Fisher Information Embedding for Video Indexing and Retrieval

Xu Chen and Alfred O. Hero
University of Michigan at Ann Arbor, Ann Arbor, MI, USA;

ABSTRACT

In this paper, we present a novel information embedding based approach for video indexing and retrieval. The high dimensionality for video sequences still poses a major challenge of video indexing and retrieval. Different from the traditional dimensionality reduction techniques such as Principal Component Analysis (PCA), we embed the video data into a low dimensional statistical manifold obtained by applying manifold learning techniques to the information geometry of video feature probability distributions (PDF). We estimate the PDF of the video features using histogram estimation and Gaussian mixture models (GMM), respectively. By calculating the similarities between the embedded trajectories, we demonstrate that the proposed approach outperforms traditional approaches to video indexing and retrieval with real world data.

Keywords: Manifold Learning, Dimensionality Reduction, Video Retrieval, Laplacian Eigenmaps

1. INTRODUCTION

In recent years, the rapid development of multimedia technology and popular online multimedia archives have motivated much research in video indexing and retrieval[1].[2] Reliable video indexing requires effective measures of similarity that are sensitive to differences in content and insensitive to variations of content-irrelevant parameters, such as view angle, range, and illumination. For example, view invariance is one of the most important issues in video indexing and retrieval due to camera motion. Previous work in view invariant video indexing rely on different view invariant methods such as null space invariants.[1] The indexing and retrieval problem becomes further complicated by the high dimensionality of the video sequences. Traditional dimensionality reduction techniques such as Principal Component Analysis (PCA) have also been employed.[3]

In this paper, we propose a novel framework that uses a recently developed non-linear dimensionality reduced technique called Fisher Information non-linear embedding (FINE)[4][5].[6] FINE inherits invariance properties from the probability distributions of the features, which here are the grey level pixels of the image. Since the probability density function is non-negative and sums to one, a direct embedding into Euclidean space is not justified. Rather, as explained in,[4] it is more appropriate to embed the feature vector probabilities into an information geometry, which is a hypersphere with respect to Hellinger metric. In[4] it is shown that this embedding is a good approximation to the optimal Fisher embedding of a smooth parametric family of densities. To the best of our knowledge, this is the first time that the Fisher information embedding has been applied to video indexing and retrieval. The main advantage of FINE embedding is its simplicity: substantial information can be extracted from low level features.

The rest of the paper is organized as follows: we first briefly introduce the framework of FINE manifold learning in Section 2. We proposed the novel video indexing and retrieval framework in Section 3. Subsequently, we discuss the process of feature extraction and estimation of the PDF of the video sequences in Section 4. In Section 5, we compare the performance to a PCA approach to traditional approaches in video indexing and retrieval. Finally, we give a brief summary of our results in Section 6.

Further author information: (Send correspondence to Xu Chen)
Xu Chen: E-mail: xhen@umich.edu
Alfred O. Hero: E-mail: hero@eecs.umich.edu

2. FINE MANIFOLD LEARNING

In this section, we give a brief review of the methods for Fisher information non-linear embedding.[4] Let M be a family of PDFs parameterized by $\theta = [\theta^1, \dots, \theta^n]$:

$$M = \{p(x|\theta)|\theta \in \Theta \subseteq R^n\} \ . \tag{1}$$

M is a statistical manifold when the Fisher information metric is used to measure the amount of information the random variable X contains with respect to the unknown parameter θ. We define the Fisher information matrix $[I(\theta)]$ with elements

$$I_{ij} = -E[\frac{\partial}{\partial \theta^i} log f(X;\theta) \frac{\partial}{\partial \theta^j} log f(X;\theta)] \tag{2}$$

The Fisher information distance between two distributions can be approximated by Hellinger distance:[4]

$$D_H(p,q) = \sqrt{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx} \ . \tag{3}$$

FINE embeds of the distances between PDFs in M into a lower dimensional Euclidean space using Laplacian Eigenmaps.[7]

3. INDEXING AND RETRIEVAL ALGORITHM

We propose the following algorithm for video indexing and retrieval:

1. Extract features: We first construct our visual codebook using Bag-of-features (Bof)[89] by randomly splitting the dataset into 2 parts, 50% for training and 50% for testing where cross-validation is conducted. By clustering the SIFT features with K-means, we select the size of our visual codebook to be 120. Therefore, for each video frame, we obtain a 120 by 1 SIFT feature based histogram by mapping SIFT features from this video frame into the visual codebook.

2. Estimate the joint feature distribution $P_i, i = 1, 2, \dots, N$ of each frame of the video with histogram estimation or Gaussian mixture models. Gaussian mixture models are implemented using the MIXMOD software[10] with a BIC model order estimator.

