LEARNING A PERCEPTUAL MANIFOLD FOR IMAGE SET CLASSIFICATION

Sriram Kumar, Andreas Savakis

Rochester Institute of Technology, Rochester, New York 14623, USA
{sk3100, andreas.savakis}@rit.edu

ABSTRACT

We present a biologically motivated manifold learning framework for image set classification inspired by Independent Component Analysis for Grassmann manifolds. A Grassmann manifold is a collection of linear subspaces, such that each subspace is mapped on a single point on the manifold. We propose constructing Grassmann subspaces using Independent Component Analysis for robustness and improved class separation. The independent components capture spatially local information similar to Gabor-like filters within each subspace resulting in better classification accuracy. We further utilize linear discriminant analysis or sparse representation classification on the Grassmann manifold to achieve robust classification performance. We demonstrate the efficacy of our approach for image set classification on face and object recognition datasets.

Index Terms— Image Set Classification, Independent Component Analysis, Grassmann Manifold, Face Recognition, Object Recognition

1. INTRODUCTION

Image set classification has attracted a lot of attention in the past decade [1-10] and finds applications in biometric authentication, surveillance and security, human computer interaction, etc. In image set classification, each set consists of a number of images that belong to the same class and can capture natural variations in the object's appearance, e.g. due to pose changes and varying illumination conditions. Compared to typical classification based on a single image, image set classification provides a richer representation of the object of interest and is well-suited for analysis of video and multi-camera data. The challenge is how to represent image sets due to the large variations displayed by images within the same class. Methods for image set representation include modeling with distributions [1], subspace learning [2, 3, 10], and dictionary learning [4].

Subspace learning methodologies, including manifold learning, suggest that high dimensional visual data can be efficiently represented using low dimensional subspaces [11]. These representations offer both efficiency due to the reduced dimensionality and improved classification accuracy due to better organization of the data in the low dimensional subspace.

Learning on the Grassmann manifold has received attention because it is efficient, discriminative and furthermore it can accommodate image set classification [7,12,13,14]. Grassmann geometry represents linear subspaces as points on the Grassmann manifold. Principal Component Analysis (PCA) is typically used to obtain sets of basis functions that represent the subspaces for an image set. However, principal components are based on global information and are limited in terms of capturing the local structure of the data within each subspace.

In this paper, we introduce a biologically inspired framework for perceptual subspace learning based on Independent Component Analysis (ICA) for Grassmann manifold construction to promote class discrimination and obtain natural means for image set comparison. The resulting **GRA**ssmann **I**CA **L**earning (**GRAIL**) approach is more robust and discriminative compared to standard Grassmann learning using PCA. The independent components [15,16] effectively capture local image characteristics in contrast to the principal components that only capture global information. Thus, the independent components provide natural discrimination capability that plays a key role in improving classification. Following this introduction, Section 2 discusses related work, Section 3 overviews the GRAIL framework, Section 4 presents results that demonstrate the effectiveness of GRAIL for face and object recognition, and Section 5 concludes the paper.

2. RELATED WORK

Image sets may be modeled as distributions [1], or alternatively, they may be modeled as subspaces [2, 3, 10], so that the distance between subspaces indicates the distance between image sets. Grassmann manifolds efficiently model image sets using subspace learning that promotes excellent class discrimination [7,12,13,14].

2.1. Grassmann Manifolds

The collection of m dimensional linear subspaces of $\mathbb{R}^D$, denoted as $\mathbb{G}(D, m)$, constitutes the Grassmann manifold. Each subspace is represented by a set of principal components obtained from images belonging to the same class. Fig. 1 shows two subspaces representing classes in image space and their mapping on the Grassmann manifold.

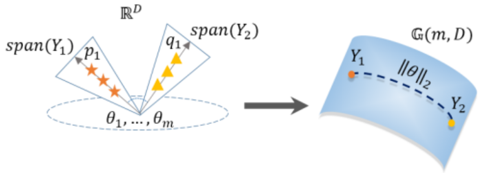


Fig 1. A conceptual illustration of the Grassmann manifold. Two image sets are described by linear subspaces in $\mathbb{R}^{\boldsymbol{D}}$. On the right, the subspaces represented by the span of their principal components are mapped as points on the Grassmann manifold.

