Geometries of sensor outputs, inference and information processing

Ronald R Coifman[1,2], Stephane Lafon[3], Mauro Maggioni[1,2], Yosi Keller[1,2], Arthur D Szlam[1,2], Frederick J Warner[1,2], Steven W Zucker[2,4]

[1]Department of Mathematics, [2]Program in Applied Mathematics, [4]Department of Computer Science, Yale University, 10 Hillhouse Ave, New Haven, CT, 06520; [3]Google Inc., 1600 Amphitheatre Pkw., Mountain View, CA, 94043.

ABSTRACT

We describe signal processing tools to extract structure and information from arbitrary digital data sets. In particular heterogeneous multi-sensor measurements which involve corrupt data, either noisy or with missing entries present formidable challenges. We sketch methodologies for using the network of inferences and similarities between the data points to create robust nonlinear estimators for missing or noisy entries. These methods enable coherent fusion of data from a multiplicity of sources, generalizing signal processing to a non linear setting. Since they provide empirical data models they could also potentially extend analog to digital conversion schemes like "sigma delta".

Keywords: Markov processes, multiscale analysis, diffusion on manifolds, Laplace-Beltrami operator.

1. FEATURE BASED FILTERING, DIFFUSIONS AND SIGNAL PROCESSING ON GRAPHS

A simple way to understand the effect of introducing similarity based diffusions on data[1–6] is provided by considering a regular gray level image in which we associate with each pixel p a vector $\nu(p)$ of features.[7,8] For example, a multi-band electromagnetic spectrum or the 5×5 sub-image centered at the pixel, or any combination of features. Define a Markov filter

$$A_{p,q} = \frac{\exp -\frac{||\nu(p)-\nu(q)||^2}{\epsilon}}{\sum_q \exp \frac{-||\nu(p)-\nu(q)||^2}{\epsilon}}\,, \tag{1}$$

where $\epsilon > 0$ is a small parameter comparable to the smallest distances between two feature vectors $\nu(p)$ and $\nu(q)$. Clearly the map ν is a bijection between pixels in the image and patches (or features). In particular every function on the pixels, such as the original image I itself, is also a function on the set of patches. With this identification, one can let the Markov filter $A_{p,q}$ act on an image.

The image I in figure 1 was filtered using the (nonlinear in the features) procedure described above where the feature vector $\nu(p)$ is the 5×5 patch around a pixel p:

$$I(p) = \sum_q A_{p,q} I(q) = \sum_q \frac{\exp -\frac{||\nu(p)-\nu(q)||^2}{\epsilon}}{\sum_q \exp \frac{-||\nu(p)-\nu(q)||^2}{\epsilon}} I(q)\,. \tag{2}$$

Observe that the edges are well preserved as patches translated parallel to an edge are similar and contribute more to the averaging procedure.[7,8] We should also observe that if we were to repeat the procedure on the filtered image we would get a numerical implementation of various nonlinear heat diffusions for image processing similar to those in PDE methods, such as those by Osher and Rudin.

It is useful to replace A by a bi-Markovian version of the form

$$A_{p,q} = \frac{\exp \frac{-||\nu(p)-\nu(q)||^2}{\epsilon}}{\omega(p)\omega(q)}$$

Send correspondence to R.R. Coifman. E-mail: coifman@fmah.com, Telephone: 1 203 432 1213

Figure 1. Left: original noisy image. Right: image denoised by application of the Markov matrix as in (1)

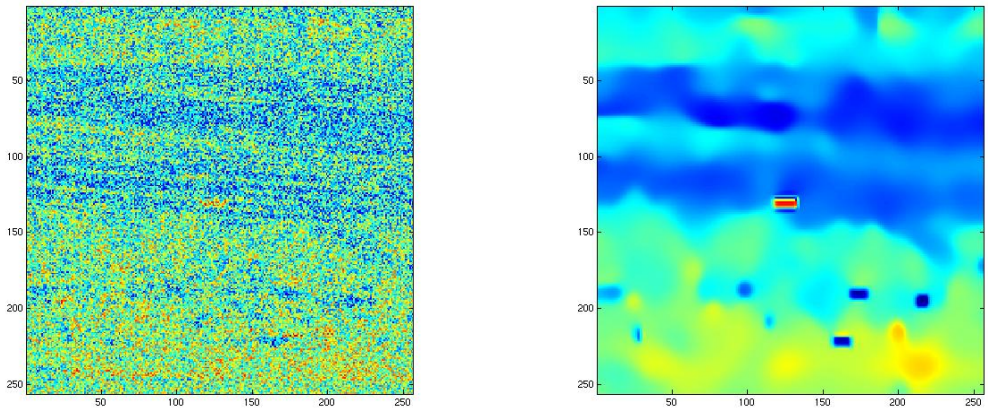

Figure 2. Left: original noisy image. Right: image denoised by application of the Markov matrix as in (1), but where features are local variances rather than pixel values in a patch around each pixel.

where the weights $\omega(\cdot)$ are selected so that A is Markov in p and q.

