Flag Manifolds for the Characterization of Geometric Structure in Large Data Sets

T. Marrinan, J. R. Beveridge, B. Draper, M. Kirby, and C. Peterson

Colorado State University, Fort Collins, Colorado, USA
kirby@math.colostate.edu

Abstract. We propose a flag manifold representation as a framework for exposing geometric structure in a large data set. We illustrate the approach by building pose flags for pose identification in digital images of faces and action flags for action recognition in video sequences. These examples illustrate that the flag manifold has the potential to identify common features in noisy and complex datasets.

1 The Mathematical Challenges of Large Data Sets

Some very intriguing problems faced by scientists today have hints, suggestions, and solutions hidden within large collections of data. Mathematicians, computer scientists and statisticians have a fundamental role to play in developing the theory, tools, and algorithms needed by the general researcher in their quest to extract meaningful information from such data sets. While the type of data can vary drastically, one seeks a range of sufficiently robust tools that can be applied across multiple disciplines.

Finding ways to compactly represent a complicated object (such as a data cloud or a high dimensional array) has allowed for knowledge discovery within massive data sets (e.g. a data set consisting of many data clouds or many high dimensional arrays). As an example, suppose a data cloud in $\mathbb{R}^n$ clusters along a k-dimensional linear space then, for comparison against other data clouds, one could identify the cloud with this k-dimensional linear space. One could then associate a single point to the data by identifying the k-dimensional linear space with a point on an appropriate Grassmann manifold. Such a map transforms the problem of comparing data clouds in one setting to comparison of points on a Grassmann manifold.

Building on the theme of the previous paragraph, suppose one identifies a portion of the information in a collection of data with a nested sequence of vector spaces. This could be natural in settings involving an ordered sequence of data or as the result of a singular value decomposition. Examples might include the spectral sheets in a hyper-spectral digital image, the frames in a video stream, or the output of a singular value decomposition applied to a data set collected under a variation of state. A nested sequence of vector spaces is known as a *flag*. One could associate a single point to the data by identifying the nested sequence of vector spaces with a point on a flag manifold. Through this representation, comparisons of multiple instances of data can be transformed to comparisons of points on a flag manifold. For both the Grassmann and flag manifolds, there is a rich collection of metrics that can be considered for purposes of comparison.

One goal of the machine learning community is to create algorithms for automated processing and interpretation of the output obtained from a collection of sensors. For example, one may be interested in detecting anomalies in human behavior such as fall detection as an aid for assisted living [8]. Another example, is concerned with automated action recognition in video sequences towards the goal of automated video-to-text algorithms [6]. This paper develops an approach based on the geometry of the flag manifold for exploiting structure and correlations within large data sets.

The paper is structured as follows. Section 2 gives the mathematical background. Section 3 describes an algorithm for producing points on flag manifolds from collections of subspaces. Section 4 provides examples illustrating the flag approach. Section 5 consists of concluding remarks.

2 Grassmann, Stiefel and flag manifolds

A *flag* is a strictly ascending sequence of subspaces of a fixed n-dimensional vector space, V. Given a flag, $V_1 \subset V_2 \subset \cdots \subset V_r \subset V$, the *signature* of the flag is the data $(d_1, d_2, \ldots, d_r, n)$ where d_i denotes the dimension of V_i. The *flag manifold* $FL(d_1, d_2, \ldots, d_r; n)$ is a manifold whose points parameterize all flags with signature $(d_1, d_2, \ldots, d_r, n)$. A flag is called complete if its signature is $(1, 2, 3, \ldots, n-1, n)$. An ordered basis, $v_1, v_2, \ldots, v_n$ of V gives rise to a complete flag by setting V_i equal to the span of $v_1, v_2, \ldots, v_i$. A complete flag in $\mathbb{R}^n$ or $\mathbb{C}^n$ can be used to build an orthonormal basis (unique up to multiplication by a unit length scalar at each step). The general linear group acts transitively on the set of all complete flags in a fixed vector space. A Grassmann manifold $G(k, n)$ is a flag manifold with signature (k, n). The projective space $\mathbb{P}^{n-1}$ is the Grassmann manifold $G(1, n)$. Thus, flag manifolds generalize Grassmann manifolds which generalize projective space.