Computing distances on the Grassmann manifold is based on geodesics instead of Euclidian distance. Geodesics can be computed using principal angles $\Theta = [\theta_1, \theta_2, \dots, \theta_m]$, where $\theta_{i=1:m} \in \left[0, {}^{\pi}/_{2}\right]$. For example, the projection metric is a measure of distance between two points on the manifold and is defined as follows.

$$d_{\text{Proj}}(Y_1, Y_2) = \left(m - \sum_{i=1}^{m} \cos^2 \theta_i\right)^{1/2} \qquad (1)$$

Although geodesics are easy to compute, they are sensitive to noise and image variability. To overcome the limitations of geodesics, kernelization is used to map in a Hilbert space where Euclidian distances are computed. There are various kernels to induce isometric embedding that map the points on the Grassmann manifold to Hilbert spaces. The projection kernel associated with this embedding is given by

$$\Phi_{\text{Proj}}(Y_1, Y_2) = tr[(Y_1 Y_1^T)(Y_2 Y_2^T)] \qquad (2)$$

The advantage of kernelization is that Euclidian distances computed in kernelized Hilbert space can be used for any type of classification.

2.2. Independent Component Analysis

Independent Component Analysis (ICA) is a generalization of PCA that is sensitive to higher order statistics. We consider the signal x which represents observations and can be written as a linear combination of the mixing matrix A and source signals s:

$$x = As \qquad (3)$$

where s denotes the unknown source signals, assumed to be independent, and A is a mixing matrix that is invertible. The ICA algorithm tries to find A, or equivalently the separating matrix W, based on the above assumptions and knowledge of observations x as follows,

$$U = Wx = W(As) \qquad (4)$$

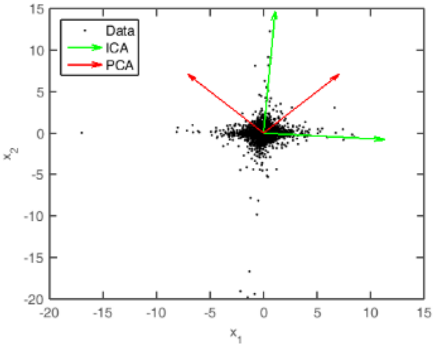


Fig 2. Visualization of the principal and independent components extracted from a data that not Gaussian distributed.

Fig. 2 shows an example where the ICA correctly identifies the components in contrast to PCA when the data is not Gaussian distributed. To speed up the process of ICA learning, the observations are preprocessed by whitening transformation. This essentially removes correlation and reduces the dimensionality. The whitening transformation matrix is obtained by performing the eigenvalue decomposition on the covariance matrix. It has been shown that in addition to removing correlation among components, the whitening transformation helps the convergence of the ICA algorithm and also improves performance in applications. The redundancy is reduced in the whitened data when compared with the raw data. After performing PCA, only the top eigenvectors were retained such that they account for 99% of the variance of that class.

To compute the independent components, we use the InfoMax algorithm proposed by Bell & Sejnowski [17]. The algorithm is based on the principle of optimal information transfer in neurons with sigmoidal transfer function.

Following the architecture proposed in [16], the data matrix X consists of faces formed by a linear combination of independent basis images S and a mixing matrix A. Let V be the matrix of dimension $(D \times m)$ be the top m eigenvectors. The ICA basis images are obtained by multiplying the weights with V^{T} i.e., $W\,V^{\mathrm{T}}$.

2.3. Sparse Representation Classification

Sparse representation classification has been widely used since it was initially introduced for face recognition [18]. We utilize sparse representation classification in kernelized space using kernel embedding. Following [14], we construct the projection kernel matrix from the training subspaces: $\Phi_{train} \in \mathbb{R}^{N_{\text{train}} \times N_{\text{train}}}$, where N_{train} is the number of training subspaces, each one corresponding to a different class. This constitutes our dictionary. In a similar way we construct the projection kernel matrix from the test subspaces $\Phi_{\text{test}} \in \mathbb{R}^{N_{\text{train}} \times N_{\text{test}}}$, where N_{test} is the number of training subspaces. The classification approach selects the class that minimizes the reconstruction error.

$$\begin{aligned} a^* = \arg\min_{a} \|\Phi_{\text{train}} a - \Phi_{\text{test}}(i)\|_2^2 + \lambda\|a\|_1 \\ s.t. \Phi_{test} = \Phi_{\text{train}} a,\ i = (1, \dots, N_{\text{test}}) \end{aligned} \tag{5}$$

where a are the sparse coefficients. Under this setting, image set comparisons are possible, as each atom in our dictionary is a subspace corresponding to one class.

