INFORMATION PRESERVING EMBEDDINGS FOR DISCRIMINATION

Kevin M. Carter[1]*, Christine Kyung-min Kim*[1]*, Raviv Raich*[2]*, Alfred O. Hero III*[1]

[1] Department of EECS, University of Michigan, Ann Arbor, MI 48109
[2] School of EECS, Oregon State University, Corvallis, OR 97331
{kmcarter,poppins,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

Dimensionality reduction is required for 'human in the loop' analysis of high dimensional data. We present a method for dimensionality reduction that is tailored to tasks of data set discrimination. As contrasted with Euclidean dimensionality reduction, which preserves Euclidean distance or Euler angles in the lower dimensional space, our method seeks to preserve information as measured by the Fisher information distance, or approximations thereof, on the data-associated probability density functions. We will illustrate the approach for multi-class object discrimination problems.

***Index Terms*—** Information geometry, statistical manifold, dimensionality reduction, classification, object recognition

1. INTRODUCTION

Object recognition and discrimination is of critical importance in many application areas such as face recognition and surveillance. The standard formulation is that of the classic classification task – one is given a set of training data $\boldsymbol{I} = [(I_1, y_1), \ldots, (I_N, y_N)]$ where $y_i \in [1, \ldots, M]$ is the class label (or object type) of image I_i, and there are M different recognized object types. The task is then to find a function $f(I) : I \rightarrow y$ which assigns a class label y given an unknown test image I.

In many applications of practical interest, one has more available information than a single image of an object. Often systems have the capability to obtain images taken at different angles of the same object. For example, a camera may capture several shots while rotating about an object, or a camera may be stationary while an object rotates in a fixed plane. In these situations, the recognition task may be modified such that the training data is now $\mathcal{I} = \{(\boldsymbol{I}^1, y_1), \ldots, (\boldsymbol{I}^N, y_N)\}$, where $\boldsymbol{I}^i = [I_1, \ldots, I_{n_i}]$ is now a collection of n_i images of the same object, potentially taken at different vantage angles. We now wish to classify an unknown set of images $\boldsymbol{I}$ with some function $f(\boldsymbol{I}) : \boldsymbol{I} \rightarrow y$. This framework is similar to that which was present by Arandjelovic *et. al.* [1] where they performed facial recognition using sets of images. Specifically, they used the Kullback-Leibler (KL) divergence to define a similarity between image sets, and used a nearest neighbor classifier over the KL-divergences between the test set and the training sets. This method showed very promising results for the facial recognition task.

In this paper we propose an information-geometric approach to the problem of object recognition. By viewing the image sets as realizations of some generative model, i.e. probability density function (PDF), we can frame the problem as that of classification on a statistical manifold (or manifold of PDFs). This information-geometric modeling allows for the comparison between PDFs with the Fisher information distance, which is the natural metric on a statistical manifold. We will show that this approach yields competitive recognition performance, distinguishing different laptops and LCD monitors. We also employ the use of Fisher information nonparametric embedding (FINE) [2, 3], which provides an information-geometric embedding of the image sets into a low-dimensional Euclidean space. This is useful for visualization and 'human in the loop' analysis.

We note that this paper is not meant to be a seminal work on object discrimination, but rather a *proof-of-concept*. The methods (FINE) that we utilize for object recognition have been similarly used for document classification [3] and flow cytometry analysis and diagnosis [4]. We now illustrate a new application using FINE, and compare performance to existing methods.

1.1. Previous Work

We note that our methods of recognition are similar to those in [1], in which they operate on a manifold of densities in the 'face space'. While never mentioned, this is essentially describing a statistical manifold, and the KL-divergence maybe used as an approximation of the Fisher information distance [5]. However, the fact that the KL-divergence is unbounded leaves it to be a very unstable measure. Given a sample from a test image set which is very dissimilar from any appearing in the training set, the KL-divergence will approach infinity. They wisely account for this by using the appropriate direction of the divergence and estimating PDFs with Gaussian mixture models (GMMs), which dampens the effect of image outliers. This, however, raises two potential concerns.

