

Recognition of Humans and their Activities using Statistical analysis on Stiefel and Grassmann Manifolds

Pavan Turaga, Ashok Veeraraghavan and Rama Chellappa
 Center for Automation Research, University of Maryland, College Park
 {pturaga, vashok, rama}@umiacs.umd.edu

Many applications in computer vision involve learning and recognition of patterns from exemplars which lie on certain manifolds. Given a database of examples and a query, the following two questions are usually addressed – a) what is the ‘closest’ example to the query in the database ? b) what is the ‘most probable’ class to which the query belongs ? The answer to the first question involves study of the geometric properties of the manifold, which then leads to appropriate definitions of distance metrics on the manifold (geodesics etc). The answer to the second question involves statistical modeling of inter- and intra-class variations on the manifold. In this paper, we concern ourselves with two related manifolds that often appear in several vision applications – the *Stiefel* Manifold and the *Grassmann* Manifold. We describe statistical modeling and inference tools on these manifolds which result in significant improvements in performance over traditional distance-based classifiers. We illustrate applications to video-based face recognition and activity recognition.

I. INTRODUCTION

The Stiefel manifold is the space of k orthonormal vectors in R^m , represented by an $m \times k$ matrix Y , such that $Y^T Y = I_k$. The Grassmann manifold is the space of k dimensional *subspaces* in R^m and can be viewed as the orbit space of the Stiefel manifold (over full rank matrices). The study of these manifolds has important consequences for applications such as dynamic textures [5], human activity modeling and recognition [6], video based face recognition [1], shape analysis [3], where data naturally lies either on the Stiefel or the Grassmann manifold.

The Stiefel Manifold $V_{k,m}$ [2]: The Stiefel manifold $V_{k,m}$ is the space whose points are k -frames in R^m , where a set of k orthonormal vectors in R^m is called a k -frame in R^m ($k \leq m$). Each point on the Stiefel manifold $V_{k,m}$ can be represented as a $m \times k$ matrix X such that $X^T X = I_k$, where I_k is the $k \times k$ identity matrix.

The Grassmann Manifold $G_{k,m-k}$ [2]: The Grassmann manifold $G_{k,m-k}$ is the space whose points are k -planes or k -dimensional hyperplanes (containing the origin) in R^m . An equivalent definition of the Grassmann manifold is as follows. To each k -plane ν in $G_{k,m-k}$ corresponds a unique $m \times m$ orthogonal projection matrix P idempotent of rank k onto ν . If the columns of an $m \times k$ matrix Y spans ν , then, $Y Y^T = P$.

II. DISTANCE METRICS AND STATISTICAL MODELS

Procrustes Distance: Two representations of points on the Stiefel manifold can be defined [2].

- Representation V_a : A point X on $V_{k,m}$ is an $m \times k$ matrix such that $X^T X = I_k$.
- Representation V_b : A point X on $V_{k,m}$ is identified with an equivalence class of $m \times k$ matrices $X R$ in $R_{m,k}$, for $R > 0$.

This is also called the Procrustes representation of the Stiefel manifold.

The squared Procrustes distance for two given matrices X_1 and X_2 on the Stiefel manifold, is the smallest squared Euclidean distance between any pair of matrices in the corresponding equivalence classes (representation V_b). Hence,

$$d_{V_b}^2(X_1, X_2) = \min_{R>0} \text{tr}(X_1 - X_2 R)^T (X_1 - X_2 R) \quad (1)$$

$$= \min_{R>0} \text{tr}(R^T R - 2X_1^T X_2 R + I_k) \quad (2)$$

Thus, for the case where R varies over the space $R_{k,k}$ of all $k \times k$ matrices, the distance is given by $d_{V_b}^2(X_1, X_2) = \text{tr}(I_k - A^T A)$, where $A = X_1^T X_2$.

Kernel Density Estimator: Given several examples from a class (X_1, X_2, \dots, X_n) on the manifold $V_{k,m}$, the class conditional density can be estimated using an appropriate kernel function. For the Procrustes distance metric $d_{V_b}^2$ the density estimate is given by [2] as

$$\hat{f}(X; M) = \frac{1}{n} C(M) \sum_{i=1}^n K[M^{-1/2}(I_k - X_i^T X X^T X_i)M^{-1/2}] \quad (3)$$

where $K(T)$ is the kernel function, M is a $k \times k$ positive definite matrix which plays the role of the kernel width or a smoothing parameter. $C(M)$ is a normalizing factor chosen so that the estimated density integrates to unity. The matrix valued kernel function $K(T)$ can be chosen in several ways. We have used $K(T) = \exp(-\text{tr}(T))$ in all the experiments reported in this paper.