3. Use FINE to embed the P_i into low dimensional Euclidean space. Given the PDFs $P = \{p_1, p_2, \dots, p_N\}$, the desired embedding dimension d is selected, e.g. for visualization of videos as trajectories $d = 3$. Subsequently, we calculate the matrix of distances D with elements $D(i,j) = \hat{D}_H(p_i, p_j)$. As the example in FINE,[4] we use Laplacian Eigenmaps[7] but here it is applied to a time indexed sequence of 2 dimensional spaces to obtain a trajectory for each video sequence.

4. Quantify the similarity between a pair of videos by computing a distance between the associated curves. Different distance measure can be used such as chamfer distance or Frechet distance.

4. ESTIMATION OF PROBABILITY DENSITY FUNCTIONS (PDFS)

We investigated two approaches for the estimation of the PDF over the video frames: histogram estimation and Gaussian mixture models. For lack of space we focus on Gaussian mixture models. The GMM has the form:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \ , \tag{4}$$

where π_k represents the weight of the kth gaussian components, μ_k and Σ_k represent the mean and variance of the kth gaussian components. We first extract the feature from each video frame using image patches. For example, for a video frame with the dimension 300 by 300, we can define 3 by 3 image patch features to obtain

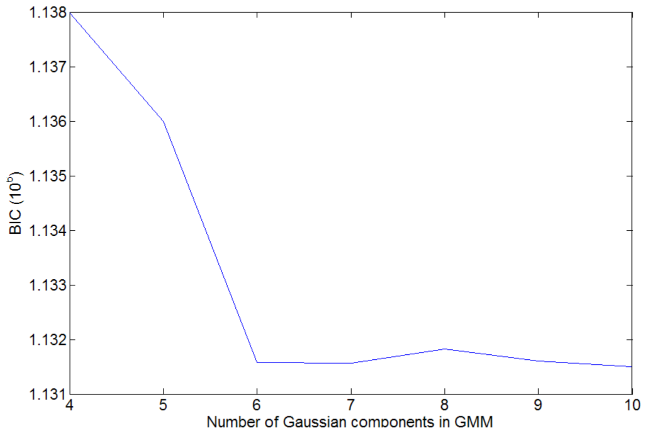

Figure 1. Plot of the BIC values versus different number of Gaussian components.

a 9 dimensional feature vector. In this case, there will be approximately 10^4 features in one video frame. The Gaussian mixture model (GMM) provides a more parsimonious and compact representation than the histogram when the PDFs of the video frames consist only a few modes.[2] By careful selection of the number of Gaussian components, one can obtain an accurate model for the feature PDF with fewer parameters than a histogram. There exist efficient algorithms, e.g. the MIXMOD algorithm[10] to implement the EM algorithm for estimating the weight, mean and variance (π_k, μ_k, σ_k) of each Gaussian components and to construct the PDFs for the features over each frame.

The number of Gaussian components in a Gaussian mixture model can be determined using a criterion to select model order such as Bayesian information criterion (BIC), Akaike information criterion (AIC).[11] Here we use BIC:

$$BIC = -2\ln(L) + k\ln(n) \ , \tag{5}$$

where L is the maximized value of the likelihood function for the estimated model, k is the number of free parameters used by the likelihood functions and n is the number of observations (the sample size). Given different candidate model orders k, the one with the lowest value of BIC is the one to be preferred.

5. SIMULATION RESULTS

Simulations were performed on the Context Aware Vision Image-based Active Recognition (CAVIAR) database[12] in order to illustrate FINE indexing and retrieval. We selected 300 video sequences from the CAVIAR dataset where each sequence consists of 100 frames. The videos have different content, such as, shopping, chasing or talking. To test the robustness of the algorithm, we create five noisy video sequences by adding different levels of salt and pepper noise. The average noise intensity ranges from 0.01 to 0.05.

In Fig.1, we plot the BIC value versus the different number of Gaussian components. As can be seen from Fig.1, when the number of Gaussian components is equal to 6, the BIC achieves its minimum value. Figs. 2, 3 illustrate the video sequences "Shopping" and "Leaving" for the video sequence "Shopping" from the CAVIAR dataset. In Fig. 4, the embedded trajectories from different video sequences are shown. We compute the cosine distances in the full dimensional space and Euclidean distances between the embedded trajectories from the video sequence "Shopping", "Leaving" and their noisy versions in Table 1. As can be seen from Table 1, the cosine distances between the embedded trajectories from the same type of video sequences are much closer than the those from different types of video sequences. To further investigate the FINE-based indexing and retrieval, we show the three dimensional embedding with the proposed method in Fig.5, where the results indicate the FINE embedding finds the natural separation of the video sequences in different classes by embedding them into lower dimensional Euclidean spaces. With the proposed method, we are able to determine the closest two frames in different video sequences in information geometrical representation. In Fig. 6, we demonstrate the two closest

Figure 2. First 10 video frames from the video segments in CAVIAR dataset for the motion event ”Shopping”.