First, since the KL-divergence is not a distance metric, they are unable to fully take a geometric approach to the problem. Secondly, GMMs are a parametric approach which requires parameter estimation for the mixtures, in terms of number of components, location, and covariance.

Our work may be considered as a more general case of [1], accounting for these potential concerns which may or may not become significant issues. We utilize the Hellinger distance as our Fisher information distance approximation, which satisfies all of properties of a distance metric. Hence, we operate in a full information-geometric manner. More importantly, the Hellinger distance is bounded, which offers a stable metric regardless of the input. Rather than parametrically estimating PDFs with a GMM, we use non-parametric kernel density estimation (KDE). While more computationally complex than GMMs, KDEs offer a better description of the PDF, and require only a single parameter (kernel bandwidth). Finally, we recognize that the approximations of the Fisher information distance are valid only in the limit, as PDFs approach one another on the manifold. Hence, we offer a geodesic approximation using graphical methods, which converges to the true information distance.

1.2. Paper Outline

This paper proceeds as follows. In Section 2, we give a motivation for the problem and a description for its difficulties. We give an overview of statistical manifolds and the Fisher information distance in Section 3, followed by a connection to the problem of object recognition in Section 4. We discuss simulation results on a real data set in Section 5, followed by conclusions in Section 6.

2. PROBLEM FORMULATION

The problem of object recognition from image sets is similar to the standard classification problem. One is given a collection of training data $\mathcal{I} = \{(\boldsymbol{I}^1, y_1), \ldots, (\boldsymbol{I}^N, y_N)\}$, where $\boldsymbol{I}^i = [I_1, \ldots, I_{n_i}]$ is a collection of n_i images $\{I_j\}$ of the same object. These images may be captured at different vantage points, showcasing different attributes of the object. We wish to classify an unknown set of images $\boldsymbol{I}$ with some function $f(\boldsymbol{I}) : \boldsymbol{I} \rightarrow y$.

Let us first illustrate the potential difficulties with this problem. Let $\mathcal{I}$ be a collection of ~ 150 image captures each of $N = 4$ unique objects. Each image is taken at a different angle, holding pitch constant while rotating the yaw (full details of image requisition can be found in Section 5.1). We use principal component analysis (PCA) on the entire collection of rasterized images (ie. $\boldsymbol{X} = [\boldsymbol{I}_1^1, \ldots, \boldsymbol{I}_{n_1}^1, \boldsymbol{I}_1^2, \ldots]$) to project each image onto the first 2 and 3 principal components of $\boldsymbol{X}$; Fig. 1 shows these results. One can naturally see a path formed which demonstrates the natural transition from one image taken at one yaw to the next taken with a slight change in yaw. It is also clear that the paths which different objects take are very similar, which would make it difficult to distinguish one from the other in most cases. Add that in practice, there may be $\ll 150$ available images per object, and the problem of differentiating image sets (i.e. recognition) becomes very difficult.

Looking at the trajectories, however, it becomes apparent that there is some generative model which governs the path. While any given point in an object trajectory may be difficult to distinguish from the path of a different object, the entire path maybe more easily discerned. We take a statistical approach by modeling each trajectory as a probability density function, which allows for an information-geometric framework to the problem.

3. STATISTICAL MANIFOLDS

Let us now present the notion of statistical manifolds, or a set $\mathcal{M}$ whose elements are probability distributions. A probability density function (PDF) on a set $\mathcal{X}$ is defined as a function $p : \mathcal{X} \rightarrow \mathbb{R}$ in which

$$p(x) \geq 0, \forall x \in \mathcal{X}$$

$$\int p(x)\, dx = 1. \tag{1}$$

If we consider $\mathcal{M}$ to be a family of PDFs on the set $\mathcal{X}$, in which each element of $\mathcal{M}$ is a PDF which can be parameterized by $\theta = \left[\theta^1, \ldots, \theta^n\right]$, then $\mathcal{M}$ is known as a statistical model on $\mathcal{X}$. Specifically, let

$$\mathcal{M} = \{p(x \mid \theta) \mid \theta \in \Theta \subseteq \mathbb{R}^d\}, \tag{2}$$