The noisy IR image in Figure 2 was filtered by N. Coult using a vector of 25 statistical features associated with each pixel.

The Markov matrix used for filtering, defines a diffusion on the graph of patches or features viewed as a subset of 25 dimensional Euclidean space. The eigenvectors of this diffusion permit us to compute all of its powers and to define a diffusion geometry and signal processing on this "image graph".[7]

For the next example consider 3 noisy sensors measuring the x, y, z coordinates of a trajectory in three dimensions. We could try to denoise each coordinate separately. Or use the position vector as as a feature vector as we did for the images above. See Figure 1.

The construction above should be viewed as signal processing on the data graph. We view all points of the trajectory as a data graph,ie data points p and q are vertices and $A_{p,q}$ is the weight of the edge connecting them

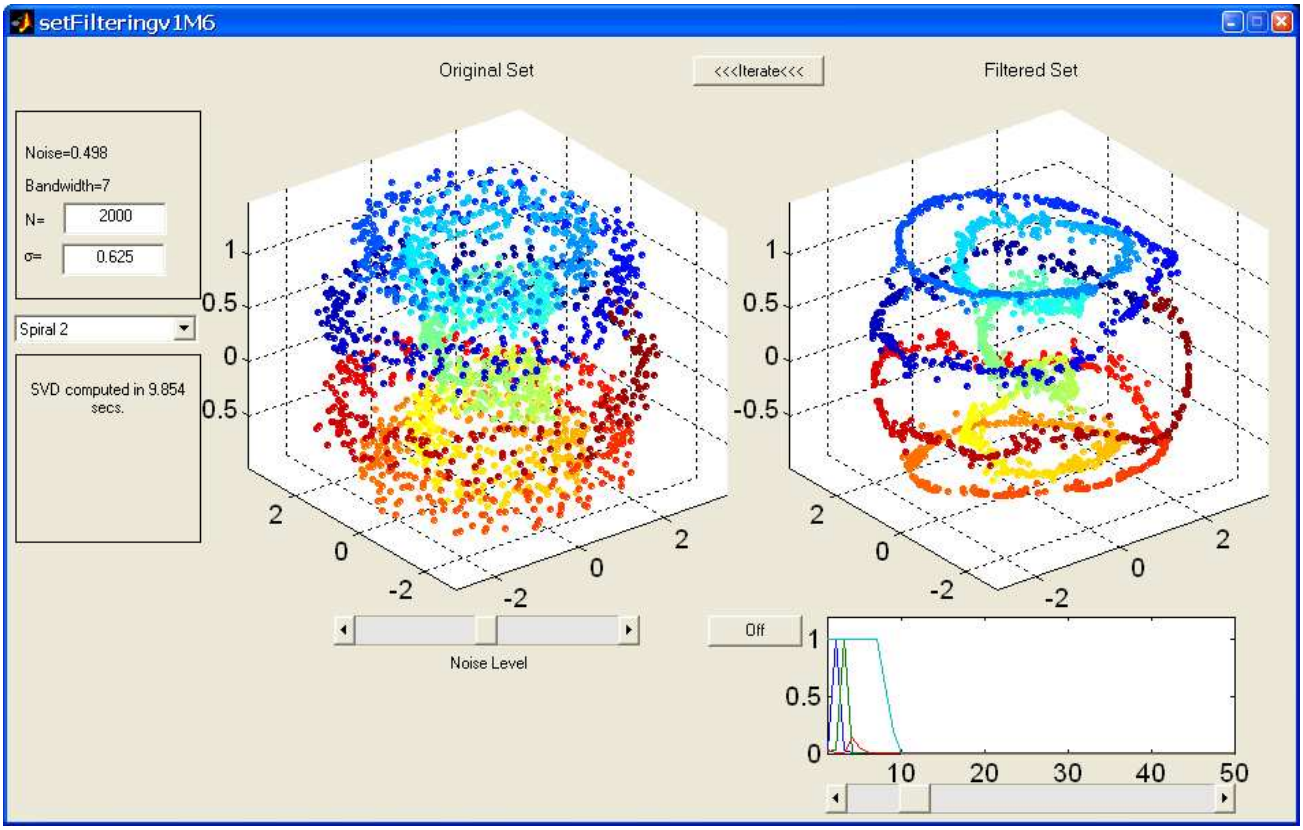

Figure 3. The green, red and blue curves are respectively the coefficients of the x, y, z coordinates , as filtered above, using less than 10 eigenvectors of the Markov matrix. These simple examples indicate that diffusion geometries are an efficient tool for sensor fusion ,and coherent signal processing for nonlinearly correlated data streams.

measuring their similarity or affinity at scale 1. We consider the eigenvectors of the Markov matrix $A_{p,q}$ defined above as a basis for all functions on this Graph. We can then expand each coordinate as a function on that graph, and restrict the expansion to the first few low frequency eigenfunctions, ie filter it and use the filtered coordinates as a clean trajectory.[1] This generalizes the simple filtering obtained above (see Figure 1).

2. DIFFUSION GEOMETRIES

