Pattern-set Representations using Linear, Shallow and Tensor Subspaces

Bernardo B. Gatto
Federal University of Amazonas, Institute of Computing (ICOMP),
Manaus, Brazil
bernardo@icomp.ufam.edu.br

Waldir S. S. Júnior
Federal University of Amazonas, Electronics and Information Research Centre (CETELI),
Manaus, Brazil
waldirjr@ufam.edu.br

Eulanda M. dos Santos
Federal University of Amazonas, Institute of Computing (ICOMP),
Manaus, Brazil
emsantos@icomp.ufam.edu.br

Abstract

Pattern-set matching refers to a class of problems where learning takes place through sets rather than elements. Much used in computer vision, this approach presents robustness to variations such as illumination, intrinsic parameters of the signal capture devices, and pose of the analyzed object. Inspired by applications of subspace analysis, three new collections of methods are presented in this paper: (1) New representations for two-dimensional sets; (2) Shallow networks for image classification; and (3) Subspaces for tensor representation and classification. New representations are proposed with the aim of preserving the spatial structure and maintaining a fast processing time. We also introduce a technique to keep temporal structure, even using the principal component analysis, which classically does not model sequences. In shallow networks, we present two convolutional neural networks that do not require backpropagation, employing only subspaces for its convolution filters. These networks present advantages when the training time and hardware resources are scarce. Finally, to handle tensor data, such as video data, we propose methods that employ subspaces for representation in a compact and discriminative way. Our proposed work has been applied to several problems, such as 2D data representation, shallow networks for image classification, and tensor representation and learning.

Keywords: Subspace representation, shallow networks, manifold learning, tensor analysis.

1 Introduction

The task of recognizing objects from one image has a limited capacity for recognition. For instance, single-view information may be insufficient to solve possible ambiguity due to camera's point of view or occlusion. In object recognition, partial occlusion or ambiguities due to the point of view are common difficulties. It is known that employing multiple images of the same object can be beneficial for recognition tasks. Often surveillance and industrial systems are equipped with multiple cameras, and most devices capture data in a continuous stream so that multiple images of an object are frequently available.

Classifying image sets has been well studied in classic computer vision literature and employed in several applications, including handling, learning, and classifying from multi-view cameras and videos, such as robot vision, where a data stream is available. In this setting, a set of patterns is a collection of images of the same object or event. This set can be unordered, where the time stamp of the collected images is not relevant,

or differently when the time of the captured patterns is necessary. A useful pattern-set model requires robustness to corrupt data; that is, some images may contain noise, occluded targets, or dropped frames. The model must also handle a variable set size properly without increasing computational complexity.

Subspace representation has been a common strategy to model pattern-sets, as a subspace alleviates the issues aforementioned by exploiting the geometrical structure under which images in a set are distributed. A subspace represents the image set with a fixed dimension, a model with mainly two advantageous points. First, statistical robustness to input noises, i.e., perturbations such as occlusion. Second, compactness (low subspace dimension), even if there are many images, it leads to a fixed complexity when processing the set as a subspace. Modern challenges exist in pattern-set modeling and classification. For instance, the formulation employing Principal Component Analysis (PCA) to model the pattern-sets may be insufficient to represent two-dimensional patterns existing in images. The conventional PCA applies patterns in the vectorial form of the concatenated two-dimensional images, weakening the pattern representation.

In this paper, among other contributions, we describe a new type of subspace that can process two-dimensional image sets without damaging its two-dimensional structures. We name such a model Two Dimensional Mutual Subspace Method (2D-MSM) as a reference to the Mutual Subspace Method (MSM) [1], a fundamental subspace classification algorithm. Similarly to MSM, in 2D-MSM, both the input and the learnable basis vectors span subspaces, and their mutual canonical angles perform their matching. The input subspace represents a set of patterns (e.g., images), which are then compared to reference subspaces.

Other difficulties exist with the current solutions. The traditional PCA cannot capture the ordering of the sets, which is essential when time defines the categories of the object. For instance, in action recognition from videos, the ordering of the patterns presents valuable information for representation and classification. Missing the relationship between the images in a video may decrease its representation.

To resolve this problem, we describe a subspace variant called the Hankel Mutual Subspace Method (HMSM). In this approach, the frames of an input video are arranged in a Hankel-like matrix. This maneuver prevents its ordering from being lost during the extraction of its basis vectors, improving the classification ability of the model when the frame's ordering is a discriminative factor.

Recently, subspaces have been incorporated in shallow neural networks, concretely as parameters of Convolutional Neural Networks (CNN). In this paper, we also describe a shallow network based on subspaces applied in image classification problems. This new concept not only learns the network weights without using backpropagation but is able to work under scarce training sample conditions. The proposed network is also equipped with a discriminative space, where the extracted features provide more reliable information for classification. We also developed a convolutional neural network able to process semi-supervised data efficiently. This learning paradigm is common in machine learning applications, where unsupervised information is abundant and supervised one is expensive to obtain.

Several applications employ data in a tensor format, such as video and audio data collected from self-driving cars, for instance. An example of tensor data is observed in action analysis from video data, where both spatial and temporal information are present in a structured form. In this scenario, the spatial and temporal information can be handled independently within different representations, such as subspaces.

Tensors can be defined as a generalization of matrices, providing a natural representation of multi-dimensional data. Applications of subspaces for learning from tensor data frequently make use of the MSM. These solutions are employed to solve gesture and action recognition problems, where video clips are expressed by 3 subspaces, where each subspace is computed from one of the tensor's modes. Despite its desirable properties, MSM cannot extract discriminative features from the tensors; therefore, a more powerful subspace representation should be developed.

Encouraged by the results obtained by the Fukunaga-Koontz Transformation (FKT) [2] in subspace-related solutions, we describe a formulation of FKT to handle tensor data. The introduced formulation has been applied to image-set modeling from tensor data to solve action learning from videos. In this solution, tensor data is decomposed into several subspaces, each one representing a particular factor. For instance, a grayscale video presents three main factors, two space dimensions and a temporal one. In this scheme, subspace-based techniques can be employed directly. We also developed another solution for tensor data when only unsupervised training data is available. A k-means algorithm is adapted to work directly on subspaces, saving time and memory when operating on a large amount of data.

Therefore, we report our advances in subspace learning by introducing new subspace representations and shallow networks. These representations may reduce the complexity of solving pattern-set classification and related problems. We explore different approaches to describe and classify pattern-sets, using diverse machine learning techniques to obtain the most suitable results. More precisely, we (1) Investigate variants of subspace-based methods that represent two-dimensional data to preserve the spatial relationship between the patterns. (2) Introduce shallow networks capable of learning the convolution kernels through subspaces without employing backpropagation. (3) Propose methods that can represent tensorial data, preserving the

temporal relationship between patterns.

Our contributions related to 2D-subspaces are: (1) The processing time to compute each subspace is reduced due to the compactness of subspaces inherited from 2D-PCA and variants, decreasing its computational cost. (2) To employ variants of nonlinear subspace techniques, we create Kernel E2D-PCA and Kernel color-PCA based on the Kernel 2D-PCA formulation. (3) We propose three versions of Kernel Two-Dimensional Subspace by employing Kernel 2D-PCA, Kernel E2D-PCA, and Kernel color-PCA.

Our contributions related to shallow networks are as follows: (1) A shallow network for handwritten character classification through the use of FKT. We generate a discriminative subspace projection to enhance the discriminability across the handwritten image classes. (2) An average pooling layer is introduced to increase the number of layers without increasing the feature dimensionality, preserving a low computational cost as the number of layers increase. (3) We propose a new type of convolutional kernel based on the orthogonalization of subspaces. We show that the basis vectors of this subspace are useful as convolutional kernels, efficiently handling supervised data.

Our contributions related to tensor analysis are as follows: (1) We propose a novel tensor data representation called n-mode GDS. (2) We incorporate the n-mode Generalized Difference Subspace (GDS) projection on the conventional product manifold, providing a tensor classification framework. (3) We optimize the proposed n-mode GDS projection on the product manifold space through a redefined Fisher score designed for tensor data. (4) We introduce an improved version of the geodesic distance, which incorporates the importance of each tensor mode for classification.

It is worth mentioning that the works related to shallow networks and tensor analysis were further expanded, and new variants capable of handling semi-supervised and unsupervised data were also introduced. These formulations are used to support new applications and cover diverse learning paradigms. The next section describes the basic idea behind subspace learning and practical examples of its applications. The detailed steps and the motivation behind subspace analysis are also given.