3. PERCEPTUAL GRASSMANN MANIFOLD

3.1. Biological Motivation Underlying the Learning Framework

The human visual system is confronted with high dimensional visual information, but has the ability to extract only perceptually pertinent features [19]. These features intrinsically lie on a low dimensional space. The biological inspiration underlying our learning framework is twofold: (i) the basis functions extracted by ICA resemble Gabor-like filters, which closely model the responses of V1 simple cells [17]. These filters are spatially localized and exhibit separate high-order dependencies. Moreover, the high order statistics contain image phase information. (ii) The manifold hypothesis states that high dimensional data resides in a low dimensional manifold embedded in a Euclidean space. Once the visual system extracts the features, they are embedded on a perceptual manifold that characterizes the intrinsic dimensionality of the data [20]. We model this process by utilizing Independent Component Analysis.

3.2. Grassmann Manifold Learning with Independent Component Analysis

The Grassmann structure (see Fig. 1) offers an efficient framework for modeling and comparing image sets. An important step in generating the Grassmann manifold is the subspace construction. We propose GRAssmann Independent component analysis Learning (GRAIL), a robust and biologically inspired approach for constructing the Grassmann manifold, as described below.

The first step is image preprocessing, which involves cropping, scaling, centering and whitening the data. Then the images available in each class are used to form a subspace using Independent Component Analysis and each subspace is mapped to a point on the Grassmann manifold. In order to make sure that the components are orthonormal we performed Gram-Schmidt orthonormalization.

The process of kernelization using the projection kernel is used to map to a Hilbert space where classification is performed with Linear Discriminant Analysis (LDA) or Sparse Representations (SR) based on minimum reconstruction error. It is notable that this framework allows image set comparisons, such as when multiple images of an individual are available for use in face recognition.

In this paper, we utilize the projection kernel (see Eq. (2)) as it was shown to have the best performance for classification problems. In our experiments we utilize a discriminative analysis framework on the Grassmann manifold as proposed in [13]. Additionally, following the approach in [14], we incorporate sparse representation classification in kernelized Hilbert space.

4. EXPERIMENTAL RESULTS

We evaluate our proposed approach on multiple image sets datasets such as Extended Yale [21], LFW [22], and ETH-80 [23]. In our experiments each class is modeled by a linear subspace extracted from a collection of images in that class.

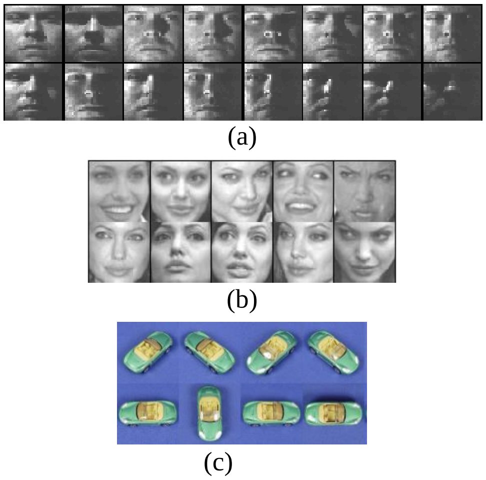


Fig 3. Examples of image sets in (a) Extended Yale, (b) LFW, and (c) ETH-80 datasets.

4.1. Datasets and Experimental Protocol

The Extended Yale Face Database B [21] consists of 38 subjects with 9 different poses and 45 different lighting conditions that vary significantly. Fig. 3(a) shows a subset of typical variations in an image set from this dataset.

The Labeled Faces in the Wild (LFW) [22] is challenging face dataset are a face database with 5749 individuals and 13,233 total face images collected from the web. The subjects vary by many parameters including pose, lighting, expression, etc. Fig. 3(b) shows a subset of typical variations in an image set from this dataset.

For the LFW dataset we only used subjects that had at least 20 images to make a valid image set comparison. Hence, we selected 62 subjects with a total of 3,024 images in total from the LFW dataset that met this criterion and performed image set comparison.

For the Extended Yale and LFW dataset, we performed two-fold cross validation. This ensures that there are enough samples per subject to construct the training and testing subspaces.