These simple examples indicate that diffusion and harmonic analysis are useful for coherent sensor integration and fusion, enabling signal processing for nonlinearly correlated data streams. Diffusion geometries enable the definition of affinities and related scales between any digital data points in (provided of course that the infinitesimal proximity in the coordinates corresponds to true affinity between data points). Moreover it enables the organization of the population of sensor output into affinity folders or subsets with a high level of affinity among their responses . In particular we claim that the eigenfunctions of the diffusion operator or equivalently a Laplacian on a graph provide useful empirical coordinates, which enable an embedding of the data to low dimensional spaces so that the diffusion distance at time t on the original data becomes Euclidean distance in the embedding, in effect providing a nonlinear version of the SVD.[1] Moreover we indicate how the diffusion at different times leads to a multiscale analysis generalizing wavelets and similar scaling mechanisms.[4,6,9]

To be specific,[1,10] let the bi-Markov matrix A defined above be represented in terms of its eigenvectors:

$$A_{p,q} = \sum_l \lambda_l^2 \varphi_l(X_p)\varphi_l(X_q)$$

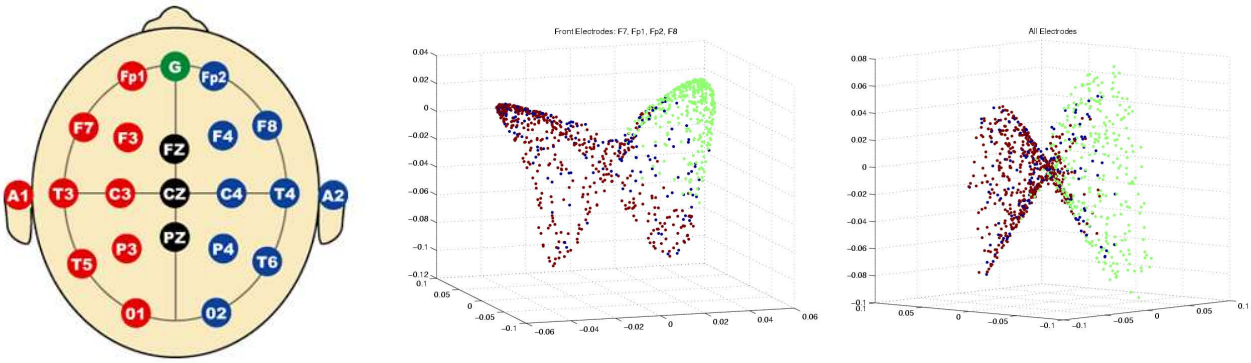

Figure 4. Left: standard position of electrodes in EEG. Middle: diffusion map of the responses to 4 electrodes, showing the nonlinear correlations and manifold-like structure of these responses. Right: diffusion map of the responses to all electrodes, exhibiting similar nonlinear correlations. In fact, the manifold structure obtained from measuring from all electrodes is very close to that obtained from 4 electrodes, suggesting that exploiting the nonlinear correlations would allow to use only 4 electrodes.