III. APPLICATIONS AND EXPERIMENTS

ARMA Modeling for Action recognition: A wide variety of time series data such as dynamic textures, human joint angle trajectories, shape sequences, video based face recognition etc are frequently modeled as autoregressive and moving average (ARMA) models [5], [6], [1]. The ARMA model equations are given by

$$f(t) = Cz(t) + w(t) \quad w(t) \sim N(0, R) \quad (4)$$

$$z(t+1) = Az(t) + v(t) \quad v(t) \sim N(0, Q) \quad (5)$$

where, z is the hidden state vector, A is the transition matrix and C is the measurement matrix. Given a sequence of observations $f(1), f(2), \dots, f(\tau)$, let $[f(1), f(2), \dots, f(\tau)] = U\Sigma V^T$ be the singular value decomposition of the data. Then, the estimates of the model parameters, (A, C) are $\hat{C} = U, \hat{A} = \Sigma V^T D_1 V (V^T D_2 V)^{-1} \Sigma^{-1}$, where $D_1 = [0 \ 0; I_{\tau-1} \ 0]$ and $D_2 = [I_{\tau-1} \ 0; 0 \ 0]$. For comparison of models, the most commonly used distance metric is based on subspace angles between column spaces of the observability matrices. Thus, a linear dynamical system can be alternately identified as a point on the Grassmann manifold corresponding to the column space of the observability matrix. We performed a recognition experiment on the publicly available INRIA dataset [7]. The dataset consists of 10 actors performing 11 actions, each action executed 3 times at varying rates while freely changing orientation. We used the view-invariant representation and features as proposed in [7]. The temporal evolution of features for an action is modeled using an ARMA model. In figure 1 (a), we show the recognition results obtained using four methods.

Video-Based Face Recognition: Video-based face recognition by modeling the ‘cropped video’ either as dynamical models (11) or as a collection of PCA subspaces [4] have recently gained popularity due to their ability to recognize faces from low resolution videos. The model parameters in both these instances (C matrix of the ARMA model or PCA subspace) are directly identifiable as points on the Grassmann Manifold. Therefore, both Procrustes distance and Kernel density methods are directly applicable to video-based face recognition. In our experiments, we model the temporal evolution of the ‘cropped video’ using an ARMA model. We tested our method on the dataset used in [1]. The dataset consists of face videos for 16 subjects with 2 sequences per subject. Subjects arbitrarily change head orientation and expressions. The illumination conditions differed widely for the 2 sequences of each subject. For each subject, one sequence was used as the gallery while the other formed the probe. The experiment was repeated by swapping the gallery and the probe data. The recognition results are reported in table 1 (b). For kernel density estimation, the available gallery sequence for each actor was split into three distinct sequences. As seen in the last column, the kernel-based method outperforms the other approaches.

Both these experiments demonstrate the strength of statistical modeling on the appropriate manifolds. Significant improvements in recognition accuracy are obtained in both cases.

| Activity | Best Dim. Red. [7] 64 ³ volume | Subspace Angles 16 ³ volume | NN-Pro-Stiefel 16 ³ volume | Kernel-Stiefel 16 ³ volume |
|--------------|--|---|--|--|
| Check Watch | 86.66 | 93.33 | 90 | 100 |
| Cross Arms | 100 | 100 | 96.67 | 100 |
| Scratch Head | 93.33 | 76.67 | 90 | 96.67 |
| Sit Down | 93.33 | 93.33 | 93.33 | 93.33 |
| Get Up | 93.33 | 86.67 | 80 | 96.67 |
| Turn Around | 96.67 | 100 | 100 | 100 |
| Walk | 100 | 100 | 100 | 100 |
| Wave Hand | 80 | 93.33 | 90 | 100 |
| Punch | 96.66 | 93.33 | 83.33 | 100 |
| Kick | 96.66 | 100 | 100 | 100 |
| Pick Up | 90 | 96.67 | 96.67 | 100 |
| Average | 93.33 | 93.93 | 92.72 | 98.78 |

(a)

| Test condition | System Distance | Procrustes | Kernel density |
|-----------------|-----------------|------------|----------------|
| Gallery1,Probe2 | 81.25 | 93.75 | 93.75 |
| Gallery2,Probe1 | 68.75 | 81.25 | 93.75 |
| Average | 75% | 87.5% | 93.75% |

(b)

Fig. 1. (a) Comparison of view invariant recognition of activities in the INRIA dataset using the dimensionality reduction methods of [7], ARMA subspace angles, Procrustes distance on the Stiefel manifold and Maximum likelihood using kernel density methods on the Stiefel manifold (b) Comparison of video based face recognition approaches ARMA model distance, Stiefel Procrustes distance, Manifold kernel density.

REFERENCES

- [1] G. Aggarwal, A. Roy-Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. *International Conference on Pattern Recognition*, 2004.
- [2] Y. Chikuse. *Statistics on special manifolds, Lecture Notes in Statistics*. Springer, New York., 2003.
- [3] C. R. Goodall and K. V. Mardia. Projective shape analysis. *Journal of Computational and Graphical Statistics*, 8(2), 1999.
- [4] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
- [5] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. *ICCV*, 2:439–446, 2001.
- [6] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Matching shape sequences in video with an application to human movement analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(12):1896–1909, Dec 2005.
- [7] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.