This paper continues as follows: The next section outlines the basic idea behind subspace learning and its motivation. In section 3, we present a two-dimensional representation for the subspace-based pattern-set classification. The Hankel subspace method for classifying gestures is presented, allowing subspace-based methods to represent ordered data. In section 4, we address the problem of learning convolutional neural network kernels through subspaces. We present a comparative study of the different convolutional kernels for shallow neural networks, and we introduce a neural network learning which does not require backpropagation. Our method of tensor decomposition enhancement using discriminative subspaces and the product of manifolds is described in section 5. We also present the technique to combine the discriminative subspaces to classify the different tensor factors into a unified manifold. The key contributions, findings, and future directions are summarized and evaluated in section 6.

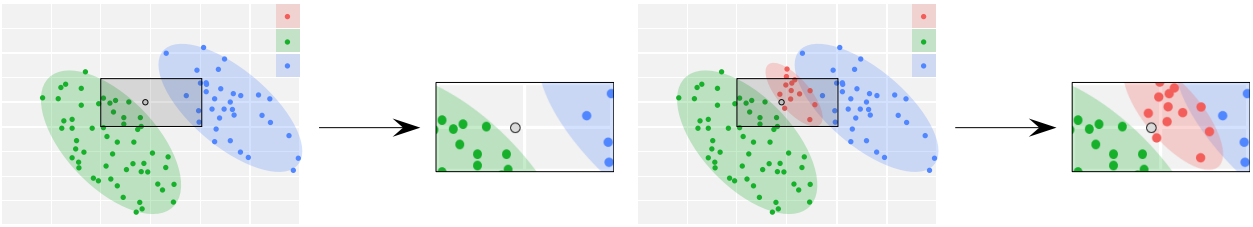
2 Background theory

In the field of machine learning and computer vision, learning algorithms are usually employed to classify single patterns in a one-to-one correspondence. In this case, given a pattern vector, a learning algorithm should indicate a semantic aspect of this pattern. However, there are applications that demand processing pattern-sets, where the classification process is held entirely in terms of collections in a set-to-one fashion. With video cameras being widely used, it is a natural choice to solve a classification problem using pattern-sets. Compared to single pattern-based methods, the pattern-set classification directly handles changes of appearance and makes decisions by comparing the query set with gallery sets. This paradigm provides advantages when patterns are described as sets, such as in face recognition from video.

2.1 Pattern-set classification

The Mutual Subspace Method (MSM) [1] is a common technique employed to represent and classify pattern-sets. We define a pattern-set as a collection of samples relating to a particular category and further represented by a subspace. In this approach, a set of patterns is analyzed in a batch instead of separately. Matching pattern-sets arises naturally in distinct circumstances, such as when the target pattern is available in a data stream, where it is possible to evaluate patterns at a time. Another practical example is when the data is contained in a bag, such as the profile pictures in a social media network. Then, it is reasonable to expect that most of the images in a profile collection belong to the same subject. Subspace analysis is one of the basic problems in the machine learning and computer vision community.

The theory of the subspace method was developed from the observation that patterns of the same object produce a compact cluster in high-dimensional vector space [3]. This compact cluster can be described by a subspace, which is generated by using PCA. It is worth mentioning that the subspace method operates by representing each class with a subspace, differently from the Eigenface [4], where only a subspace is computed



(a) Single pattern analysis.

(b) Multiple pattern analysis.

Figure 1: In single vector pattern analysis, the classification is based on the minimum distance between an input pattern and a distribution of reference patterns, which may not reflect the desired properties for classification. The minimum distance is unstable due to the variations in point of view or illumination. Differently, multiple patterns provide stability, revealing desired properties for classification.

to embed the patterns. The MSM has been applied in several applications, including audio data [5], and image sets [6]. The advantages of subspace-based methods include their high compactness ratio and their flexibility to handle different types of data.

2.2 The Mutual Subspace Method

To represent a pattern-set by a subspace, we adopt the observation that a set of images lies in a cluster, which a set of orthonormal basis vectors (computed by Singular Value Decomposition (SVD)) can efficiently represent. Given an $N \times N$ feature matrix $\mathbf{X}$, we may conduct a decomposition for extracting information regarding its linear correlations and geometric structures. The SVD decomposition will produce a set of eigenvectors $\mathbf{U} = \{u_1, u_2, \ldots, u_N\}$, and a set of eigenvalues $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. In algebraic terms, each vector in $\mathbf{U}$ represents an axis and each value in Λ describes how important each axis is w.r.t reconstruction of the features in $\mathbf{X}$. Another useful idea of $\mathbf{\Lambda}$ is that it represents how much the vectors in $\mathbf{X}$ are correlated, which is a valuable guidance towards the redundant and non-redundant information. The employed SVD decomposition in this arrangement can be computed as follows:

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \ . \tag{1}$$

The matrix $\mathbf{U}$ is an $N \times N$ matrix where each column is a singular vector. Then, $\mathbf{\Lambda}$ is a $N \times N$ matrix where the main diagonal presents the singular values in descending order. The ordered nature of the singular values contained in $\mathbf{\Lambda}$ can be directly employed to reveal the importance of each singular vector in $\mathbf{U}$. The analysis of $\mathbf{\Lambda}$ is useful in various problems, such as dimensionality reduction, signal filtering, and feature extraction. By understanding the importance of each singular vector in $\mathbf{U}$, it is possible to select a small set of them $\mathbf{U}'$ by removing all but the top K singular values in the diagonal of $\mathbf{\Lambda}$. It is worth mentioning that $\mathbf{U}$ satisfies $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$.

Figures 1 and 2 present the contrast between classifying a single pattern and multiple patterns. A single pattern can be described as a point in a high-dimensional feature vector space where an image pattern is handled as an N-dimensional vector. The minimum distance is very unstable because the input pattern may fluctuate due to the variations in point of view or illumination. Differently, multiple patterns provide stability. Therefore, a method based on the similarity between the pattern-sets is slightly influenced by the variations discussed above. Besides, it is worth noting that the similarity between the pattern-sets may reflect the similarity between 3D shapes of 3D objects.

2.3 Selecting basis vectors

As mentioned before, the basis vectors generated by SVD may represent a set of patterns compactly. The following criteria can be utilized to obtain the compactness ratio of this transformation:

$$\mu(K) \leq \sum_{i=1}^{K}(\lambda_i) / \sum_{i=1}^{D}(\lambda_i) \ . \tag{2}$$

In the above equation, K is the number of the selected basis vectors which will span a subspace, λ_i corresponds to the i-th eigenvalue of $\mathbf{X}\mathbf{X}^\top$. Then, $D = \text{rank}\left(\mathbf{X}\mathbf{X}^\top\right)$. It is useful to set K as small as possible to achieve a minimum number of orthonormal basis vectors, maintaining low memory requirement. In addition, $\mu(K)$ should be fixed in a form that best represents each set of images and also satisfying the

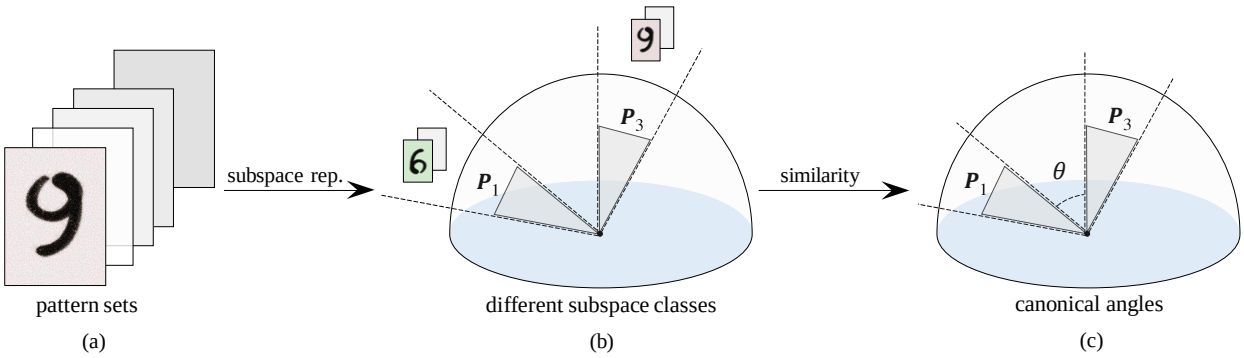


Figure 2: Here is an attempt to display the MSM algorithm schematically since it is not possible to draw the subspaces in a high-dimensional space. (a) An unordered set of images representing a particular character is processed (b) The subspaces are produced by extracting the basis vectors from the set of patterns. (c) The canonical angles are employed to achieve the similarity between $\mathbf{P}_1$ and $\mathbf{P}_3$.

application requirements. In practical terms, we should select $\mu(K)$ to meet the trade-off of compactness ratio and representativity of the subspace. So far, there is no precise solution to determine the minimum number of basis vectors that best represents a set of patterns.

