = \frac{w_\sigma(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)}$, $i, j = 1, ..., n$. Because $\sum_{\mathbf{x}_j \in X} p(\mathbf{x}_i, \mathbf{x}_j) = 1$, P is a transition matrix of a Markov chain on the graph of the members of X. Taking powers of P in steps $t \geq 1$, produces probabilty functions $p_t(\mathbf{x}_i, \mathbf{x}_j)$ which measure the probability of transition from $\mathbf{x}_i$ to $\mathbf{x}_j$ in t steps. Since w_σ is symmetric, P has a sequence of n eigenvalues

$$1 \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$$

and a collection of $1 \leq d_r \leq n$ right eigenvectors $\{\phi_{d_r}\}$ so that for each fixed $t \geq 1$,

$$P^t \phi_{d_r} = \lambda_{d_r}^t \phi_{d_r}.$$

Each eigenvector is a signal over the data points and the eigenvectors form a new set of coordinates on X. For any choice of t, the mapping

$$\Psi_t : \mathbf{x}_i \to \left(\lambda_1^t \phi_1(\mathbf{x}_i), ..., \lambda_{d_r}^t \phi_{d_r}(\mathbf{x}_i)\right)^T$$

is an isometric embedding of X into $\mathbb{R}^{d_r}$ and the function

$$\alpha(\mathbf{x}_i, \mathbf{x}_j) := ||\Psi(\mathbf{x}_i) - \Psi(\mathbf{x}_j)||, \; i, j = 1, ..., n$$

defines a metric on the graph given by the nodes of X. Here $||.||$ denotes the Euclidean norm in $\mathbb{R}^{d_r}$.

σ sensitivity The parameter σ gives the rate at which the similarity between two points decays. The choice of this parameter is something for which there is no good theory. Several heuristics have been proposed and they boil down to trading off sparseness of the kernel matrix (small sigma) with adequate characterization of true affinity of two points. The reason that spectral clustering methods work in general, see for example [13] is that with sparse kernel matrices, long range affinities are accomodated through the chaining of many local interactions as opposed to standard Euclidean distance methods - e.g. correlation - that impute global influence into each pair wise affinity metric, making long range interactions wash out local interactions. Ng and Jordan use a cross-validation approach to selecting sigma, this involves running their algorithm for a large set of values for sigma and choosing the one that results in the 'tightest' clusters, where tightness is determined by some cluster metric such as a closeness of the reindexed distance matrix to a block diagonal matrix. One of the interesting features of our proposed method, (see Section 3) is that it chooses an optimal t, which compensates experimentally for sensitive σ,

Diffusion via dual root minimal spanning trees. The method we describe relies on a simple but elegant idea. Starting with two random walks on different points $\mathbf{x}_i$ and $\mathbf{x}_i$ in X, when will two paths generated hit each other? More precisely, given $\mathbf{x}_i \in X$, we compute a greedy minimal spanning tree and define the distance d between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ as the number of greedy iterations required so that two greedy minimal spanning trees rooted on each point $\mathbf{x}_i$ and $\mathbf{x}_j$ in X will intersect. We set σ to be $1/C\max(\max(\text{Ahop}))$ where $C > 1$ and Ahop is the matrix of all pairwise distances - the hitting times between diffusions from different pairs of points. (In [6] C is taken as 10 but the method there easily works for any fixed $C > 1$). This is an adaptive normalization in the sense that it makes the kernel decay on the order of $1/C$ of the maximum of the hitting times. An affinity matrix P is calculated with the weight w_σ with distance given by the hitting time between points x and y and the eigenvectors of P are used for a dimension reduction map.

3 Proposed method

We are given a set of labeled video sequences of 12 frames $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ as training data and an unlabeled set of testing data $\mathbf{x}_i$, $i = 1, \ldots, m$ where each $\mathbf{x}_i \in \mathbb{R}$, $i = 1, ..., n+m$ represents a d dimensional spatial temporal feature which can be extracted from the ith video sequence. Also $y_i \in \{1, ..., p\}$ is a class label associated to $p \geq 1$ activities. In total, $s \geq 1$ people perform p different actions. The data for p people consititutes traing data while the remaining data for $s - p$ people is for testing. Thus the training data consists of $n = p^2$ video sequences $\mathbf{x}_i \in \mathbb{R}^d$ while the testing data consists of $m = (s-p)p$ video sequences $\mathbf{x}_i \in \mathbb{R}^d$.