The ETH-80 [23] benchmark dataset for object recognition task has images of 8 object categories with each category including 10 different object instances. Each object instance has 41 images captured under different views, which form an image set. Fig. 3(c) shows a subset of typical variations in an image set from this dataset. For comparing our results with literature we followed the protocol given in [6], wherein we perform 10 trials in which the training image sets have five object instances from each category and remaining five object instances for testing. In each trial the object instances are randomly chosen.

All the images were resized into 32×32 intensity images. We extracted linear subspaces using ICA for each image set. In Fig. 4, we present sample ICA components extracted from the ETH-80 "cup" and "horse" image set.

Table 1: Image set classification results on Extended Yale

Method	Accuracy
k-NN [14]	94.16%
LDA [14]	92.98%
LPP [14]	99.14%
GDA [14]	100%
GRAIL + LDA (ours)	100%
GRAIL + SR (ours)	100%

Table 2: Image set classification results on LFW dataset

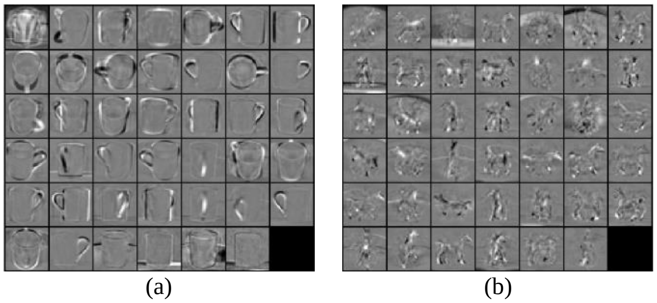
Features	Method	Accuracy
Linear Subspaces LTP features	LDA [14]	43.20%
	SR [14]	69.53%
	GDA [14]	58.87%
	GSR [14]	96.77%
Linear Subspaces Raw Image	MSM [24]	71.00%
	GRAIL + LDA (ours)	91.94%
	GRAIL + SR (ours)	89.52%

Table 3: Image set classification results on ETH-80 dataset

Method	Accuracy
MSM [24]	87.8%
DCC [6]	90.5%
MDA [6]	89.0%
GDA [6]	92.8%
COV + LDA [6]	94.5%
CDL [8]	92.5%
GRAIL + LDA (ours)	94.0%
GRAIL + SR (ours)	90.8%

4.2. Results

We compare the results our proposed approach with published results of various image set classification methods, such as Mutual Subspace Method (MSM), Discriminant Canonical Correlation analysis (DCC), Manifold Discriminant Analysis (MDA), Grassmann Discriminant Analysis (GDA), Covariance Discriminative Learning (CDL) and Grassmannian Sparse Representation (GSR). Tables 1 and 2 show face recognition rates on the Extended Yale and LFW face datasets respectively, and Table 3 shows the recognition rates on the ETH-80 dataset, all based on image set comparisons. For the Extended Yale dataset the proposed GRAIL approach outperforms the standard linear subspace methods and attains 100% classification, as does GDA. It is interesting to note that Grassmann approaches (with PCA/SVD or ICA) perform extremely well on datasets without using any handcrafted features, such as Histogram of Oriented Gradients (HOG) or Local Binary Patterns (LBP), while the results in [14] used Local Ternary Pattern (LTP) features.



(a) (b)

Fig 4. Sample visualizations of one instance of (a) "cup" and (b) "horse" image set in the ETH-80 dataset.

In the challenging LFW dataset the proposed approach performs well without any handcrafted features but does not outperform the results in [14] where LTP features are used. When only the normalized face images were used, the GDA approach (based on PCA) gave poor accuracy (61.29%) in contrast to GRAIL which obtained 91.94%. The subspace construction using PCA is limited to second order statistics and performs well when images in the image set are not appreciably different. However, the images in LFW vary in pose, background, color, saturation, and image quality. Hence, the GRAIL approach, which extracts independent components based on higher order statistics, exploits the nonlinearity of data and outperforms the PCA based Grassmann approach.

Finally, in the object recognition task GRAIL outperforms GDA. The independent components capture spatially local features which are essential for object recognition tasks. For image set classification, in both ETH-80 and LFW, GRAIL with LDA performs better than GRAIL with SR.