and define the diffusion map $\Phi_m^{(t)}$ at time t into m dimensional Euclidean space by

$$X_p \mapsto \Phi_m^{(t)}(X_p) := (\lambda_1^t \varphi_1(X_p), \lambda_2^t(X_p), \ldots, \lambda_m^t \varphi_m(X_p)) \tag{3}$$

For a given t we determine m so that λ_{m+1}^t is negligible. The diffusion distance[1] at time t between $X_p^{(t)}$ and $X_q^{(t)}$ is given as

$$d_t^2(p,q) = A_{p,p} + A_{q,q} - 2A_{p,q} = \sum_l \lambda_l^{2t}(\varphi_l(X_p) - \varphi_l(X_q))^2 = ||\Phi_m^{(t)}(X_p) - \Phi_m^{(t)}(X_q)||^2 .$$

This map enables us to represent geometrically an abstract set of measurements on a sensor array (measurement space) as we illustrate on the following EEG example.[11]

The 20 electrodes measure coherent electrical activity in the brain. Mapping the configuration space of the measurements of 4 electrodes leads to the same configuration as for all 20. In the linear case this will be obtained by de-correlating the outputs , here however different locations of sources result in a different attenuation vectors ,or linear de-correlations. Here the first three nontrivial eigenvectors are used to map the data to three dimensions (diffusion map), see Figure 4. The implications are obvious 4 electrodes suffice to get essentially the same measurements , redundancy is useful to obtain a clean version.[11]

3. MULTISCALE STRUCTURES AND THE EMERGENCE OF ABSTRACT SENSOR FEATURES

It is possible to build a multiscale decomposition of a data graph simply by organizing the data into affinity folders where the affinity is measured through the diffusion distance at different time scales A simple algorithm[9] is obtained as follows Let x_j^{l+1} be a maximal sub-collection of points in $\{x_j^1\}$ (key-points at scale 1) such that $d_{t_l}(x_j^{l+1}, x_i^{l+1}) \geq \frac{1}{2}$, where x_j^0 are the original points, and $t_l = a2^l$, $l = 0, 1, 2, \ldots$. Then clearly each point is at distance less than a half at scale l from one of the selected key-points allowing us to create a folder labeled by the key-point. It is easy to modify to obtain a tree of disjoint folders by viewing each key point as the folder of points nearest to it, and reinterpret the distance as distance between folders.

When applied to text documents (equipped with semantic coordinates), this construction builds an automatic folder structure with corresponding keywords characterizing the folders.[4,7] While for text documents folders are just collection of related documents, and abstractions are collection of words in a given class, the situation is