Spatial temporal features $\mathbf{x}_i$, $i = 1, .., n$ are associated with the nodes of a graph with a natural metric α given in Subsection 2.2. As a classifier, the reduced features $\mathbf{z}_i$ are then matched using a k-nearest neighbor classifier in the reduced space with $k = 3$. The main idea in this paper is to use class labels to find the optimal hitting time for dimension reduction using diffusion via random walks. More precisely, for each fixed positive integer t, we define a map

$$t : \mathbf{x}_i \to \mathbf{z}_i^{(t)}, \, i = 1, ..., n$$

Here, each $\mathbf{z}_i^{(t)} \in \mathbb{R}^d$ for some $1 \leq d \leq 900$. Then for each t, we compute a leave one out cross validation prediction error estimate (see [8]) on the points $\mathbf{z}_i^{(t)}$ using the labels y_i given by

$$CV^{(t)} := \frac{1}{n} \sum_{i=1}^{n} L(\hat{f^{-i}}(\mathbf{z}_i^{(t)}), y_i)$$

where $\hat{f^{-i}}$ is the fitting function computed with the ith part of the data removed and L is 1 if $\hat{f^{-i}}(\mathbf{z}_i^{(t)}) = y_i$ and 0 otherwise. We then select the optimal t value t^{opt} defined as

$$t^{\text{opt}} := \text{argmin}(CV^{(t)})$$

which yields the smallest cross validation value. (We may assume always that in practice we are given a finite collection of t values so that a minimum exists). For t^{opt}, we now perform random walk diffusion dimension reduction on the high dimensional spatial temporal feature space and produce n, d dimensional reduced training data $(\mathbf{z}_i^{t^{\text{opt}}}, y_i)$, $i = 1, \ldots, n$ and m, d dimensional reduced testing data $\mathbf{z}_i^{t^{\text{opt}}}$. Now we compute for each $i = 1, ..., m$, a $k = 3$ nearest neighbor classifier fit $\hat{y}_i$ for each testing data point $\mathbf{z}_i^{t^{\text{opt}}}$. See [8] and finally we approximate a prediction risk by the formula

$$R_{\text{pred}} := \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}_i, y_i).$$

We note the method of [3] emphasizes eigenvectors corresponding to larger eigenvalues. Our method compensates for the the sensitive weight width parameters σ by choosing an optimal weight for the eigenspace. Finally we compare our new method with the same method (with $t = 1$) random walk diffusion and dual root minimal spanning tree diffusion.

4 Experimental Results

In this experiment, the video sequences were recorded using a single stationary monochrome CCD camera mounted in such a way that the actions are performed parallel to the image plane. The data set consists of actions performed by $s = 29$ different people. Each person performed $p = 8$ activities, as shown in Figure 5: walk, run, skip, line-walk, hop, march, side-walk, side-skip. The location and size of the person in the image plane is assumed to be available (e.g., through tracking). Each activity sequence by each person includes a full cycle of the activity. The number of frames per sequence therefore depends on the speed of each action. The entire image is processed as explained in Section 3.3.

In our experiments, we used the data for eight of the 29 subjects for training (64 video sequences). This leaves a test data set of 168 video sequences performed by the remaining 21 subjects. The training instances have label. The number of selected frames was arbitrarily set to 12. So, the full dimension d of the space is 775*12 dimensions.

Fig. 3. Frames from walk, run, skip, march, line-walk, hop,side-walk,side-skip actions respectively

We look at (a) Random walk diffusion+KNN); (b) Dual rooted diffusion+KNN);

4.1 Experimental Analysis

The results of Table 1 are considerably lower and less sensitive to the choice of σ than those of Table 2 and Table 3. The reason for this is that the choice of optimal hitting times compensates for the sensitivity of σ by choosing optimal weighted eigenspaces. On the otherhand, the results of Table 2 and Table 3 show that the classifier method using random walk diffusion(for $t = 1$) and dual rooted minimal spanning tree diffusion is sensitive to the choice of σ. The results of Table 2 and Table 3 show that, in general, the method of random walk diffusion

Table 1. Proposed method using cross validation to select best heating time t, and based on Lafon's Random walk diffusion+KNN

	d=3	d=5	d=10	d=20	d=100	d=150	d=200	
$\sigma = 6$	53.57%	53.57%	46.43%	51.79%	53.57%	57.74%	70.24 %	73.21%
$\sigma = 8$	57.14%	45.24%	48.81%	44.05%	44.05%	46.43%	44.05%	50%
$\sigma = 10$	54.67%	56.55%	51.19%	43.45%	41.07%	44.64%	47.02%	44.05%
$\sigma = 12$	50.6%	53.57%	38.69%	38.69%	36.31%	39.88%	40.48%	41.67%
$\sigma = 14$	57.74%	50.6%	39.88%	44.64%	47.02%	42.26%	42.26%	41.67%

Table 2. Random walk diffusion+KNN

	d=3	d=5	d=10	d=20	d=100	d=150	d=200	
$\sigma = 6$	53.57%	53.57%	46.43%	51.79%	56.55%	59.52%	73.21%	84.52%
$\sigma = 8$	58.93%	45.24%	48.81%	50.6%	48.81%	59.52%	69.64%	87.5%
$\sigma = 10$	57.14%	56.55%	51.19%	44.64%	44.64%	58.33%	69.05%	86.9%
$\sigma = 12$	62.5%	55.95%	45.83%	42.86%	44.64%	60.12%	70.83%	87.5%
$\sigma = 14$	57.74%	50.6%	39.88%	44.64%	47.02%	60.12%	72.62%	87.5%