2.4 Computing the similarity between subspaces

A popular procedure for measuring the similarity between subspaces is by computing the principal angles, also known as canonical angles [7, 8]. Jordan introduced the theory of computing the canonical angles between subspaces. Since then, the theory has been developed and improved as a helpful tool in many applications. The canonical angles provide information concerning the relative location of two subspaces in a Euclidean space, which gives a clue regarding how similar two subspaces are. For example, given two subspaces, $\mathbf{P}$ and $\mathbf{Q}$, then the set of principal angles between these subspaces can be described as follows.

Let p be the dimension of subspace $\mathbf{P}$, and q be the dimension of subspace $\mathbf{Q}$. If $p \geq q \geq 1$, then exist q canonical angles $\theta_1, \theta_2, \ldots, \theta_q \in [0, \pi/2]$ between the subspaces $\mathbf{P}$ and $\mathbf{Q}$. In practical terms, SVD can be applied to compute the eigenvalues of $\mathbf{U}^\top \mathbf{V}$; then, the principal angles are readily available by computing the cosine of the eigenvalues. Let $\mathbf{U}$ and $\mathbf{V}$ be orthonormal bases of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Then:

$$\mathbf{R}\mathbf{\Sigma}\mathbf{S} = \mathbf{U}^\top \mathbf{V} \ . \tag{3}$$

The matrix $\mathbf{\Sigma}$ provides the set of eigenvalues, $\sigma_1, \sigma_2, \ldots, \sigma_q$, in its main diagonal, with $0 \leq \sigma_1 \leq \sigma_2 \leq \ldots, \leq \sigma_{q-1} \leq \sigma_q \leq 1$. After obtaining the set of eigenvalues, the canonical angles are computed as follow:

$$\theta_i = \cos^{-1}(\sigma_i) \ , \quad \forall \ i = 1, \ldots, \min(p,\ q) \ . \tag{4}$$

The next section describes more sophisticated subspace-based methods, including kernel methods and discriminant analysis. We introduce a faster version of MSM, which employs two-dimensional patterns directly, without using a vectorization process. Additionally, we present an MSM version which efficiently represents ordered patterns, which is essential to represent gesture and actions from videos.

3 New subspace representations

Principal Component Analysis [9] usually achieves subspace representations that minimizes the mean square error. The PCA subspace representation simplifies the classification between a set of reference images and an input image vector through the use of multiple canonical angles [10]. Following this main concept, Kernel Orthogonal Mutual Subspace Method (KOMSM) [11] is an extension of MSM that is able to handle nonlinear patterns. In KOMSM, the model discriminability is further enhanced by the orthogonalization method proposed by Fukunaga-Koontz Transformation [2].

KOMSM has been used in many applications due to its flexibility in dealing with multiple class problems and straightforward implementation [12]. However, its performances are not satisfactory for more advanced systems, wherein more complicated structures should be classified. In short, these methods employ PCA to generate the subspaces as follows: First, each two-dimensional image from a set is reshaped to a one-dimensional vector. Then, a covariance matrix is computed from these reshaped images. And finally, a set

of basis vectors is generated from this covariance matrix through eigenvector decomposition. This reshaping procedure leads to a very high dimensional vector space, increasing the overall computational complexity.

To overcome these drawbacks and motivated by 2D-PCA [13], we propose a Kernel Two-Dimensional Subspace (K2DS) and a Two-Dimensional Mutual Subspace Method (2D-MSM) to speed up the learning and the matching processing times. The main difference between PCA and 2D-PCA is that 2D-PCA employs the image matrix directly, without vectoring the patterns, to generate the covariance matrix, which is smaller than the covariance matrix produced by PCA. Since KOMSM and MSM systematically operate on the basis vectors produced by PCA, replacing PCA with 2D-PCA reduces the memory cost since the basis vectors produced by 2D-PCA are more compact. Consequently, K2DS and 2D-MSM are much more efficient than KOMSM and MSM in memory and time complexity. This section introduces the concept of nonlinear two-dimensional subspaces, achieving improvements over conventional KOMSM and MSM [14, 15].

Even though subspace-based methods can achieve high performance when applied to image set recognition, these methods cannot cope with temporal information [16], as required for an efficient gesture representation, for instance. The temporal information may contain discriminative information since its ordering may represent different gesture categories. We also propose a new method based on clustering and sample selection to reduce computational complexity and simultaneously preserve the temporal information to solve this problem. This new representation is mainly based on the Hankel matrix formulation, where the image patterns can be stored in a manner where the ordering of the images is preserved. We select representative samples from each image gesture set to compound its corresponding Hankel matrix in this approach. By exploiting this strategy, we obtain a smaller covariance matrix, compared to the traditional methods, where we can easily extract its basis vectors.

Problem formulation: Let C be the number of image sets, which are given by $\mathbf{A} = \{\mathbf{A}_m\}_{m=1}^{C}$, where $\mathbf{A}_i$ is a set containing M two-dimensional images and each $\mathbf{A}_i$ set belongs to one of the C classes. Then, we assume a nonlinear mapping that represents each $\mathbf{A}_i$ set in terms of its variance. This nonlinear transformation is in such a way that the M images are converted into $k-$dimensional orthonormal vectors ordered by its accumulated energy, where $k \ll M$. This new representation, $\{\Phi_i\}_{i=1}^{C}$, provides a more compact manner to represent each $\mathbf{A}_i$ set, and its computational classification cost is greatly reduced. Each Φ_i basis vector spans a reference subspace $\mathbf{P}_i$, where its compactness ratio is empirically defined by choosing the first k vectors, ordered by its accumulated energy. Finally, for a given set of two-dimensional test images, $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^{M}$, the task is to compute a subspace $\mathbf{Q}_\mathbf{Y}$ that represents $\mathbf{Y}$ in terms of its variance and predicts its corresponding image set based on the nearest $\mathbf{P}_i$ reference subspace. Figure 3 shows the overall schematic flowchart of our proposed method.

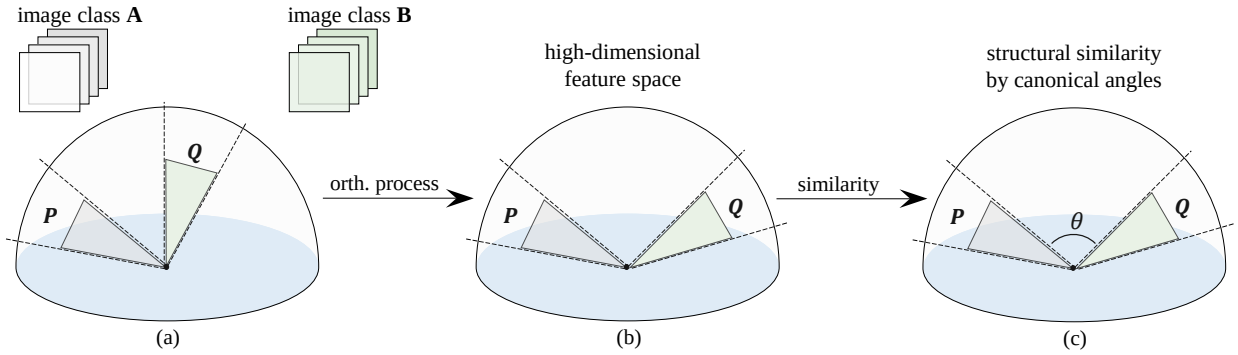


Figure 3: Conceptual illustration of the proposed method. (a) Training image sets **A** and **B** are expressed as subspaces. (b) Pattern distribution of each class is represented by nonlinear subspaces, which are generated by kernels 2D-PCA and variants. The procedure employed will generate K2DS and variants. The nonlinear subspaces are orthogonalized in high dimensional feature space, increasing the robustness of the method. (c) The similarities between the input nonlinear subspace and reference nonlinear subspaces are calculated using canonical angles. Then, the class assigned to the set of input images is the class with the highest similarity.