The *Stiefel Manifold*, $S(k, n)$, parametrizes orthonormal k-frames in a fixed n dimensional inner product space space V [10]. It can be viewed as a homogeneous space for the action of a matrix group. There is a natural projection, $F : S(k, n) \to G(k, n)$ where an orthonormal k-frame is sent to its span. The fiber of F over a point $P \in G(k, n)$ is the set of k-frames in the k-dimensional space determined by P. Similarly, there are projection maps from $S(k, n)$ to any flag manifold $FL(d_1, d_2, \ldots, d_r; n)$ with $d_r \leq k$.

If $V = \mathbb{R}^n$, then $S(1, n)$ corresponds to the unit hypersphere S^{n-1} and $S(2, n)$ corresponds to the unit tangent bundle on the unit hypersphere. In general, $S(k, n)$ can be identified with $O(n)/O(n-k)$, $S(n, n)$ corresponds to $O(n)$, and $S(n-1, n)$ corresponds to $SO(n)$. In a similar manner, $Gr(k, n)$ can be identified with $O(n)/O(k) \times O(n-k)$ and $FL(d_1, d_2, \ldots, d_r; n)$ can be identified with $O(n)/O(d_1) \times O(d_2 - d_1) \times \cdots \times O(d_r - d_{r-1}) \times O(n -$

d_r). Through these identifications, concrete descriptions of the tangent and normal bundles to Grassmann, flag and Stiefel manifolds are available [1,13].

3 Flag Manifolds from Data

There are many approaches for encoding and representing the structure in a data matrix as a point on a Grassmann or flag manifold. At a higher level, what kinds of statistical tools should be developed for the purpose of analyzing or representing a data cloud consisting of points on multiple Grassmann manifolds? An important early step is to find algorithms that produce single points on a flag manifold that represents common structure in such a data cloud. For additional details concerning special manifold statistics, see [9,12].

An Algorithm for Computing a Flag from a Collection of Subspaces. Let $[X]$ denote the column space of a matrix X. From a collection of subspaces $\mathcal{D} = \{[X_1], \ldots, [X_N]\}$ of an n-dimensional vector space V, we utilize an optimization algorithm to associate a point on a flag manifold to $\mathcal{D}$. A full description of the optimization algorithm and its properties is found in [5]. The algorithm finds an ordered collection of orthonormal vectors, u^j, by solving

$$\begin{aligned} [u^{(j)}] := \quad & \operatorname*{arg\,min}_{[u] \in Gr(1,n)} \quad \sum_{[X_i] \in \mathcal{D}} d_{pF}([u], [X_i])^2 \\ & \text{subject to} \quad [u^{(j)}] \perp [u^{(l)}] \qquad \text{for } l < j, \end{aligned} \tag{1}$$

where $d_{pF}([u], [X_i])$ is the projection Frobenius norm (which can be written in terms of the principal angles between the two subspaces $[u]$ and $[X_i]$). As illustrated by Björck and Golub, the vector of principal angles, Θ, between the subspaces $[X]$ and $[Y]$ can be found as the inverse sines of the singular values of the matrix $Q_X^T Q_Y$, where Q_X and Q_Y are unitary bases for $[X]$ and $[Y]$ respectively [3].

Solving Eq. 1 leads to the set $\{[u^{(1)}], [u^{(2)}], \ldots, [u^{(r)}]\}$ where r is the dimension of the span of the elements in $\mathcal{D}$. Recalling that $d_{pF}([X], [Y]) = \| \sin \Theta \|_2$, the sequence of optimizers can be found analytically as is shown in [5]. From these one-dimensional subspaces, the point $P \in FL(1, 2, \ldots, r, n)$ associated with $\mathcal{D}$ is then,

$$P = \left[u^{(1)} \right] \subset \left[u^{(1)} | u^{(2)} \right] \subset \ldots \subset \left[u^{(1)} | \ldots | u^{(r)} \right] \subset V. \tag{2}$$

Principal angles have been widely used for comparing points on Grassmannians. We propose using them to compare points on flag manifolds obtained from the approach above. If individual elements of a flag are of interest, the distance between subspaces can be measured using a variety of metrics, such as the geodesic distance based on arc length, i.e. $d([X], [Y]) = \|\Theta\|_2$. Similarly, we can define metrics between flags with the same signature by taking

Data Set	Sample Size	Number of Samples	Classes
Mind's Eye Tracklets	$32 \times 32 \times 48$	308 tracklets	'carry', 'gesture', 'leg-motion', 'loiter' 'loiter-group', 'turn', 'walk', 'walk-group'
Images of Letters	45×45	60 images	'a', 'f', 'w', 'x' each in 15 fonts
PIE faces	277×299	15288 images	56 subjects, 21 illuminations, 13 poses

Table 1: Descriptions of the three data sets used for experiments.

functions of the principal angles between each of their elements of the appropriate size, such as the sum of the geodesic distances between matching elements.