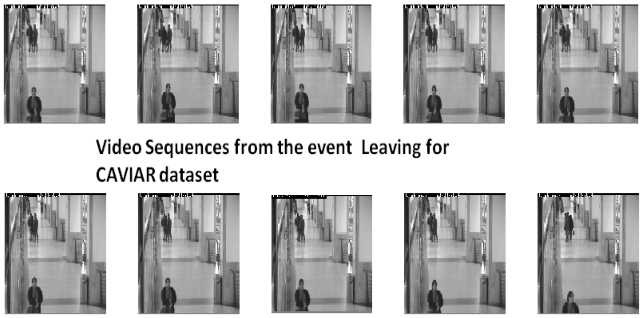

Figure 3. First 10 video frames of the video segment from CAVIAR dataset for the motion event ”Leaving”.

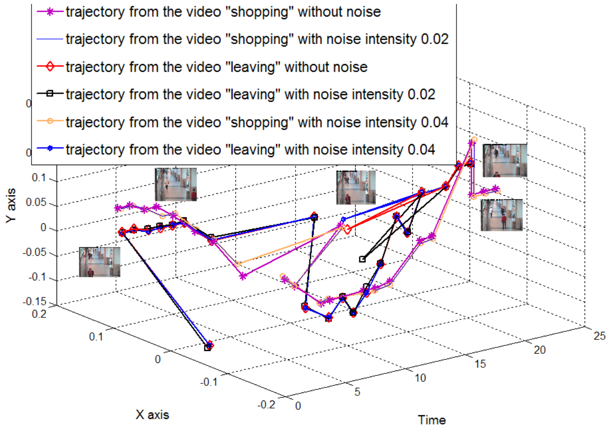

Figure 4. The six embedded trajectories from the two classes of video sequences ”shopping” and ”leaving” with various degrees of salt and pepper noise.

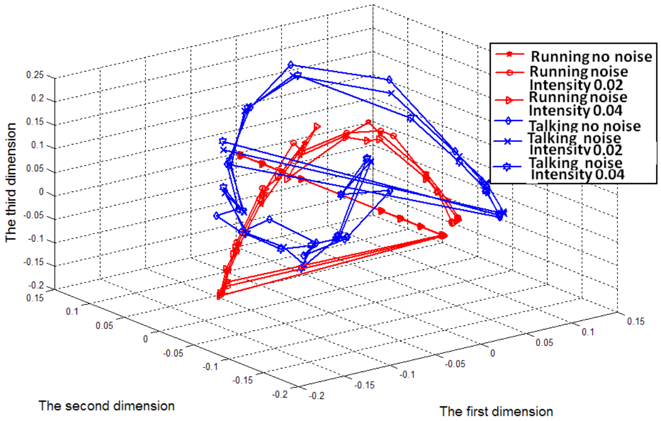

Figure 5. The six embedded trajectories from the two classes of the video sequences ”running” and ”talking” with various degrees of salt and pepper noise using FINE in 3 dimensions for PDF estimation.

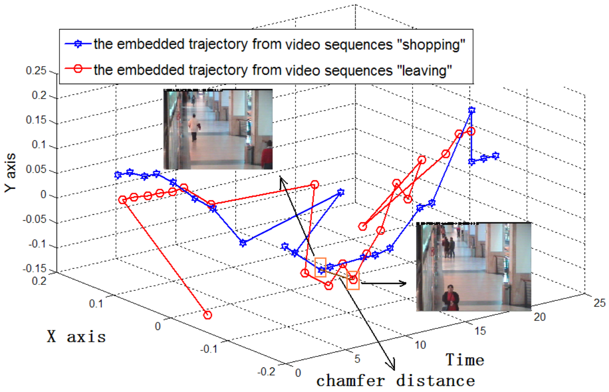

Figure 6. The chamfer distance between two embedded trajectories for the video sequences ”shopping” and ”leaving” determines the two most similar video frames in these two video sequences.

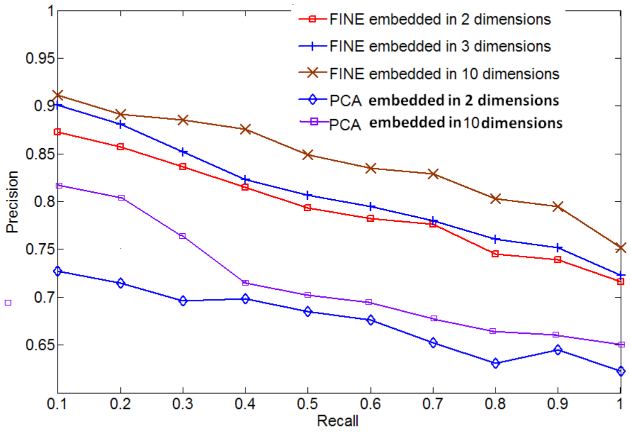

Figure 7. The precision and recall curves with Fisher information embedding into different dimensions and Principal component analysis (PCA) for video retrieval.