3.1 Generating Nonlinear Subspaces via K2D-PCA

Encouraged by the efficiency of KOMSM framework and the advantages of 2D-PCA and variants, we propose a novel framework that inherits the capabilities above. Besides, there is no work regarding the applications of nonlinear subspaces produced by K2D-PCA and variants of 2D-PCA for image set classification.

The K2D-PCA generalizes 2D-PCA by first mapping the data nonlinearly into a higher dimensional dot product space $\mathcal{F}$. Given a set of samples $\mathbf{x}_i$, with $i = \{1, \ldots, N\}$, let $\mathbf{x}_i = (\mathbf{x}_i^1, \ldots, \mathbf{x}_i^S)^T$ to be $S \times U$ matrix,

where $\mathbf{x}_i^j$, with $j = \{1, \ldots, S\}$ is the j row of $i-$th sample. Suppose that ϕ is an implicit nonlinear mapping which maps the $\mathbf{x}_i^j \in \mathbb{R}^N$ into a higher or even infinite dimensional Hilbert space, such as $\phi \colon \mathbb{R}^N \to \mathcal{F}$ and $\mathbf{x}_i^j \to \phi(\mathbf{x}_i^j)$, where ϕ is a nonlinear function and $\mathcal{F}$ is very high dimensional. The implicit feature vector ϕ does not require explicit computation, which can just be obtained by computing the dot product of two vectors in $\mathcal{F}$. The dot product can be calculated through a kernel function:

$$K_F\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle \ , \tag{5}$$

and the ϕ-mapping sample is defined as: $\phi(\mathbf{x}_i) = (\phi(\mathbf{x}_i^1), \ldots, \phi(\mathbf{x}_i^s))^\top$, where $i = \{1, \ldots, N\}$. Kernel 2D-PCA aims to perform the 2D-PCA in the feature space $\mathcal{F}$. We should compute $\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$ to perform the 2D-PCA in the nonlinear mapped patterns. At this point, we need to choose a form for the kernel function $\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle = K_F$. In our proposed method, we use a Gaussian kernel, since this kernel is indicated for the images sets: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)$, where the value of σ is determined by experimentation. The function ϕ maps an input pattern onto an infinite feature space $\mathcal{F}$. It is worth remarking that a linear subspace generated by this kernel approach can be regarded as a nonlinear subspace in the input space [11, 17].

3.2 Orthogonalization of Subspaces

We will now explain the procedure to compute the orthogonalization matrix $\mathbf{O}$ in order to orthogonalize the C classes M-dimensional subspaces with the orthogonal basis vectors $\{\phi_i\}_{i=1}^M$. This orthogonalization procedure enhances the difference between the class subspaces, increasing the recognition rate of the framework. Let the projection matrix corresponding of the i-th class subspace $\mathbf{P}_i = \sum_{i=1}^M \phi_j \phi_j^\top$, where ϕ_j is the $j-$th orthogonal basis vector of $\mathbf{P}_i$. Next, the total projection matrix is defined as: $\mathbf{G} = \sum_{i=1}^R \mathbf{P}_i$. By applying singular value decomposition on the total projection matrix $\mathbf{G}$, we obtain the whitening matrix $\mathbf{O}$:

$$\mathbf{O} = \mathbf{\Lambda}^{-1/2} \mathbf{D}^\top \ , \tag{6}$$

where $\mathbf{D}$ is a matrix whose i-th column vector is the eigenvector of $\mathbf{G}$ corresponding to the i-th highest eigenvalue, and $\mathbf{\Lambda}$ is a diagonal matrix with the i-th highest eigenvalue of $\mathbf{G}$ as the i-th diagonal component. After the whitening process and obtaining a set of basis vectors that best approximate each subspace to its corresponding set of images, we can compute the similarity between them. This procedure is achieved by applying subspace similarity or principal angles [18]. This procedure is similar to the one described in eq (4).

We expect that the proposed methods K2DS and variants will reduce the computational complexity of KOMSM, achieving a faster processing time from the improvements above.

3.3 Computational Advantage

The main difference of 2D-MSM from traditional MSM is that 2D-MSM does not require transforming image matrices into vectors. Thus, it reduces the computational complexity of constructing the subspaces and reduces the computation time of the matching. All these aspects make the proposed algorithm superior to MSM in terms of computational time. Besides, the process of extracting the basis vector of each 2D-PCA variant determines its processing time and the dominant complexity of each algorithm. In 2D-PCA and variants, their time requirements and the computational complexity are similar.

3.4 Experimental Results Summary

We conducted image set matching experiments on 7 datasets: ALOI, RGB-D for the object recognition task, Honda/UCSD, YouTube Celebrities (YTC), PubFig83 and CMU-MoBo (CMU MoBo gait database) for the face recognition and ASL Finger Spelling dataset.

In our results, the classification time of K2DS is about 3 times faster than the learning time and matching time of KOMSM, revealing that the computational cost to obtain the subspaces from K2D-PCA employed by K2DS is more efficient than the computational cost of KOMSM. For the 2D-MSM, the classification time of 2D-MSM (and most of its variants) is about 4 times faster than the learning time and matching time of MSM, revealing that the computational cost to obtain the subspaces from the covariance matrix employed by 2D-MSM (and variants) is more efficient than the covariance matrix used by MSM. Finally, K2DS and 2D-MSM achieved comparable accuracy with the ones provided by KOMSM and MSM.

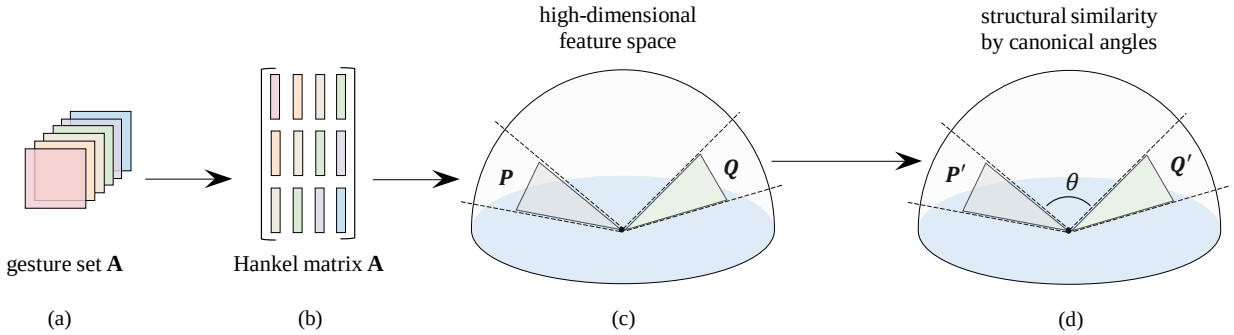


Figure 4: Conceptual figure of our method. (a) An ordered subset of images representing a gesture $\mathbf{A}$ is handled, where a selection criterion is employed to reduce the number of images. (b) Then, the Hankel matrix $\mathbf{H_A}$ is created from the set of the selected images. (c) After that, we extract the basis vectors from the Hankel matrix $\mathbf{H_A}$ to produce $\mathbf{P}$ and its soft weights. Then, we orthogonalize the subspace to achieve a subspace $\mathbf{P'}$. (d) The structural similarity between $\mathbf{P'}$ and a reference subspace $\mathbf{Q'}$ is computed.

4 Hankel Subspaces

Although controlling machines employing gesture recognition is useful, it includes many difficulties; for instance, the distribution of a gesture largely varies depending on viewpoints due to its multiple joint structures. To solve these problems, we introduce the Hankel Mutual Subspace Method [19, 20] based on the Hankel matrix formulation to describe pattern sets and the MSM framework, as illustrated in Figure 4.

The problem formulation of matching time-aware pattern-sets is similar to the problem of pattern-set matching observed in the previous section, except that the ordering of the patterns should be preserved since some gesture classes present their semantic information correlated to the pattern ordering.

4.1 Hankel Matrix-based Gesture Representation

A gesture that is handled as a time series of vectors can be regarded as the output of a Linear Time Invariant (LTI) system of unknown parameters. Then, given a sequence of output measurements $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$, its block-Hankel matrix is: $\widetilde{\mathbf{H}}_\mathbf{A} = \left[(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{n-1})^\top,\ (\mathbf{A}_2, \mathbf{A}_3, \ldots, \mathbf{A}_n)^\top, \ldots,\ (\mathbf{A}_{m+1}, \mathbf{A}_{m+2}, \ldots, \mathbf{A}_M)^\top\right]$, where n is the maximal order of the system, M is the temporal length of the sequence, and it holds that $M = n + m - 1$. Finally, the Hankel matrix can be normalized: $\mathbf{H_A} = \widetilde{\mathbf{H}}_\mathbf{A} / \left(||\widetilde{\mathbf{H}}_\mathbf{A}\ \widetilde{\mathbf{H}}_\mathbf{A}^\top||_F\right)^{1/2}$.