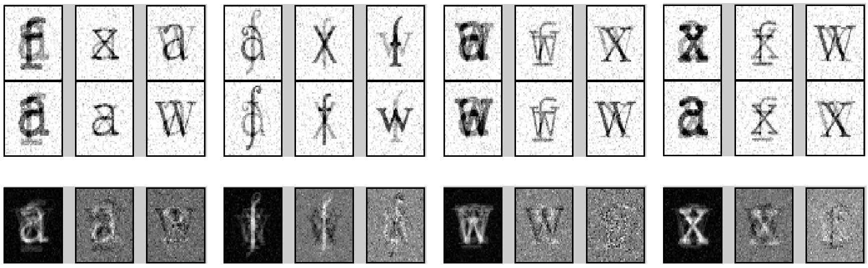
4 Numerical Experiments

The experiments in this paper are meant to illustrate the ability of the flag representation to organize data with multiple semantically meaningful forms of variation. To this end, we will explore three sets of image and video data. Each experiment will look to isolate one simple form of variation. Success in the task will be measured by percentage of test samples that correctly identified with a flag containing the intended information. A concise description of each data set can be seen in Table 1.

Illustrative example. The first data set consists of 60 images of the letters 'a', 'f', 'w', and 'x' depicted in 15 fonts. The flag representation will isolate a common letter in a set that contains other distracting letters and noise.

For each letter we randomly choose 6 images from our set of 60; 3 images of one letter and 1 image of each of the other 3 letters. The images are raster-scanned to create vectors in $\mathbb{R}^{2025}$. Three 2-dimensional subspaces of $\mathbb{R}^{2025}$ are produced from the 6 images. Each 2-dimensional subspace is formed as the span of two random linear combinations of an image of the main letter, an image of a different letter, and added Gaussian noise. Thus the 6 images in a set create 3 points on $\mathrm{Gr}(2, 2025)$ such that the main letter is a common feature in each Grassmann point. For each set of 3 Grassmann points, we create a flag that helps to expose the similarity between the points. Examples of the four sets and the first three vectors from their associated flags can be seen in Fig. 1.

The flags created from subspaces containing each letter can be used to identify instances of the same letter in fonts that were not used to create the flags. In fact, calculating the closest flag to a test sample did a better job of classifying novel instances of a letter than using a nearest neighbor classifier with the raw images that were used to train the flags. Over 10 trials, classification using the flags had an average success rate of $\approx 83\%$, while classification using the nearest image had a success rate of $\approx 65\%$.



(a) The letter ‘a’ (b) The letter ‘f’ (c) The letter ‘w’ (d) The letter ‘x’

Fig. 1: Four sets of images each with one common feature. Each image is the superposition of two letters and noise, and each column of images is used to create a point on $\mathrm{Gr}(2, 2025)$ that spans two letters. Below the sets are the first three images from the associated flag.

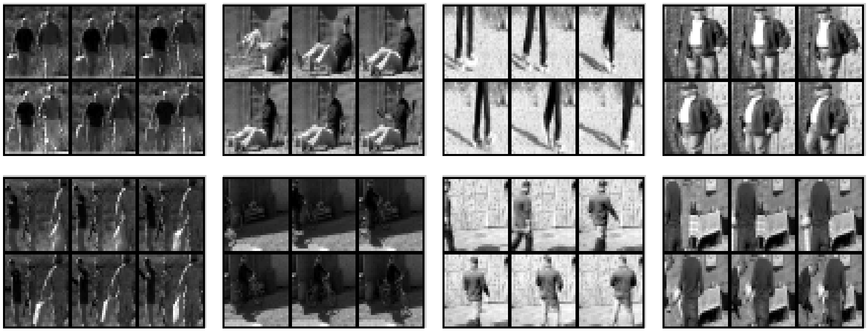


Fig. 2: Mind’s Eye video sequences illustrating six frames of eight doubly labeled actions. From left-to-right, top-to-bottom the labels are: ‘carry/walk’, ‘gesture/sit-up’, ‘leg-motion/walk’, ‘loiter/walk’, ‘loiter/group-patdown’, ‘turn/ride-bike’, ‘walk/turn’, and ‘walk-group/bend’.