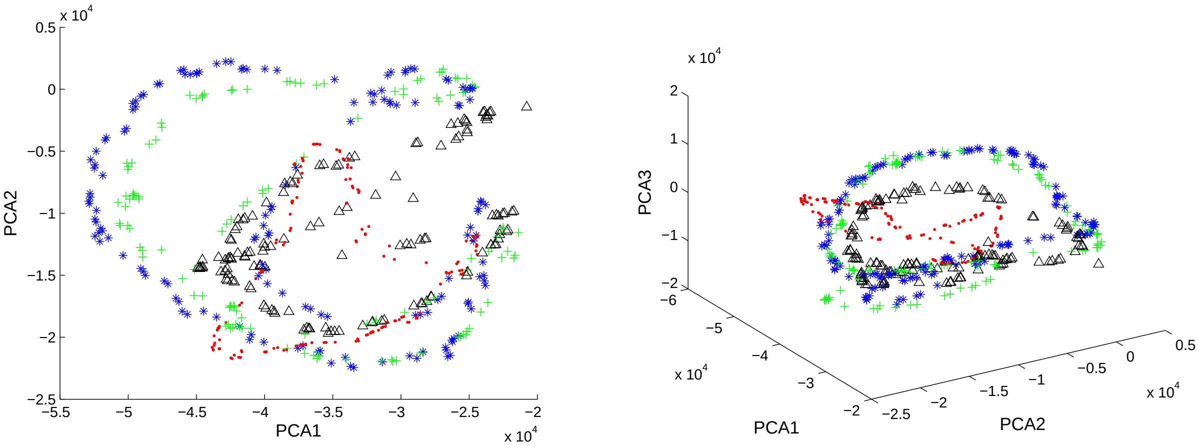
with $p(x \mid \theta)$ satisfying the equations in (1). Additionally, there exists a one-to-one mapping between θ and $p(x \mid \theta)$.

Given certain properties of the parameterization of $\mathcal{M}$, such as differentiability and C^∞ diffeomorphism (details of which are described in [6]), the parameterization θ is also a coordinate system of $\mathcal{M}$. In this case, $\mathcal{M}$ is known as a statistical manifold. In the rest of this report, we will use the terms 'manifold' and 'statistical manifold' interchangeably.

3.1. Fisher Information Distance

For a parametric family of probability distributions on a statistical manifold, it is possible to define a Riemannian metric using the Fisher information matrix $[\mathcal{I}(\theta)]$, which measures the amount of information a random variable contains in reference to an unknown parameter θ. The Fisher information distance between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ is:

$$D_F(\theta_1, \theta_2) = \min_{\substack{\theta(\cdot): \\ \theta(0)=\theta_1 \\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{d\theta}{dt}\right)^T [\mathcal{I}(\theta)] \left(\frac{d\theta}{dt}\right)}\, dt, \tag{3}$$



(a) First 2 PCs

(b) First 3 PCs

Fig. 1. Projected each image onto the first principal components (PCs). It is clear that there is some trajectory which is followed by each object, corresponding to the change in yaw in each image.

where $\theta = \theta(t)$ is the parameter path along the manifold [5, 6]. Note that the coordinate system of a statistical manifold is the same as the parameterization of the PDFs (i.e. θ). Essentially, (3) amounts to finding the length of the shortest path – the geodesic – on $\mathcal{M}$ connecting coordinates θ_1 and θ_2.