4.2 Creating Hankel Subspaces

To represent an ordered image set $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$ by subspace and preserve spatial and temporal information, we introduce the concept of Hankel subspace for gesture recognition. Subspace-based methods exploit the fact that a set of images lies in a cluster, which can be efficiently represented by a set of orthonormal basis vectors [1]. We assume that the same formulation can be regarded for Hankel subspaces so we can achieve a novel representation for gesture-based image recognition. Given a normalized Hankel matrix $\mathbf{H_A}$ from the ordered image set $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$, we can compute an auto-correlation Hankel matrix as: $\mathbf{C_A} = \mathbf{H_A}\mathbf{H_A}^\top$, when $\mathbf{C_A} \in \mathbb{R}^{K \times K}$, its eigendecomposition generates a set of eigenvectors $\mathbf{\Phi_A} = \{\phi_i\}_{i=1}^{K}$ that spans $\mathbf{P_A}$.

4.3 Selecting Samples

When creating a Hankel matrix, the number of images in a set and its dimension are crucial factors in terms of computational resources. To alleviate this issue, we introduce two approaches based on sample selection.

Random sample selection: Here, we randomly select images from the set, preserving its original order. We adopt this temporal sampling scheme in the image sequence since close images in time hardly change their appearance, containing high level of redundant information to identify the gesture that is being performed. This strategy also allows us to deal with sample reduction with a straightforward implementation.

Clustering selection: The second approach employs a clustering strategy, where the centroids obtained by k-means clustering are employed to represent the set, decreasing its number of images. The use of k-means clustering was previously employed for kernel dimensionality reduction in [21]. The advantage of using clustering is that the k centroids of the clusters will represent most of the relevant gesture information for discrimination, eliminating redundant images, achieving a good accuracy with low computational cost.

The orthogonalization process employed to improve the discriminant ability of the method is similar to the one described in Eq. (6), and the matching process is described in Eq. (4).

4.4 Experimental Results Summary

We employed Cambridge gesture dataset [22] for general gestures classification and Human-Computer Interaction (HCI) dataset [23], which contains computer interface gestures. We compare HMSM and variants with several state-of-the-art subspace-based methods. HMSM and variants achieved competitive accuracy, similar to discriminative methods. This indicates that the temporal information extracted by the Hankel representation is compelling, even when random samples are selected as long as the temporal order is preserved. We found out that k-means clustering is more efficient than random sample selection. Selecting the centroids obtained by k-means conserves the gesture manifold's structural information more than random selection. As a final remark, we would like to emphasize that HMSM and its variants do not employ any learning scheme different from the compared methods. This demonstrates the effectiveness of employing Hankel subspace for gesture representation.

5 Shallow neural networks based on subspaces

Deep learning-based approaches, especially those using deep Convolutional Neural Networks (CNN), have been employed in image classification problems. Learning through deep neural networks has received significant attention due to its improvements over hand-crafted features. The central concept of deep learning is that relevant information required to recognize image patterns can be obtained hierarchically. Despite encouraging results, the fine-tuning of deep neural network parameters is time-consuming. Many shallow networks have been proposed based on PCA, where convolutional kernels are obtained from PCA, ICA, or DCT basis vectors. For instance, PCANet (PCA network) [24] uses a CNN architecture with no pooling layers, no activation functions, and without using backpropagation to learn its weights. Although only PCA or Linear Discriminant Analysis (LDA) basis vectors define the convolutional kernels, they present competitive performance compared to the state-of-the-art results achieved in several image classification tasks.

These models do not provide discriminative features in more complicated computer vision problems. To improve the discriminant potential of such networks, we propose a shallow network based on the Fukunaga-Koontz Transform (FKT) [25] to generate discriminative features and handle complex distributions. It is worth noting that there is no method using FKT in a shallow network approach.

Instead of creating the transformation matrix from the sum of the auto-correlation matrices, we utilize the sum of the projection matrices, producing more stable features since the subspaces can have their dimensions independently estimated. Therefore, instead of employing PCA or LDA to learn the convolutional kernels, we use the subspace generated by FKT. Using the FKT decorrelation subspace, we build the FKNet [26], a shallow network able to minimize the correlation between different image classes. In FKNet, the training images are firstly compressed as subspaces to minimize their within-class distance. Besides, the decorrelation subspace based on the compressed data is more robust to outliers. Therefore, it is expected that such convolutional kernels can reveal more discriminative information compared to related shallow networks [27].

Fukunaga-Koontz Network: Figure 5 shows the conceptual diagram of the proposed shallow network. FKNet processes images as follows. An input image is processed by a convolutional feature extraction layer, followed by a mean-pooling or other convolutional layers. Then, binary hashing is applied to the produced features to achieve dimensionality reduction. Finally, block-wise histogramming is employed to attain relative rotation invariance and create the final feature vector.

5.1 Representation by image patches

Given a dataset $\mathbf{X}$ consisting of N labeled training images of size $H \times W$, we extract patches of size $K_1 \times K_2$ from $\mathbf{X}$. This procedure is performed by taking a patch around each pixel from each one of the N training images. Here, we denote the set of image patches as $\mathbf{P}$. Given that each image patch will have size $K_1 \times K_2$, the set $\mathbf{P}$ will contain $N_{\mathbf{P}} = HWN$ patches. It is worth noting that, after collecting the patches of all the images, FKNet does not perform the mean-removal operation on $\mathbf{P}$, as employed in PCANet, since this operation modifies the subspace obtained.

5.2 Computing image patches subspaces

To create subspaces, we will use the patch set $\mathbf{P} = \{p_i^j\}_{i,j=1}^{N_j,C}$, where C stands for the number of classes and N_j is the number of patches in the $j-$th class. In this C class classification problem, it is required to compute C feature matrices $\{\mathbf{A}_j\}_{j=1}^{C}$. For each feature matrix $\mathbf{A}_j$, we compute the auto-correlation matrix

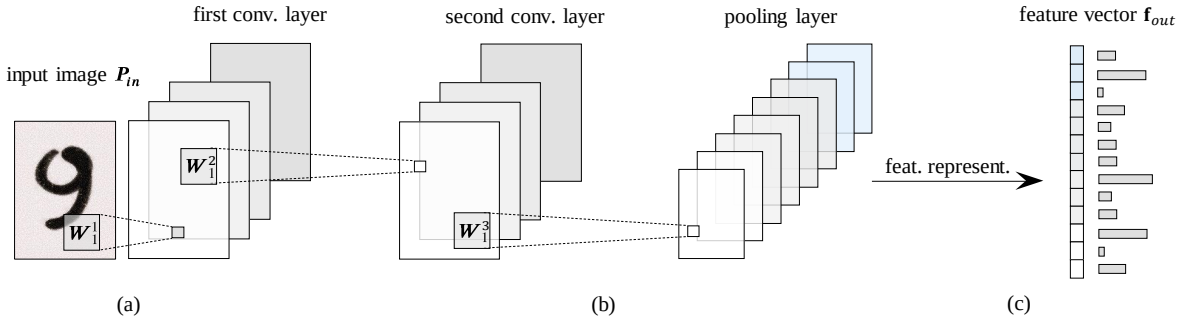


Figure 5: The proposed shallow network architecture: A convolutional feature extraction layer processes an input image based on FK convolutional layer, followed by another FK layer. Then, an average pooling layer is employed. Finally, binary hashing and a block-wise histogramming produce the final feature vector.