Video Sequences. The second example uses portions of video clips, called *tracklets*, that were extracted from larger and longer videos filmed as part of DARPA’s Mind’s Eye program. The tracklets have been automatically cropped and registered to focus on an action of short duration (48 frames). The goal of this experiment is to automatically recognize a single action contained in a tracklet that depicts two actions being performed simultaneously. The tracklets have been hand labeled with the actions they contain. Examples of frames from some of these tracklets can be seen in Figure 2. Each video contains at least one of the labels listed in Table 1.

In order to demonstrate the flag’s ability to model the dominant form of variation in a set of subspaces, we compare classification accuracy using flags

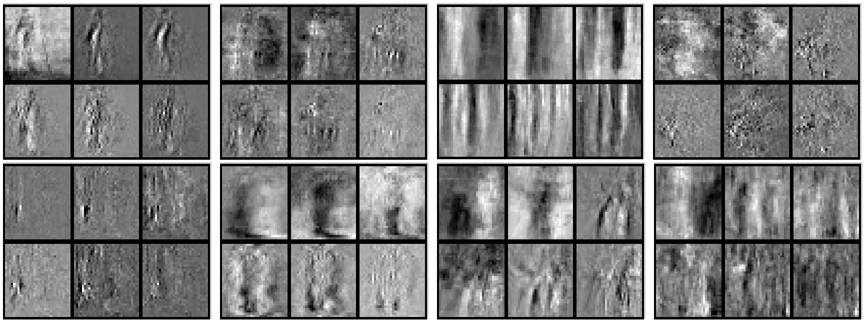


Fig. 3: Mind's Eye flags left-to-right, top-to-bottom: 'carry', 'gesture', 'leg-motion', 'loiter', 'turn', 'walk', 'walk-group'.

versus a nearest neighbor approach. To begin, each 3-way array is unfolded into a matrix of size 1024 × 48 with each column representing a frame of the video. The column space of each matrix is used to represent the tracklet. From the 308 available tracklets, we select K training samples from each action class. It is important to note that some of the training samples for one class may share their second label with another one of the classes being modeled. For example, a video labeled 'walk/turn' could be used as a training sample for either class 'walk' or 'turn', but not both.

Using the K training samples from each of the 8 classes, we create a flag for each class. The tracklets that are not used for training make up the test set. Each test video is compared to the 48-dimensional component in each flag. Using the geodesic distance based on arc length, a test sample is given the label of the nearest flag. The classification is considered a success if the label matches one of its given labels. Similarly, each test video is compared to the $8 \times K$ videos that were used to create the flags. In this case a test sample is given the label of the class that contained the nearest video. The results of this experiment can be seen in Fig. 4. The left graph in Fig. 4 shows the accuracy for the classification as the number of training samples, K, increases. Surprisingly, the accuracy is comparable for each method. One would expect superior performance using nearest neighbors given the amount of data required. The graph on the right side of Fig. 4 shows the precision rate for the individual classes. We can see that some classes do much better than others, which is partially a product of the number of samples available from each class. The resulting flags for these actions are shown in Fig. 3.

Pose flags. In the final example we create flags to represent poses from a subset of images in the CMU-PIE database [11]. The subset of images used is described in Table 1. The example focuses on the images that have ambient lighting turned on with neutral expressions. By grouping the remaining images in a structured way, we create a flag that represents a single pose.

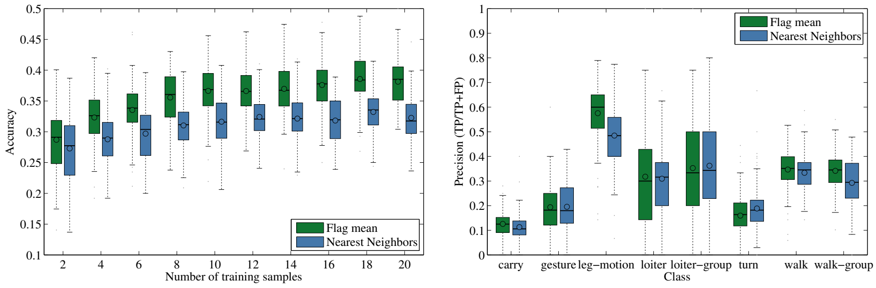
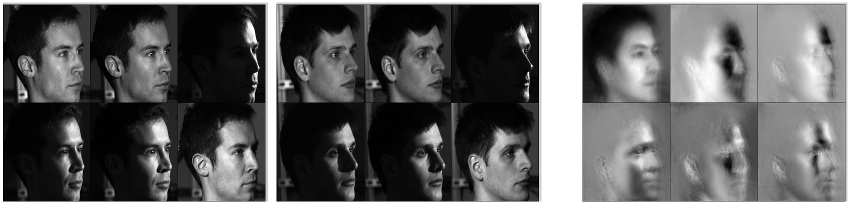