While the Fisher information distance cannot be exactly computed without a priori knowledge about the parameterization of the manifold, the distance between PDFs p_1 and p_2 may be approximated with the Hellinger distance,

$$D_H(p_1, p_2) = \sqrt{\int \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 dx}, \quad (4)$$

which converges to the Fisher information distance,

$$2D_H(p_1, p_2) \to D_F(p_1, p_2)$$

as $p_1 \to p_2$ [5]. This measures allow for the approximation of the information distance in the absence of the geometry of the statistical manifold on which the PDFs lie. For additional measures of probabilistic distance, some of which approximate the Fisher information distance, and a means of calculating them between data sets, we refer the reader to [7, 8].

If p_1 and p_2 do not lie closely together on the manifold, these approximations become weak. A good approximation can still be achieved if the manifold is densely sampled between the two end points.Using a graphical model, we may define the path between p_1 and p_2 as a series of connected segments. The geodesic distance may then be approximated as the sum of the length of those segments. Specifically, given the collection of N PDFs $\mathcal{P} = \{p_1, \ldots, p_N\}$ and using an approximation of the Fisher information distance $\hat{D}_F(p_1, p_2)$ as $p_1 \to p_2$, we can now define an approximation function G for all pairs of PDFs:

$$G(p_1, p_2; \mathcal{P}) = \min_{M,\mathcal{P}} \sum_{i=1}^{M-1} \hat{D}_F(p_{(i)}, p_{(i+1)}), \quad p_{(i)} \to p_{(i+1)} \forall i. \quad (5)$$

Intuitively, this estimate calculates the length of the shortest path between points in a connected graph on the well sampled manifold, and as such $G(p_1, p_2; \mathcal{P}) \to D_F(p_1, p_2)$ as $N \to \infty$. This is similar to the manner in which Isomap [9] approximates distances on Riemannian manifolds in Euclidean space.

4. OBJECT RECOGNITION

Given the information-geometric framework we have created, by approximating distances on statistical manifolds, we may now extend to the task at hand of object recognition. Specifically, we are given $\mathcal{I} = \{(\boldsymbol{I}^1, y_1), \ldots, (\boldsymbol{I}^N, y_N)\}$ as training data, and we may estimate the PDFs of each $\boldsymbol{I}^i$ as $p_i(\boldsymbol{I})$, for $i \in [1, N]$. This is performed using kernel density estimation (KDE) on the rasterized image sets; specific details of the KDE implementation may be found in [8]. Once the object class PDFs are estimated with the training data, test sets are classified by minimizing the information divergence between test and training sets. Let $\boldsymbol{I}$ be a test image set with estimated PDF $p(\boldsymbol{I})$, our classifier $y = f(\boldsymbol{I})$ is

$$f(\boldsymbol{I}) = \arg\min_i G(p(\boldsymbol{I}), p_i(\boldsymbol{I}); \mathcal{P}). \quad (6)$$

This may be essentially viewed as a 1-nearest neighbor classifier, using the Hellinger distance as an appropriate metric.

Algorithm 1 Fisher Information Nonparametric Embedding

```
Input: Collection of data sets $\mathcal{I} = \{\boldsymbol{I}^1, \ldots, \boldsymbol{I}^N\}$; the desired embedding dimension $d$
for $i = 1$ to $N$ do
    Calculate $\hat{p}_i(\boldsymbol{I})$, the density estimate of $\boldsymbol{I}^i$
end for
Calculate $G$, where $G(i, j)$ is the geodesic approximation of the Fisher information distance between $p_i$ and $p_j$
$\boldsymbol{Y} = \text{mds}(G, d)$
Output: $d$-dimensional embedding of $\mathcal{X}$, into Euclidean space $\boldsymbol{Y} \in \mathbb{R}^{d \times N}$
```

4.1. Visualization

Suppose now that visualization of several image sets is desired for 'human in the loop' analysis of the data. For example, as a recognition system stays online and accumulates new test data, an analyst may be interested in the comparative relationship between objects. For this task we refer the reader to Fisher information nonparametric embedding (FINE) [3], which finds an information-geometric embedding of PDFs into a low-dimensional Euclidean space. FINE operates by performing multidimensional scaling (MDS) on the matrix of pairwise dissimilarities formed by the information divergences (e.g. Hellinger distances). FINE is not tied to any specific method of MDS, and has been implemented with both classical MDS [10] and Laplacian Eigenmaps [11] for unsupervised dimensionality reduction. The full description of FINE may be found in Algorithm 1.