Table 1. Comparison of the cosine distances in full dimensional space and Euclidean distance in lower dimensional space between different embedded trajectories for CAVIAR videos. ”Shopping (A)” represents the embedded trajectory from the video sequences ”shopping” without noise. ”Shopping(B)” and ”Shopping(C)” represent the embedded trajectories from the video sequences ”shopping” with the noise intensity 0.02 and 0.04 respectively. Similar annotation is used for video sequences ”Leaving”.

Cosine\Euclidean Distance	Shopping(A)	Shopping(B)	Shopping(C)	Leaving(A)	Leaving(B)	Leaving(C)
Shopping(A)	0\0	0.39\0.037	0.50\0.037	1.33\0.53	1.32\0.51	1.29\0.50
Shopping(B)	0.39\0.037	0\0	0.31\0.0007	1.33\0.54	1.34\0.52	1.35\0.516
Shopping(C)	0.50\0.037	0.31\0.0007	0\0	1.35\0.54	1.30\0.52	1.36\0.516
Leaving(A)	1.33\0.53	1.33\0.54	1.35\0.54	0\0	0.31\0.058	0.42\0.076
Leaving(B)	1.32\0.51	1.34\0.52	1.30\0.52	0.31\0.058	0\0	0.37\0.019
Leaving(C)	1.29\0.50	1.35\0.516	1.36\0.076	0.42\0.076	0.37\0.019	0\0

frames in two video sequences ”Shopping” and ”Leaving” where we can identify that both of them correspond to the similar motions that people are leaving the shops. We compare the retrieval performance of our method with a canonical Euclidean dimensionality reduction approach based on principal component analysis (PCA) on the same dataset to plot precision and recall curves in Fig. 7. We apply PCA to the distance matrix and reduce it to the same two dimension as in FINE. It can be seen from Fig. 7 that the performance of our approach outperforms this simple application of PCA. This can be attributed to the fact that PCA is a linear Euclidean method that is not well justified for dimensionality reduction in information geometries.

6. CONCLUSION

We have presented a novel dimensionality reduction framework using Fisher information embedding. This framework is based on statistical manifold learning and provides a powerful tool for dimensionality reduction and video indexing and retrieval when there is no explicit Euclidean representation for the distances between videos. The method requires computation of the feature PDFs for frames of the video sequences. By embedding the Fisher information into a Euclidean space, we obtain a simple visual comparison between the embedded trajectories. The simulations demonstrated the superiority of the proposed approach compared to PCA. Future work will consider more sophisticated image features such as Gabor wavelets and a larger population of videos.

7. ACKNOWLEDGEMENT

The research in this paper was supported in part by ARO Grant W911NF-09-1-0310.

REFERENCES

[1] Chen, X., , Schonfeld, D., and Khokhar, A., “Robust null space representation and sampling for view-invariant motion trajectory analysis,” *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[2] Zhou, X., Zhuang, X., Yan, S., Chang, S., Johnson, M., and T.Huang, “Sift-bag kernel for video event analysis,” *ACM conference on Multimedia* (2008).

[3] Bashir, F., Schonfeld, D., and Khokhar, A., “Real-time motion trajectory-based indexing and retrieval of video sequences,” *IEEE Transactions on Multimedia* (2009).

[4] Carter, K., Raich, R., and Hero, A., “Fine: Information embedding for document classification,” *IEEE Intl Conf. on Acoustics, Speech and Signal Processing* (2008).

[5] Carter, K., Raich, R., and O.Hero, A., “Spherical laplacian information maps (slim) for dimensionality reduction,” *IEEE Workshop on Statistical Signal Processing (SSAP)* (2009).

[6] Sricharan, K., Raich, R., and Hero, A. O., “Manifolds, sparsity, and structured models: When can low-dimensional geometry really help?,” *NIPS Workshop* (2009).

[7] Belkin, M. and Niyogi, P., “Laplacian eigenmaps and spectral techniques for embedding and clustering,” *Advances in Neural Information Processing System* (2002).

[8] Fei-Fei, L. and Perona, P., "A bayesian hierarchical model for learning natural scene categories," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005).

[9] Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W., "Discovering objects and their location in images," IEEE International Conference on Computer Vision (ICCV) (2005).

[10] Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F., "Model-based cluster and disciminant analysis with the mixmod software," *Computational Statistics and Data Analysis* (2006).

[11] Lanterman, A., "Schwarz, wallace, and rissanen: Intertwining themes in theories of model selection," *International Statistical Review* (2001).

[12] "http://homepages.inf.ed.ac.uk/rbf/caviardata1/," *CAVIAR dataset* (2003).