Fig. 4: Mind's Eye data classification: The figure on the left shows overall accuracy vs. number of training samples used per class. The figure on the right shows the precision for each class when 4 training samples were used.



(a) Two *data points* used to create the PS/I flag.

(b) The PS/I flag.

Fig. 5: An example of a PS/I-flag created from 34 sets of PIE images.

We find a basis for a collection of vectorized images that share a single pose, a single subject, and whose lighting conditions differ. The span of this basis approximates an illumination subspace. If we then create a flag out of subspaces that share a pose, but contain different subjects, we get a model for that pose that appears independent of subject or illumination. We refer to such a flag as a PS/I-flag to indicate the ordering of the variation. That is, the pose is consistent across all subspaces, the subject is consistent within each subspace, and the lighting varies within each subspace. An example of a PS/I-flag can be seen in Fig. 5b. Images from two of the subspaces used to create the flag are shown in Fig. 5a. If we train a flag for each of the 13 poses in the PIE database using half of the subjects and illumination conditions, we can recognize the pose of the remaining images with near perfect accuracy. This organization is one way to create flags from the PIE images. A similar technique can be employed to recognize the other forms of variation as well.

5 Conclusions

In this paper we investigated the representation of several data sets in terms of flag manifolds. The flag manifolds were used to classify unlabeled patterns and did so with an accuracy comparable to nearest neighbor classification.

We infer that the geometric structure characterized by the flag captures information inherent in the data. We note that, in general, nearest neighbor classifiers perform very well, use all the available data, and grow in complexity as more data is collected. In contrast, flags also perform well, serve as prototypes and, as we have seen, have significant representational differences.

Acknowledgments. This research was partially supported by the NSF: CDS&E-MSS-1228308, DMS-1322508, DOD-USAF: FA9550-12-1-0408 P00001 and by DARPA: N66001-11-1-4184. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. T. Arias, A. Edelman, S. T. Smith, The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl*, 20:303–353, 1998.
2. J.R. Beveridge, B. Draper, J-M. Chang, M. Kirby, H. Kley, C. Peterson, Principal angles separate subject illumination spaces in YDB and CMU-PIE. *IEEE TPAMI*, 29(2), 351-363, 2008.
3. Ake Björck and Gene H. Golub, Numerical methods for computing angles between linear subspaces. *Math. Comp.*, 27:579–594, 1973.
4. J-M Chang, M. Kirby, H. Kley, C. Peterson, J.R. Beveridge, B. Draper, Recognition of digital images of the human face at ultra low resolution via illumination spaces. *Springer Lect. Notes Comp. Sci.*, 4844:733–743, 2007.
5. B. Draper, M. Kirby, J. Marks, T. Marrinan, and C. Peterson. A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32, 2014.
6. Tom Geller, Seeing is not enough. *Comm. ACM*, 54(10):15–16, 2011.
7. David Mumford, Pattern theory: The mathematics of perception, *Proceedings of ICM,* Beijing, 1:401–422, 2002.
8. F. Nater, H. Grabner, L. Van Gool, Exploiting simple hierarchies for unsupervised human behavior analysis., In *IEEE CVPR*, 2014-2021, 2010.
9. X. Pennec, Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. of Math. Imaging and Vision*, 25:127–154, 2006.
10. I.M. James, The Topology of Stiefel Manifolds, *London Math Society Lecture Note Series,* No. 24, 1977.
11. T. Sim, S. Baker, M. Bsat The CMU pose, illumination, and expression (PIE) database. In *Proc. 5th Int. Conf. Auto. Face and Gesture Recog.*, 2002.
12. P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *IEEE TPAMI, Vol. 33(11)*, 2273-2286, 2011.
13. V. S. Varadarajan, Lie groups, Lie algebras, and their representations. *Springer-Verlag Graduate Texts in Mathematics,* Vol. 102, 1984.