Note that once embedded in a Euclidean space, one may use other learning methods for classification. For example, while one cannot easily define a linear classifier between PDFs, this becomes a trivial task once those PDFs are embedded into Euclidean space with FINE.

5. SIMULATIONS

5.1. Data Setup

The data we will analyze was collected at Tech-edge building, in the Air Force Research Laboratory. The experiment was performed with 4 unique objects – 3 different model laptops and an LCD monitor. Each object was positioned on a swiveling desk, with a stationary camera (Canon VB-50iR) located above and to the left side of the object. The desk was then spun by a rope (so that no person is in the scene) and the camera captured still frames of the object at 15 fps with a 640×480 resolution, for roughly 10 seconds. An illustration of these retrieved data sets may be found in Fig. 2. Note that for each trial, the object was placed at the same location on the desk, and the desk was spun at an (attempted) equal speed.

Given the lack of unique objects, but the well sampled trajectories of the objects with changes in yaw, we may artificially manufacture "new" realizations of unique objects by subsampling along the trajectory. Specifically, let $\boldsymbol{I} = [I_1, \ldots, I_n]$, ordered according to change in object yaw, and let l be the sample spacing. Rather than having only 1 image set for the object, we can create n/l image sets by subsampling in the following manner:

$$\boldsymbol{I}^j = [I_j, I_{j+l}, I_{j+2l}, \ldots], \tag{7}$$

which generate uniformly spaced, i.i.d. realizations along the yaw trajectory. Although artificially generated, this is statistically equivalent to capturing a sequence of images from identical items which have been positioned differently (with respect to yaw). Note that each manufactured set has entirely unique images, so no two estimated PDFs will be identical. This is key as it simulates the setting for this object recognition task.

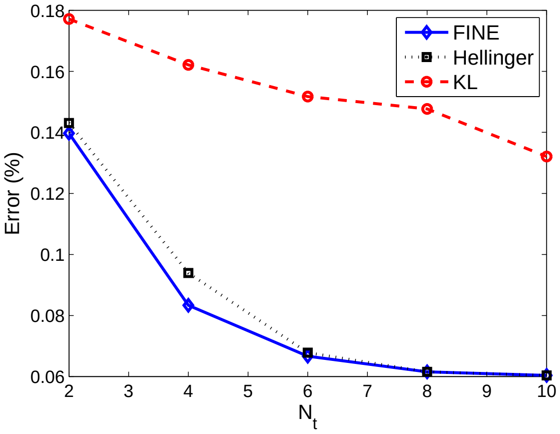


Fig. 3. Classification error rates for object recognition using different information divergences. The stability of the Hellinger distance for low sample sizes shows superior performance, garnering even better rates when using the geodesic approximation (FINE).

5.2. Results

We first wish to study the effect of test sample size on recognition capability. We begin by partitioning our training set to $\sim$ 10 sample images for each of the 4 objects, obtained with subsampling using (7). Next, we partition our test set using $\sim N_t$ samples per test object, with $N_t \in [2, 10]$. Given the small sample sizes, we preprocess the data by projecting each image onto the first 10 principal components of the entire collection. To test recognition capabilities, we use the 1-NN classifier (6) and plot the classification error, over a 10-fold cross validation, in Fig. 3. We also compare to the method presented in [1], which classifies by maximizing the KL-divergence between test set and training set. Note that we have modified the method to use a KDE rather than GMM