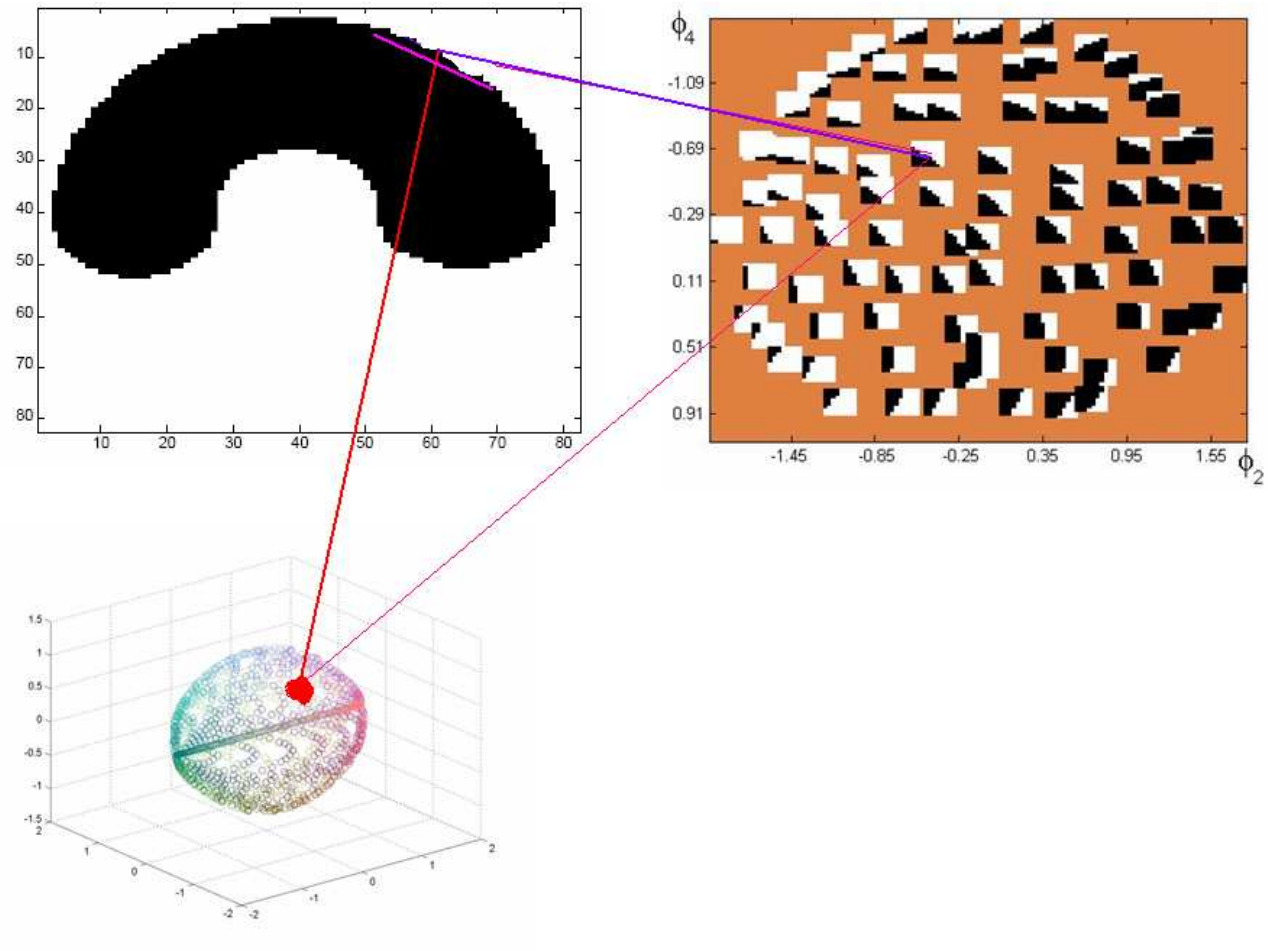

Figure 5. Organization of the set of patches from an image. Top left: original image; bottom right: diffusion map of the set of patches from the image; top right: the patches from image.

even more interesting for sensor outputs where the language occurs through self organization of data into affinity folders.

To illustrate this point consider Figure 5

To organize the black and white image in Figure 5 we have considered all 8×8 patches as our primitive data set forming the graph.[1, 7, 9] The first 2 eigenfunctions map them to the top right image,the first three to the image at the bottom left ,we see that only two parameters emerge, the orientation and the number of black pixels. If we now pick a little diffusion neighborhood say red patch on the 3d graph, it corresponds exactly to a little curved edge on the boundary of the original black spot on the image . While simple, observe that the organization is automatic requiring no a priori geometric information, and a rudimentary visual cortex has emerged only through observation of 8×8 patch data TODO.

One can modify this basic construction of a hierarchical scale decomposition in order to build scaling functions and wavelets on the graph/manifold,[4, 12, 13] which provide filters restricting the frequency content of a function to bands of eigenfunctions of the diffusion or Laplace operator on the graph.

4. SENSOR FUSION

For a heterogeneous sensor system each category of sensors can be parametrized and normalized in its intrinsic diffusion coordinates. A new graph is then created combining the relevant diffusion coordinates emanating from

	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
zero	**0.90**	0.00	0.00	0.00	0.00	0.00	0.06	0.04	0.00	0.00
one	0.00	**0.99**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
two	0.00	0.00	**0.96**	0.01	0.02	0.00	0.00	0.00	0.00	0.00
three	0.00	0.00	0.00	**0.99**	0.00	0.00	0.00	0.00	0.00	0.01
four	0.00	0.00	0.00	0.04	**0.96**	0.00	0.00	0.00	0.00	0.00
five	0.00	0.00	0.00	0.00	0.00	**0.97**	0.00	0.00	0.02	0.01
six	0.06	0.00	0.00	0.00	0.00	0.00	**0.90**	0.04	0.00	0.00
seven	0.03	0.00	0.00	0.00	0.00	0.00	0.03	**0.93**	0.00	0.00
eight	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	**0.95**	0.03
nine	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.02	**0.96**

Figure 6. Classification results for the diffusion spelling scheme combining both channels, over 50 random trials.