Table 3. Dual rooted diffusion+KNN

	d=3	d=5	d=10	d=20	d=100	d=150	d=200	
$\sigma = 1/6 * max(max(Ahop))$	64.29%	64.88%	60.12%	63.69%	66.07%	65.48%	74.4%	83.33%
$\sigma = 1/8 * max(max(Ahop))$	61.9%	61.31%	61.31%	63.69%	61.31%	64.88%	73.81%	88.1%
$\sigma = 1/10 * max(max(Ahop))$	61.31%	63.69%	62.5%	57.14%	62.5%	66.67%	70.83%	83.93%
$\sigma = 1/12 * max(max(Ahop))$	60.71%	60.12%	62.5%	60.12%	61.31%	66.07%	74.4%	82.74%
$\sigma = 1/14 * max(max(Ahop))$	59.52%	60.12%	61.31%	63.1%	59.52%	64.29%	75%	82.74%

(for $t = 1$) does better than dual rooted minimal spanning tree diffusion over a range of dimensions and σ and are comparible for large dimensions and varying σ.

5 Summary

In this paper, we studied the problem of classification of spatial temporal feature actions from video. Our new idea is to use class labels to find optimal hitting times $t \geq 1$ for dimension reduction using diffusion via random walks. We used our methods on real data, and compared our new method with the same method (with $t = 1$) for random walk diffusion and dual root minimal spanning tree diffusion. The results by our former method are shown to be considerably better as the choice of optimal hitting time compensates for sensitive weight width parameters in affinity matrices and chooses a natural weighted eigenspace. An interesting idea in this work is that we do not need to assume that the space consisting of all features is given by a manifold. Although our method is implemented on heat type kernels, our method works for any non negative, symmetric, positive definite kernel. Experimental results also show that taking any fixed k in the nearest neighbor fit produces results of the same order.

Acknowledgement The authors want to thank Professor A. Hero from University of Michigan and Dr. T. Wittman from University of Minnesota for useful discussions. S.B. Damelin is supported, in part by grants EP/C000285 and NSF-DMS-0439734 and thanks the Institute for Mathematics and Applications for their hospitality.

References

1. M. Belkin, P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Neural Computation, June 2003; **15** (6), pp. 1373-1396.
2. R. Chellappa, A. Chowdhury and S. Zhou, *Recognition of Humans and Their Activities Using Video*, Morgan Claypool, 2005.
3. R. R. Coifman, S. Lafon, *Diffusion Maps*, submitted to Applied Computational and Harmonic Analysis, 2004.
4. S. B. Damelin and M. Werman, Learning curved Manifolds via Diffusion Map embeddings into Riemann spaces and their applications, manuscript.
5. J. Davis and A. Bobick, *The Representation and Recognition of Action Using Temporal Templates*, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, pp. 928-934, June 1997.
6. S. Grikschat, J. Costa, A. Hero and O. Michel, *Dual rooted diffusions for clustering and classification on manifolds*, manuscript.
7. N. Jin and F. Mokhtarian, *Human Motion Recognition based on Statistical Shape Analysis*, Proc. IEEE International Conference on Advanced Video and Signal based Surveillance, pp. 4-9, 2005.
8. T. Hastie, R. Tibshirhani and J. Friedman, *The Elements of Statistical Learning*, Springer 2001.

9. R. Kohavi and G. John, *Wrappers for feature subset selection*, Artificial Intelligence, **97**, pp. 273-324, 1997.
10. Y. Ma, B. Miller, P. Buddharaju and M. Bazakos, *Activity Awareness: From Predefined Events to New Pattern Discovery*, 2006 IEEE International Conference on Vision Systems, New York, USA, January 5-7, 2006.
11. O. Masoud, N.P. Papanikolopoulos, *A method for human action recognition*, Image and Vision Computing, **21**, no. 8, pp. 729-743, Aug. 2003.
12. G. Medioni, I. Cohen, F. Bremond, S. Hongeng and R. Nevatia, *Event Detection and Analysis from Video Streams*, IEEE Transactions on pattern analysis and machine intelligence, **23**. no 8, 2001.
13. A. Ng, M. Jordan and Y. Weiss, *On spectral clustering:analysis and an algorithm*, Neural Information Processing Systems, **14**, 2001.
14. F. Porikli and T. Haga, *Event detection by Eigenvector Decomposition using object and feature frame*, CVPR workshop, pp. 114-114, 2004.
15. B. Scholkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.
16. V. D. Shet, D. Harwood and L. S. Davis, *VidMAP: Video Monitoring of Activity with Prolog*, IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Como, Italy, September 2005.
17. V. Vu, F. Bremond and M. Thonnat, *Video surveillance: human behaviour representation and on-line recognition*, The Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Podere d'Ombriano, Crema, Italy, September 2002.
18. H. Zhong, J. Shi and M. Visontai, *Detecting Unusual Activity in Video*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR), 2004.
19. T. Wittman, Manifold learning matlab demo, http://www.math.umn.edu / ∼ wittman/mani/

Y. Ma, Honeywell Labs, 3660 Technology Drive, Minneapolis, MN 55418
email: Yunqian.Ma@honeywell.com

S. B. Damelin, Institute for Mathematics and its Applications, University of Minnesota, 400 Lind Hall, 207 Church Hill, Minneapolis, MN 55455
email: damelin@georgiasouthern.edu