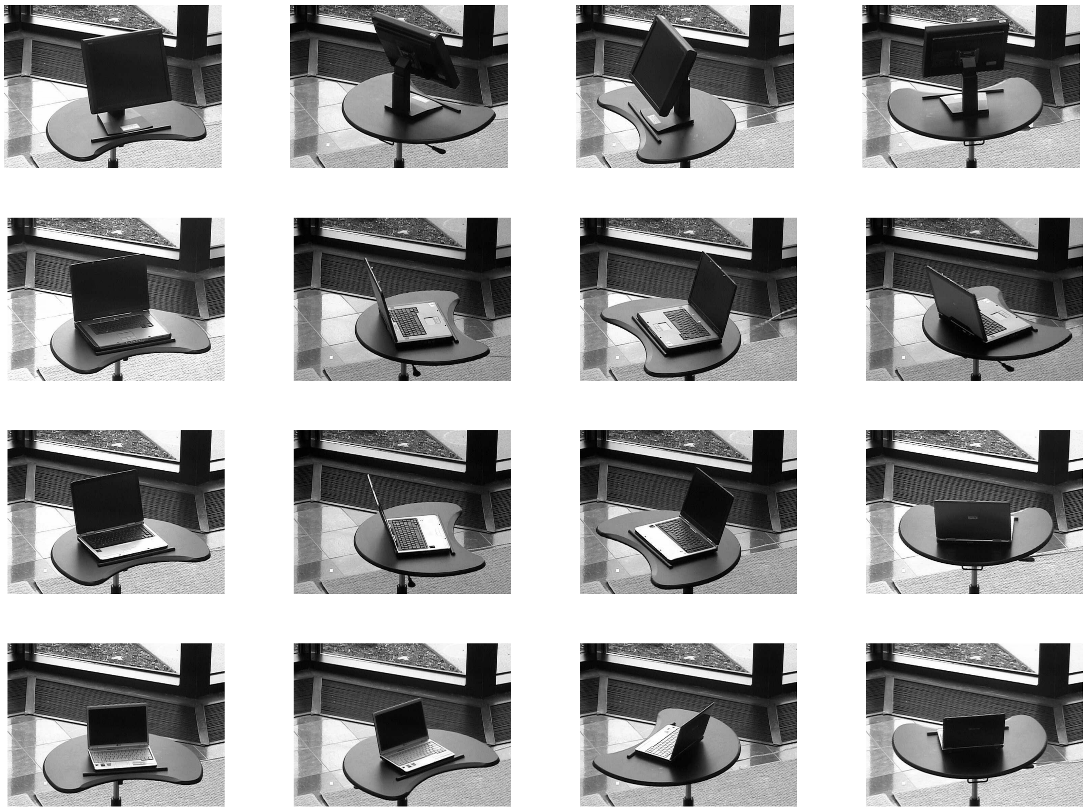


Fig. 2. Sample images from the image sets. The objects rotate on the table, giving the camera different capture angles. Pitch remained constant while yaw changed with the rotation.

for density estimation. While this may cause a minor change in performance, we aim to keep as many factors constant as possible for a fair comparison. Additionally, given the low number of samples we are considering, a GMM offers very little difference to a KDE.

It is clear that the proposed method (FINE) outperforms the KL method. To ease concerns that the performance gain is strictly due to the geodesic distance approximation $G(p, p_i; \mathcal{P})$, which may not be practically available in all cases, we also illustrate classification performance using the strict Hellinger distance $D_H(p, p_i)$. There is a slight decrease in performance, which shows that there is indeed some gain from the geodesic approach, but performance is still far superior to that of the KL-divergence. We believe this is due to the instability of the KL measure, which is highlighted when dealing with low sample size. As the sample size increases, and the PDFs are better estimated, we believe both methods would perform comparably.

Finally, we illustrate the embedding obtained with FINE of the data. For this case we used $l = 7$, such that each test image set had roughly 70% the number of sample images as the training sets. The embedding results are shown in Fig. 4, and the natural clustering is visually identified. Each point represents a unique image set $\boldsymbol{I}$, and the points corresponding to training sets are denoted with the symbol o. Note that this embedding was entirely unsupervised. This visualization, which is entirely based on the natural information-geometry between the image sets, is useful for comparing objects. One may notice that two of the laptop image sets are similarly embedded, while the other two are clearly separated. It is logical that the points corresponding to the LCD monitor lie furthest away from the points representing laptop image sets, as they are the most dissimilar. We can not visually decipher the reason 2 laptops seem so close, but note that they are still distinguishable even in 2 dimensions.