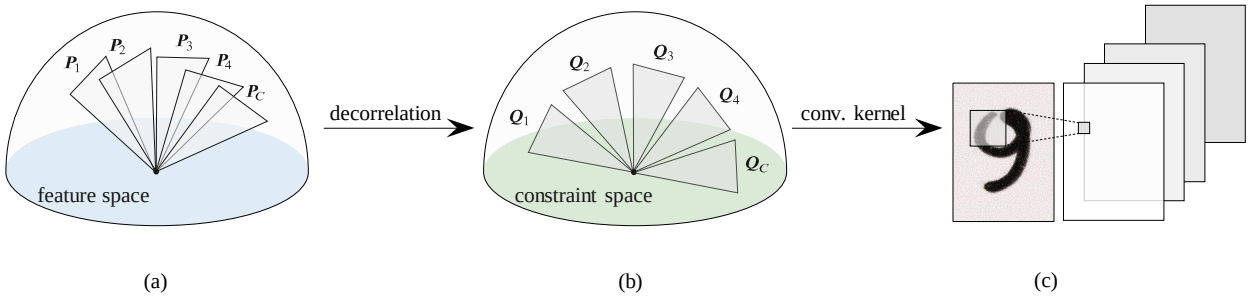


Figure 6: The decorrelation process generated by Fukunaga-Koontz transform and its application in this chapter. (a) Image sets form clusters in a low-dimensional space, which can be represented by $\mathbf{P}_i$ subspaces. These subspaces, however, are not optimal for classification due to lack of discriminative mechanism. (b) FTK is employed to decorrelate the subspaces. (c) When subspaces $\mathbf{P}_1, \mathbf{P}_2 \ldots, \mathbf{P}_C$ represent image patches, the FKT transformation matrix can be used as a convolutional kernel.

$\mathbf{C}_j = \mathbf{A}_j{}^\top \mathbf{A}_j$. Equipped with all C auto-correlation matrices, we can move forward to calculate the matrix $\mathbf{U}_j$ of eigenvectors which diagonalizes the auto-correlation matrix $\mathbf{C}_j$ as follows: $\mathbf{D}_j = \mathbf{U}_j{}^{-1}\mathbf{C}_j\mathbf{U}_j$ with $j = 1, \ldots, C$ and each $\mathbf{U}_j$ is a $K_1K_2 \times K_1K_2$ matrix satisfying $\mathbf{U}_j\mathbf{U}_j{}^\top = \mathbf{U}_j{}^\top\mathbf{U}_j = \mathbf{I}$. The columns of $\mathbf{U}_j$ that correspond to nonzero singular values compound a set of orthonormal basis vectors for the range of $\mathbf{C}_j$. $\mathbf{D}_j$ is the diagonal matrix of eigenvalues of $\mathbf{C}_j$. Unlike PCANet, FKNet creates a subspace for each class independently, exploiting its intrinsic characteristics in a more effective way.

5.3 FKT for image patches subspaces decorrelation

Once equipped with all the C image patches subspaces $\mathbf{P}_j$ and their R_j dimensions have been computed, we can now use FKT to generate the matrix $\mathbf{F}$ that can decorrelate the subspaces. Then, each set of basis vectors $\mathbf{U}_j$ spans a reference subspace $\mathbf{P}_j$, where its compactness ratio is empirically defined by choosing the first R_j vectors, ordered by its accumulated energy, as shown in Eq. (2). The method to generate the matrix $\mathbf{F}$ that efficiently decorrelates the C R_j-dimensional classes subspaces is explained as follows. First, we compute the total projection matrix as: $\mathbf{G} = \sum_{j=1}^{C} \mathbf{U}_j\mathbf{U}_j^\top$. The eigendecomposition of the total projection matrix $\mathbf{G}$ produces a $K_1K_2 \times K_1K_2$ decorrelation matrix $\mathbf{F}$. This procedure is better described by the following equation: $\mathbf{F} = \mathbf{\Lambda}^{-1/2}\ \mathbf{B}^\top$, which is the orthogonalization process in (6). Figure 6 illustrates the procedure to construct FTK and its application as convolutional kernels.

5.4 Fukunaga-Koontz convolutional kernels

After obtaining $\mathbf{P}$ and $\mathbf{F}$, we can now compute the FK convolutional kernel. In our formulation, each basis vector of $\mathbf{F} = \{\mathbf{w}_1, \ldots, \mathbf{w}_{N_\mathbf{F}}\}$ will be a convolutional kernel in the network. According to this formulation, the definition of the Fukunaga-Koontz convolutional kernel is: $\mathbf{W}_l = \text{map}_{K_1\times K_2}(\mathbf{w}_l)$ with $l = \{1, 2, ..., L_S\}$, where the operator $\text{map}_{K_1\times K_2}(\cdot)$ maps an input vector $y \in \mathbb{R}^{K_1K_2}$ onto a matrix $\mathbf{Y} \in \mathbb{R}^{K_1\times K_2}$ and L_S is

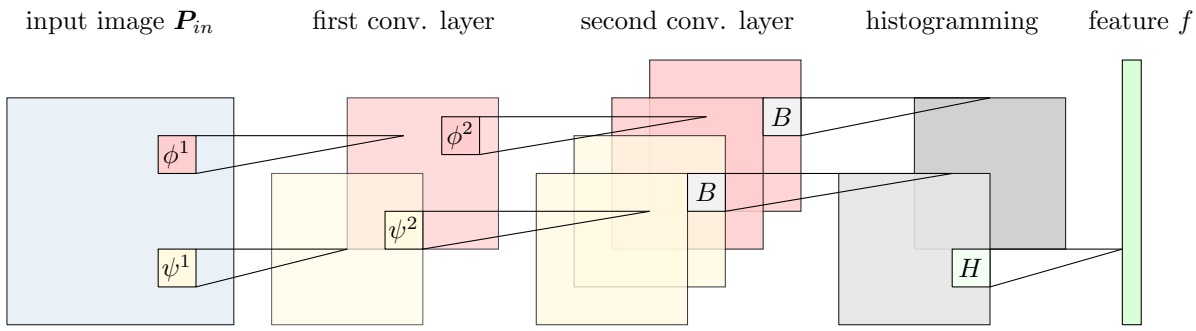


Figure 7: Conceptual figure of the semi-supervised network. The network employs two distinct filters banks based on PCA and GDS. To reduce the high dimensionality of the feature vectors and increase rotation invariance, the proposed method is followed by a feature mapping that includes binarization and block-wise histogram. Similar to most shallow networks, the classification is performed by linear SVM.

the number of convolutional kernels in the S-th convolutional layer.

Given an input image $\mathbf{P}_{in}$, the output image $\mathbf{Y}_l$ of a convolutional layer is obtained by the following operation: $\mathbf{Y}_l = \rho(\mathbf{W}_l * \mathbf{P}_{in})$ with $l = \{1, 2, ..., L_S\}$, where $*$ refers to a convolution with zero-padding in the boundary of the image patch and $\rho(\cdot)$ is an average pooling operator, which may or may not be present in a particular layer, defined by a $B_1 \times B_2$ window, where B_1, $B_2 \in \mathbb{N}^+$.

5.5 A semi-supervised neural network based on subspaces

In supervised machine learning, classifiers employ labeled data to create models. However, labeled data is often challenging and expensive to obtain in many practical situations. For example, real-world remote sensing, medical image analysis. In contrast, models based on unsupervised learning are generated from unlabeled data, usually readily available and can be obtained at a low cost.

There is often no consensus on how to employ labeled and unlabeled data to improve machine learning models due to the large imbalance between labeled and unlabeled data. Therefore, most classification methods produce models based only on labeled datasets, neglecting unlabeled data. There is in literature a class of learning techniques called semi-supervised learning that aims to solve this problem. This class may be categorized as supervised learning, though it also uses unlabeled data for training. In general, these techniques employ a large amount of unlabeled data with a small number of labeled ones. Many studies show that this kind of combination can significantly enhance learning accuracy over unsupervised learning.

Many solutions based on deep representations have achieved state-of-the-art results. However, this comes with a cost of a considerable number of parameters to be trained, requiring a large amount of data to be employed for training, which can lead to a high computational cost, even when computational resources are equipped with GPU are available. As a result, the computational complexity required from most of the deep learning architectures prevents some computer vision applications from fully employing the capabilities of deep CNN.

As an alternative, shallow networks have been proposed to exploit the advantageous characteristics of deep learning models while lightening the computational cost associated with their training. Although these networks hold hierarchical structures, their weights are obtained through non-derivative methods, giving them a processing time advantage over the traditional deep network models by several orders of magnitude. For instance, PCA or LDA are employed to replace the convolutional kernels of a CNN in many computer vision frameworks. While presenting a simple architecture, this strategy exhibited performance comparable to the state-of-the-art for several image classification tasks.