different species of sensors as coordinates. As an example of integration of audio and video sensors we recorded several grayscale movies depicting the lips of a subject reading a text in English and retained both the video sequence and the audio track. Each video frame was cropped into a rectangle of size 140×110 around the lips and was viewed as a point in $\mathbb{R}^{140\times 110}$. We took the log of the power spectrum of the window between two frames as the audio vectors. We used a small vocabulary of ten words, zero, one, two, ... nine for training and testing a simple classifier. To each spoken digit corresponded a small trajectory, i.e. "spelling" in the diffusion geometry of the combined model. The combined graph was built from a feature representation of the data based on appending the first 5 dimensional diffusion embedding of the audio channel with the first 5 dimensional embedding of the video stream. A new graph is constructed from this collection of points in 10 dimensions , this graph is then embedded in lower dimensions and the trajectories of words on it (diffusion spelling) gives a classification (see Figure 6) substantially superior to either audio or video alone.

Observe that the goal here is to do ab initio learning with no a priori assumptions or knowledge.

5. ANALYSIS OF NOISY OR CORRUPT DIGITAL DATA IN MATRICES

As seen above an affinity structure on a collection of points in Euclidean space leads to diffusion geometries. More generally this data-driven geometric self organization also enables to analyze any data matrix according to its intrinsic row or column structure. This procedure is useful even for purposes of achieving more efficient numerical analysis, an analysis which generalizes the singular value decomposition, the fast multipole methods and various other numerical compression methods. We claim that it is useful to view a data matrix as a function on the tensor product of the graph build from the columns of the data with the graph of the rows of the data In other words the original data matrix becomes a function of the joint inference structure (Tensor Graph), and can be expanded in terms of any basis functions on this joint structure.

As is well known any basis on the column graph can be tensored with a basis on the row graph, but other combined wavelet bases can also be obtained. As seen above we can use the rows as well as the columns of the data to build two graphs which are then merged to a single combined structure (this procedure was done above for any two graphs permitting a fusion of two different structures). A simple matrix processing or filtering scheme is provided below: given data entries $d(q, r)$ where, for illustration we can think of the rows q as sensors and the columns r as responses:

$$D(q, r) = \sum_{\alpha,\beta} \delta_{\alpha,\beta} \varphi_\alpha(q) \varphi_\beta(r) \,, \tag{4}$$

where φ_α is a (e.g. wavelet) basis on Q, and $\varphi_\beta(r)$ is a (wavelet) basis on R. In the formula above

$$\delta_{\alpha,\beta} = \sum_{q,r} d(q,r)\varphi_\alpha(q)\varphi_\beta(r)\,,$$

where we accept this sum (as validated) only if various randomized averages using subsamples of our data lead to the same value of $\delta_{\alpha,\beta}$. In the calculation of D we only use accepted estimates for $\delta_{\alpha,\beta}$.

The wavelet basis can of course be replaced by tensor products of scaling functions or any other approximation method in the tensor product space, including other pairs of bases, one for q the other for r, including graph Laplacian eigenfunctions (we observe in passing that the singular value decomposition is a particular case of this construction). A direct method for filtering d or estimating D without the need to build basis functions can be implemented as at the beginning of this paper.

Define a Markov matrix $A = a[(r,q),(r',q')]$ (corresponding to diffusion on $Q \times R$) as

$$a[(r,q),(r',q')] = \frac{\exp\left(\frac{||\nu(r)-\nu(r')||^2}{\epsilon} + \frac{||\mu(q)-\mu(q')||^2}{\delta}\right)}{\sum_{r,q}\exp\left(\frac{||\nu(r)-\nu(r')||^2}{\epsilon} + \frac{||\mu(q)-\mu(q')||^2}{\delta}\right)} \tag{5}$$

Where the vector $\nu(r)$ is response column vector corresponding to the column r, and $\mu(r)$ is a sensor row vector.

The parameters epsilon, delta are chosen after randomized validation as described above. We can have an alternate definition of D as follows.

$$D(r,q) = \sum_{r,q} a[(r,q),(r',q')]d(r,q)\,.$$

Observe that the distances occurring in the exponent can be replaced by any convenient notion of distance or dissimilarities, and that any polynomial in A can be used to obtain a better filtering operation on the raw data.

A new combined graph can also be formed by embedding the graph $Q \times R$ into Euclidean space ,say by the diffusion embedding , followed by an expansion of the data $d(q,r)$ on this new structure, or by filtering as above on the new structure.