6. CONCLUSIONS

In this work we attempted to use an information-geometric approach to the problem of object recognition. By modeling each image set as a realization of some PDF on a statistical manifold, we were able to approximate the Fisher information distance between PDFs through the use of the Hellinger distance. The classification task was then performed using a 1-NN classifier between the test PDFs and the training PDFs. This formulation was shown to offer promising results for the

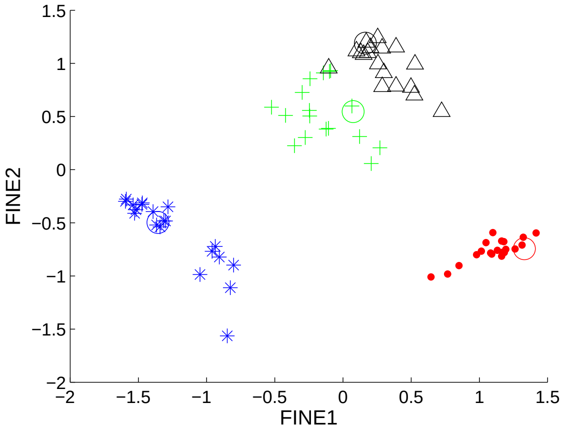


Fig. 4. Embedding of the image sets with FINE. We can see that two of the laptops (△ and +) are very similar, while the third laptop (⋆) and LCD monitor (·) are clearly separable.

object recognition task. Using Fisher information nonparametric embedding, we were able to reconstruct the statistical manifold in an low-dimensional Euclidean space. This enables visualization for 'human in the loop' analysis, as well as the ability to use learning methods which operate in Euclidean space (e.g. linear classifier) which have no straightforward connection to PDF representations.

We compare our methods to those found in [1], which is a similar framework focusing on the Kullback-Leibler divergence. We have shown that due to the low sample size, this measure is unstable and yields poor recognition performance. When more samples are available, the KL divergence becomes a more stable measure and is much more useful, as was illustrated in [1]. We have also shown that the geodesic approximation of the Fisher information distance, through the use of the Hellinger distance, yields improved performance to that of the strict Hellinger distance. We once again stress that this work is meant as a *proof-of-concept*, and we simply illustrate an example where we cover the shortcomings of a leading algorithm.

In future work, we wish to extend our methods towards the problem of automated face recognition. Additionally, we plan to continue the work on object recognition and obtain more thorough data sets and compare to several leading methods.

7. SPECIAL THANKS

This work was partially supported by the AFRL ATR Center through a summer internship of the second author and a Signal Innovations subcontract to the University of Michigan.

8. REFERENCES

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proceedings IEEE Conf. On Computer Vision and Pattern Recognition*, June 2005, pp. 581–588.

[2] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Fine: Fisher information non-parametric embedding," *IEEE Transactions on Pattern Analyis and Machine Intelligence*, 2008, submitted.

[3] K. M. Carter, R. Raich, and A. O. Hero, "Fine: Information embedding for document classification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, April 2008, pp. 1861–1864.

[4] W. G. Finn, K. M. Carter, R. Raich, and A. O. Hero, "Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: Treating flow cytometry data as high-dimensional objects," *Cytometry Part B: Clinical Cytometry*, 2008, in press.

[5] R. Kass and P. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley Series in Probability and Statistics. John Wiley and Sons, NY, USA, 1997.

[6] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191, American Mathematical Society and Oxford University Press, 2000, Translations of mathematical monographs.

[7] S.K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 917–929, June 2006.

[8] K. M. Carter, R. Raich, and A. O. Hero, "An information geometric framework for dimensionality reduction," Tech. Rep., University of Michigan, 2008, arXiv:0809.4866.

[9] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[10] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, London, 1994.

[11] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems, Volume 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002.