Even though shallow networks have been successfully applied in various recognition tasks, such methods can only describe either supervised or unsupervised data and cannot efficiently exploit both. Here, we describe a semi-supervised network proposed to solve this issue [28]. This semi-supervised network employs filter banks produced by both PCA and GDS, which preserve the discriminative information among different classes, generating more efficient representations.

Accordingly, this network can operate on both labeled and unlabeled data, improving the performance when only small volumes of labeled data are available. This network is called dual flow subspace network (DFSNet) due to its flexibility in handling both learning paradigms. In addition to its advantages, semi-supervised learning is of theoretical interest since it makes it possible to understand the mechanisms of human learning.

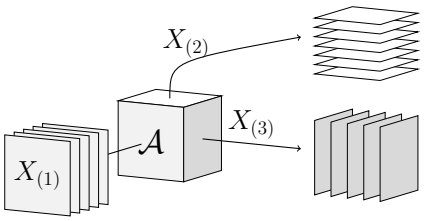


Figure 8: Illustration of the unfolding procedure of a 3-mode tensor. The unfolding of the 3-mode tensor $\mathcal{A}$ produces 3 sets of matrices $X_{(1)}$, $X_{(2)}$ and $X_{(3)}$.

In summary, using a semi-supervised neural network based on subspaces present the following contributions: 1) a new type of filter bank based on GDS. Unlike PCA, the filter banks produced by GDS can efficiently handle labeled data and 2) a semi-supervised shallow network based on PCA and GDS, presenting a flexible framework. Fig. 7 shows the flowchart of the semi-supervised neural network.

We presented the advantages of the proposed semi-supervised neural network over current shallow networks by experimental results using CIFAR−10 and ETH−80 databases for object recognition, LFW and FERET databases for face recognition and NYU Depth V1 database for scene recognition.

5.6 Experimental Results Summary

The experimental results show that by employing Fukunaga-Koontz transform for convolutional kernels, FKNet provides competitive classification results when compared to related shallow networks. To show its flexibility, FKNet was evaluated on a face verification task using the LFW dataset. In this experiment, FKNet was demonstrated to be competitive, where FVF, MBSIF-OB and other shallow networks were employed as baselines. The processing time measurement by the proposed network is efficient. For instance, CNN required about 3 hours to generate a 4 convolutional layers model using the EMNIST training dataset. On the other hand, FKNet obtained a comparable model using less than 17 minutes on the same hardware, which is approximately one order of magnitude faster.

It is also observed that one benefit of using the proposed network is that the number of convolutional kernels employed is much smaller than the ones used by a CNN. Besides, FKNet inherits the fast processing time exhibited by the shallow networks investigated, which is faster than the processing time obtained by CNN, suggesting that the proposed shallow network can replace CNN when processing time is a requirement [29, 30].

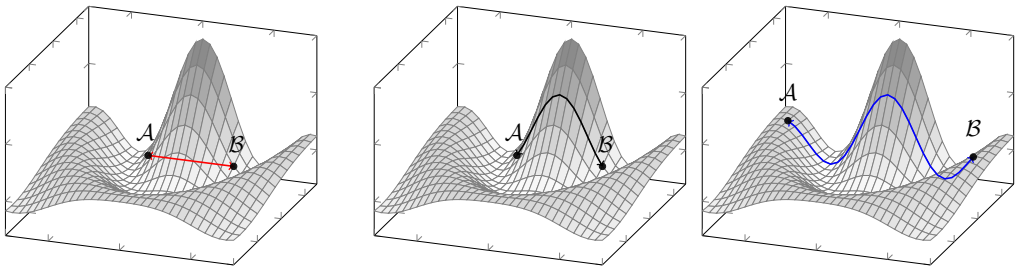
6 Tensor analysis based on subspaces

Tensors, which can be defined as a generalization of matrices, allow a natural representation of multidimensional data. For instance, video data is intuitively described by its correlated images over the time axis. Vectorization and concatenation of the video pixels may be applied to produce a practicable representation. However, it does not present a natural pattern representation. Also, the vectorization procedure may degrade the spatio-temporal relationship between pixels of a video tensor data, causing information loss.

The order of a tensor is linked to its dimensions, also known as ways or modes. Tensor unfolding is a procedure that reorganizes the tensor data to permit the analysis of each mode separately, possibly revealing correlations that were not immediately observed. This tensor unfolding procedure is shown in Figure 8. The tensor unfolding maneuver is relevant for the interpretability of the modes, as in medical image analysis.

Product Grassmann Manifolds (PGM) is one example of the use of subspaces to represent tensor data, as in action recognition problems [31, 32]. PGM extracts subspaces from tensor data and represents them as a point on the product space of n Grassmann manifolds, where each subspace corresponds to a point on one of the Grassmann manifolds. The classification is performed based on the chordal distance [33, 34] on the product manifold. Since PGM is a direct extension of MSM, it inherits the main disadvantage of MSM: absence of a discriminative mechanism.

We introduce the n-mode Generalized Difference Subspace projection (n-mode GDS), which can extract discriminative information from tensor data and provide suitable subspaces for tensor data classification. We employ the GDS projection, which acts as a feature extractor for MSM. Since GDS represents the difference among class subspaces, the GDS projection can increase the class subspaces' angles toward orthogonal status. Under this formulation, we can efficiently express tensor data as a point on a product manifold, simplifying the tensorial data representation and inheriting the main characteristics of GDS. Once the n-mode GDS is embodied into the product manifold, we can represent the relationship between all modes of a tensor in



(a) Euclidean distance. (b) Geodesic distance. (c) Effect of n-mode GDS.

Figure 9: The geodesic and the Euclidean distance on the PGM. (a) The Euclidean distance is the distance calculated when directly connecting $\mathcal{A}$ and $\mathcal{B}$. (b) Differently, the geodesic distance exploits the manifold surface, reflecting the actual distance between $\mathcal{A}$ and $\mathcal{B}$. (c) By employing the n-mode GDS projection, we improve the geodesic distance, since discriminative information is uncovered.

a unified design. Besides, we can go further and evaluate each mode separately, providing information to create a flexible measure of similarity [35].

7 n-mode Generalized Difference Subspace

Multi-dimensional data is usually represented by a set of modes (n-mode tensor) to reduce computational complexity. Given two n-mode tensors $\mathcal{A}$ and $\mathcal{B}$, we can formulate the tensor matching problem in two steps. First, we create a convenient representation, where $\mathcal{A}$ and $\mathcal{B}$ can be expressed in a compact and informative manner. Second, we establish a mechanism to produce a reliable measure of similarity between these representations, allowing the comparison of $\mathcal{A}$ and $\mathcal{B}$.

7.1 Tensor Representation by Subspaces

The tensors $\mathcal{A}$ and $\mathcal{B}$ present distinct properties in each mode. For instance, in video data, where $n = 3$, we have two spatial modes and a temporal one. Thus, each mode must be analyzed independently, according to its factors. To simplify this procedure, we employ the unfolding process. We denote by $\mathbf{X} = \{X_i\}_{i=1}^{n}$ the set of unfolded images corresponding to the mode-1, mode-2 and mode-3 unfolding of $\mathcal{A}$, respectively. The same procedure is conducted on the tensor $\mathcal{B}$, resulting in $\mathbf{Y} = \{Y_i\}_{i=1}^{n}$.

Eigen-decomposition can be exploited to derive a set of eigenvectors for each element of $\mathbf{X}$ and $\mathbf{Y}$. It is expected that the eigenvectors associated with the largest eigenvalues of each element of $\mathbf{X}$ and $\mathbf{Y}$ accurately represent their elements in terms of variance maximization [9]. After selecting these eigenvectors, we obtain the following sets $\mathbf{U}_X = \{U_i\}_{i=1}^{n}$ and $\mathbf{U}_Y = \{U_i\}_{i=1}^{n}$, respectively.

Now that we have $\mathbf{U}_X$ and $\mathbf{U}_Y$, which span the n-mode subspaces $\mathbf{P} = \{P_i\}_{i=1}^{n}$ and $\mathbf{Q} = \{Q_i\}_{i=1}^{n}$, we can employ a mechanism to extract more discriminative information from $\mathcal{A}$ and $\mathcal{B}$. Here, we should create a set of subspaces $\mathbf{D} = \{D_i\}_{i=1}^{n}$, whereby projecting the sets $\mathbf{P}$ and $\mathbf{Q}$. We adopt GDS [36] since it provides a reasonable balance between robustness and computational complexity, considering that it is mainly based on eigen-decomposition. Once we have projected the n-mode subspaces $\mathbf{P}$ and $\mathbf{Q}$ onto $\mathbf{D}$, we obtain the sets $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$. After we select the similarity function, we have the essential components to represent and measure the similarity between $\mathcal{A}$ and $\mathcal{B}$.