5.1. Markov Decision Processes

In the papers[14, 15] the multiscale analysis construction of diffusion wavelets is applied to Markov Decision Processes. Informally, and in a simplified version, one or more agents explore a given *state space* S by taking actions in each state from a set of actions A, and collect different *rewards* R, that we assume, to simplify the presentation, to depend only on the location and not on the action. Suppose we can model the state space as a finite graph (S,E,W) (the uncountable or continuous case can be handled as well), with edges E and weights W, and that the agent(s) explore the state space randomly accordingly to the Markov process P^π, parametrized by a (*policy*) π, which maps each state to a probability distribution of actions for that state. The reward function R is a real-valued function on S. The expected long term sum of discounted rewards when the agent follows the policy π is a function V^π on S, called (state) *value function*. It satisfies the so-called Bellman equation $V^\pi = R + \gamma P^\pi V^\pi$, $\gamma \in (0,1]$ being the discount factor, and hence $V^\pi = (I - \gamma P^\pi)^{-1}R$. In terms of potential theory, $(I - P^\pi)^{-1}$ is the Green's function (or fundamental matrix) of the "Laplacian" $I - P^\pi$, and V^π is the potential generated by the "charge" R under the diffusion P^π. Suppose for simplicity that P^π is reversible: it is then similar to a symmetric matrix T^π that generates a Markov diffusion semigroup $\{(T^\pi)^t\}$. The diffusion multiscale analysis allows to efficiently compute $(P^\pi)^t(x,y)$ for arbitrary t, medium and large, for one or multiple agents; it allows to effectively approximate the value function V^π, which is often piecewise smooth, performing a very useful dimensionality reduction,[14] where *ad hoc* basis functions were previously constructed by hand and were only available in particularly simple geometries. Finally, it allows to solve Bellman's equation directly, to high precision, in an efficient way. In[15] this method is compared with classical direct methods (often unfeasible because of their computational complexity of $\mathcal{O}(|S|^3)$), and with optimized iterative solvers.

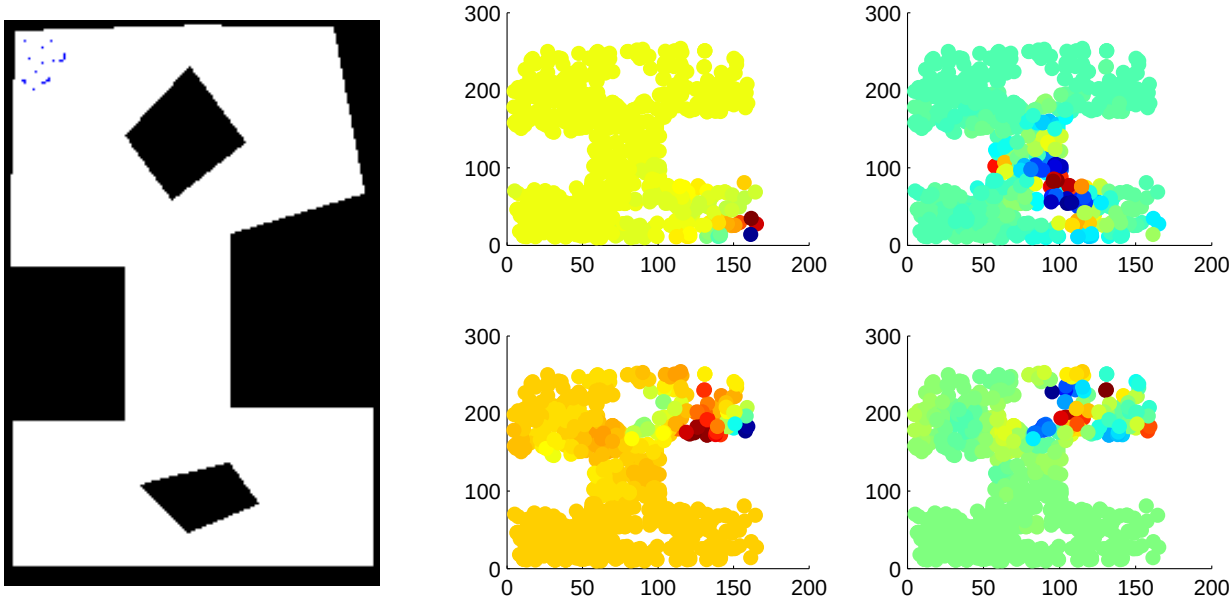

Figure 7. Left: continuous state space for a MDP, the actions are movements in the four cardinal directions, blue points represent positive rewards. Right: after a random exploration by the agent, multiscale bases functions are constructed on the state space: the color is proportional to the value of various scaling functions, which are automatically adapted to the state space. The value function can be projected onto this basis, in fact if the value function is piecewise smooth, only few elements of the basis (a number independent of the number of samples!) will be required to approximate the value function to a given precision.