5. CONCLUSION

In this paper we propose Grassmann Independent component analysis Learning (GRAIL), a biologically inspired framework for robust and discriminative image set classification. The motivation behind using ICA over PCA for subspace learning was that ICA captures local image features and offers better discrimination among classes. We further incorporated sparse coding within the GRAIL framework by formulating the problem in kernelized Hilbert space using Grassmann kernels. This natural extension further adds robustness to the algorithm. Testing on standard datasets for face and object recognition demonstrates that in general GRAIL classifies image sets more effectively than PCA-based Grassmann methods.

6. REFERENCES

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, “Face recognition with image sets using manifold density divergence,” *Computer Vision and Pattern Recognition, CVPR*, 2005.

[2] R. Wang, S. Shan, X. Chen, and W. Gao, “Manifold-manifold distance with application to face recognition based on image set,” *Computer Vision and Pattern Recognition, CVPR,* 2008.

[3] M. Harandi, C. Sanderson, S. Shirazi, and B. Lovell, “Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching,” *Computer Vision and Pattern Recognition, CVPR,* 2011.

[4] Y. Hu, A. Mian, and R. Owens, “Face Recognition Using Sparse Approximated Nearest Points between Image Sets,” *IEEE Trans. Pattern Anal. Machine Intel., (PAMI),* 2012.

[5] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, “Image sets alignment for video-based face recognition,” *Computer Vision and Pattern Recognition, CVPR,* 2012.

[6] R. Wang, H. Guo, L. Davis, and Q. Dai, “Covariance discriminative learning: A natural and efficient approach to image set classification,” *Computer Vision and Pattern Recognition, CVPR,* 2012.

[7] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell, “Improved image set classification via joint sparse approximated nearest subspaces,” *Computer Vision and Pattern Recognition, CVPR,* 2013.

[8] J. Lu, G. Wang, and P. Moulin, “Image set classification using multiple order statistics features and localized multi-kernel metric learning,” *Int. Conf. Computer Vision, ICCV,* 2013.

[9] M. Hayat, M. Bennamoun, and S. An, “Learning non-linear reconstruction models for image set classification,” *Computer Vision and Pattern Recognition, CVPR,* 2014.

[10] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, “Multi-Manifold Deep Metric Learning for Image Set Classification,” *Computer Vision and Pattern Recognition, CVPR,* 2015.

[11] C. Deng, H. X. H. Yuxiao, H. Jiawe and T. Huang, “Learning a spatially smooth subspace for face recognition,” *Computer Vision and Pattern Recognition, CVPR,* 2007.

[12] P. Turaga, A. Veeraraghavan and R. Chellappa, “Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision,” *Computer Vision and Pattern Recognition, CVPR* ,2008.

[13] J. Hamm and D. D. Lee, “Grassmann discriminant analysis: a unifying view on subspace-based learning,” *Proc. 25th International Conference on Machine learning, ICML,* 2008.

[14] S. Azary and A. Savakis, “Grassmannian Sparse Representations,” *Journal of Electronic Imaging*, May 2015

[15] A. Hyvarinen, J. Karhunen, E. Oja, *Independent Component Analysis*, Wiley, New York, 2001

[16] M. Stewart Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face Recognition by Independent Component Analysis,” *IEEE Trans. Neural Networks,* 2002.

[17] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural computation,* vol. 7, no. 6, pp. 1129-1159, 1995.

[18] J. Y. Wright, A. Y., A. Ganesh, S. S. Sastry and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI),* vol. 31, no. 2, pp. 210-227, 2009.

[19] H. S. Seung and D. D. Lee, “The manifold ways of perception,” *Science*, 290(5500), pp. 2268-2269, 2000.

[20] J. D. Tenenbaum, “Mapping a manifold of perceptual observations,” *Advances in Neural Information Processing Systems, NIPS,* pp.682-688, 1998.

[21] K.-C. Lee, J. Ho and D. J. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” *IEEE Trans, Pattern Analysis and Machine Intelligence (PAMI),* vol. 27, no. 5, pp. 684-698, 2005.

[22] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments,” University of Massachusetts, Amherst, Technical Report 07-49, 2007

[23] Leibe, Bastian, and Bernt Schiele. “Analyzing appearance and contour based methods for object categorization,” *IEEE Int. Conf. Computer Vision and Pattern Recognition, CVPR,* 2003.

[24] Nishiyama, Masashi, Osamu Yamaguchi, and Kazuhiro Fukui. "Face recognition with the multiple constrained mutual subspace method." *Audio-and Video-Based Biometric Person Authentication.* Springer Berlin Heidelberg, 2005.