6. CONCLUSIONS AND DISCUSSION

It is quite clear from the preceding descriptions that the data graph can be equipped with informative geometric structures which coherently integrate data and enable inference and interpolation. One of our main goals is to efficiently regress empirical functions on a data set, we have indicated various methods to build and approximate empirical functions, admitting natural extensions (generalization) off the known measured data. We also indicated that signal processing on data could be achieved without any knowledge of the data model, by letting the intrinsic data geometry emerge through a natural process of affinity diffusion. Modern sensor systems such as radar, hyperspectral, MRI and others actually do not measure images but much more elaborate vectors, the images are built to allow understanding and further processing, in reality we should let the intrinsic geometry of the measurements participate in the information extraction. Such an approach has been developed by our team for hyperspectral imaging.

We also observe that in the context of compressed sensing where the sensor inputs are randomly encoded. The projection into a random coded subspace while maintaining the relative affinity of the original data points permits rebuilding the data geometry by tools described above.

7. ACKNOWLEDGEMENTS

RC is partially supported by DARPA and NSF. MM is partially supported by NSF Grant DMS 0512050. SWZ is partially supported by AFRL.

REFERENCES

1. S. Lafon, *Diffusion maps and geometric harmonics.* PhD thesis, Yale University, Dept of Mathematics & Applied Mathematics, 2004.
2. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comp. Harm. Anal.* , 2004.

3. R. Coifman and S. Lafon, "Geometric harmonics," *Appl. Comp. Harm. Anal.* , 2004 2004. Submitted.
4. R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Tech. Rep. YALE/DCS/TR-1303, Yale Univ., Appl. Comp. Harm. Anal.* , Sep. 2004. to appear.
5. R. R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data. part i: Diffusion maps," *Proc. of Nat. Acad. Sci.* , pp. 7426–7431, May 2005.
6. R. R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data. part ii: Multiscale methods," *Proc. of Nat. Acad. Sci.* , pp. 7432–7438, May 2005.
7. R. R. Coifman and M. Maggioni, "Multiscale data analysis with diffusion wavelets," Tech. Rep. YALE/DCS/TR-1335, Dept. Comp. Sci., Yale University, September 2005.
8. A. D. Szlam, *Non-stationary analysis of datasets and applications.* PhD thesis, Yale University, Dept of Mathematics & Applied Mathematics, 2006.
9. R. R. Coifman, M. Maggioni, S. W. Zucker, and I. G. Kevrekidis, "Geometric diffusions for the analysis of data from sensor networks," *Curr Opin Neurobiol* **15**, pp. 576–84, October 2005.
10. M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation* **6**, pp. 1373–1396, June 2003.
11. E. Causevic, R. R. Coifman, R. Isenhart, A. Jacquin, E. R. John, M. Maggioni, L. S. Prichep, and F. J. Warner, "QEEG-based classification with wavelet packets and microstate features for triage applications in the ER," Oct 2005. ICASSP 05.
12. M. Maggioni, J. C. Bremer Jr, R. R. Coifman, and A. D. Szlam, "Biorthogonal diffusion wavelets for multiscale representations on manifolds and graphs," August 2005. Proc. SPIE Wavelet XI.
13. M. Maggioni, A. D. Szlam, , R. R. Coifman, and J. C. Bremer Jr, "Diffusion-driven multiscale analysis on manifolds and graphs: top-down and bottom-up constructions," August 2005. Proc. SPIE Wavelet XI.
14. S. Mahadevan and M. Maggioni, "Value function approximation with diffusion wavelets and laplacian eigenfunctions," in *University of Massachusetts, Department of Computer Science Technical Report TR-2005-38; Proc. NIPS 2005*, 2005.
15. M. Maggioni and S. Mahadevan, "Fast direct policy evaluation using multiscale analysis of markov diffusion processes," in *University of Massachusetts, Department of Computer Science Technical Report TR-2005-39; submitted*, 2005.