7.2 Generating the n-mode GDS Projection

In a m-class classification problem, $\mathbf{P} = \{P_{ij}\}_{i,j=1}^{n,m}$ denotes the set of all n-mode subspaces spanned by $\mathbf{U} = \{U_{ij}\}_{i,j=1}^{n,m}$. Then, we can now develop the n-mode GDS projection $\mathbf{D} = \{D_i\}_{i=1}^{n}$ that act on $\mathbf{P}$, to extract discriminative information. Since each mode subspace reflects a particular factor, it is essential to handle each one independently and compute a model that reveals hidden discriminative structures. In traditional GDS, this procedure is performed by removing the overlapping components that represent the intersection between the subspaces. In mathematical terms, the GDS projection can be described as extending the difference vector between two vectors in a multi-dimensional space.

Figure 9 shows the advantages of using the n-mode GDS projection on the PGM. To compute the n-mode GDS, we compute the sum of the projection matrices of each i-mode subspace as follows: $G_i = \frac{1}{m}\sum_{j=1}^{m} U_{ij}U_{ij}^{\top}$, for $1 \leq i \leq n$. Since G_i has information regarding all class subspaces in a particular mode, it is beneficial to decompose it to exploit discriminative elements. Applying eigen-decomposition

to G_i, we obtain: $G_i = V_i \Sigma_i V_i^\top$, for $1 \leq i \leq n$, where the columns in $V_i = \{\phi_1, \phi_2, \ldots, \phi_{R_i}\}$ are the normalized eigenvectors of G_i, and Σ_i is the diagonal matrix with corresponding eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_{R_i}\}$ in descending order, where $R_i = \text{rank}(G_i)$. We can define $D_i = \{\phi_{\alpha_i}, \ldots, \phi_{\beta_i}\}$, where $\alpha_i < \beta_i \leq R_i$. The n-mode GDS dimension is defined by maximizing the mean canonical angles between n-mode class subspaces.

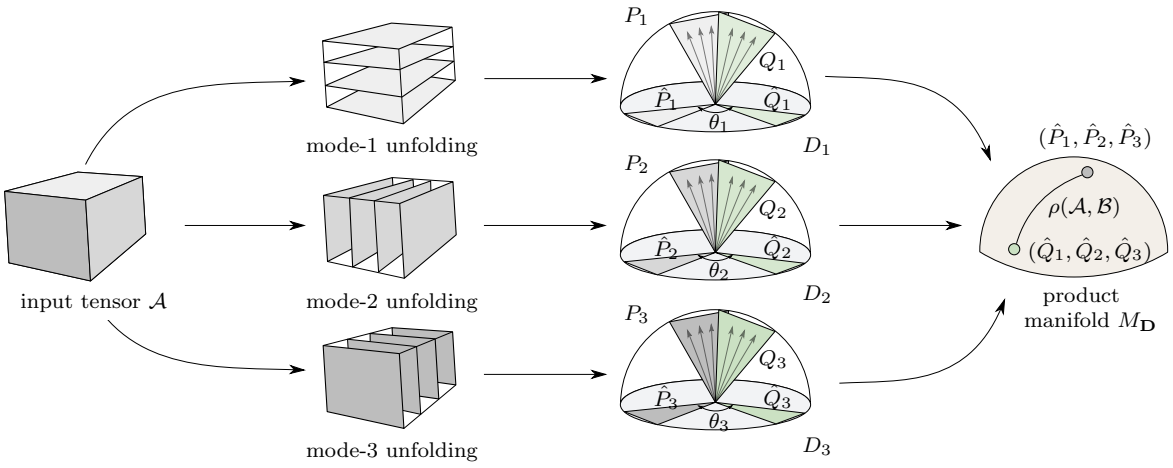


Figure 10: Conceptual figure of the n-mode GDS projection. First, we unfold the 3−mode tensor $\mathcal{A}$ and compute its subspaces from the unfolded modes. Then, we project the subspaces onto the n-mode GDS. The product of manifolds can be exploited to represent the projected subspaces. Finally, the chordal distance $\rho(\mathcal{A}, \mathcal{B})$ determines the similarity between tensors $\mathcal{A}$ and $\mathcal{B}$ [35].

7.3 Representing the n-mode Subspaces $\hat{\mathbf{P}}$ on the Product Manifold

We introduce the product manifold to describe $\hat{\mathbf{P}}$ into a single manifold. This manifold consists of the product of the projected n-mode subspaces onto the n-mode GDS. However, in traditional PGM, the subspaces are generated directly from the tensors by employing n-mode SVD. Although convenient, the generated subspaces may not be ideal for classification. In contrast, we project $\mathbf{P}$ onto $\mathbf{D}$ before applying the product manifold. Our objective is to achieve more efficient subspaces for classification. Therefore, given a set of manifolds $\mathbf{M} = \{M_i\}_{i=1}^n$ composed by $\hat{\mathbf{P}}$, Eq. (7) describes the product manifold:

$$M_{\mathbf{D}} = M_1 \times M_2 \times \ldots \times M_n = \left(\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_n\right), \tag{7}$$

where $\times$ denotes the Cartesian product, M_i is a i-mode manifold and $\hat{P}_i \in M_i$. Here, tensor data can be regarded as a point on the product manifold $M_{\mathbf{D}}$, as shown in Figure 10. A benefit of employing $M_{\mathbf{D}}$ is that it allows working directly with geodesics through the use of the geodesic distance. The geodesic distance between two points is the length of the geodesic path, which is the shortest path between the points that lie on the surface of the manifold. Once obtained the average canonical angles of all the available modes $\bar{\theta} = \{\bar{\theta}_i\}_{i=1}^n$ (see Eq. (3) and Eq. (4)), we can introduce the weighted geodesic distance based on the product manifold, which is defined as:

$$\rho(\mathcal{A}, \mathcal{B}) = \left(\sum_{i=1}^{n} \left(w_i \bar{\theta}_i\right)^2\right)^{1/2}, \tag{8}$$

where we estimate w_i by using the Fisher score since each mode will provide a different separability index reflecting the importance of each mode regarding classification.

7.4 Tensor learning by unsupervised subspaces

A typical example of tensor data in computer vision is observed in action analysis from video data, where both spatial and temporal information is present in a structured form. The spatial and temporal information can be handled independently within different representations in this scenario.

As mentioned before, tensors can be defined as a generalization of matrices, providing a natural representation of multi-dimensional data. For example, a video clip can be expressed by correlated images over the time axis.

Clustering has shown to be a valuable tool to reveal underlying data structures. A straightforward solution is to implement a clustering algorithm where vectorized tensor data are employed. However, such a solution usually provides poor clustering accuracy. It results in intractable computational times since data vectorization breaks the tensor's spatial and time structure (when available) [37].

After an extensive literature review, we could not find many works where the FKT projection was employed for unsupervised learning of tensor data. Besides, regularization on unsupervised tensor learning framework seems to be a novel topic, which may enhance clustering accuracy. Our main assumption is that the discriminant ability of the Fukunaga-Koontz transform is enhanced through the eigenspectrum regularization analysis.

Here we list some of the main advantages offered by the proposed method 1) Low-computational complexity representation for tensor data, which is inherited from the SVD decompositions. More precisely, the time complexity is linear according to the number of modes. 2) Low training data requirements; the compact subspace representation requires few patterns to express a complex category since the subspaces are linear combinations of the patterns, containing the eigenvectors and its linear combinations of all available images in a particular mode. 3) Flexibility for handling state-of-the-art handcrafted features.

Therefore, our contributions are as follows: 1) A new framework for tensor data clustering which provides flexibility to adapt to any existing clustering algorithm with low computational cost inherited from subspace learning. 2) An efficient eigenspectrum regularization scheme for multilinear clustering. 3) A new formulation of the mean between two tensors in terms of the product of spaces. 4) A Fisher score for unsupervised learning of tensorial data.

Fig. 11 shows the flowchart of the proposed framework. First, we use a tensor unfolding technique to represent the data in the subspace scheme. Next, the TFKT is applied to the subspaces, followed by the eigenspectrum regularization. The Fisher score adapted for handling tensor data is used for optimizing the cluster parameters. Finally, clustering is performed on the product of manifolds.

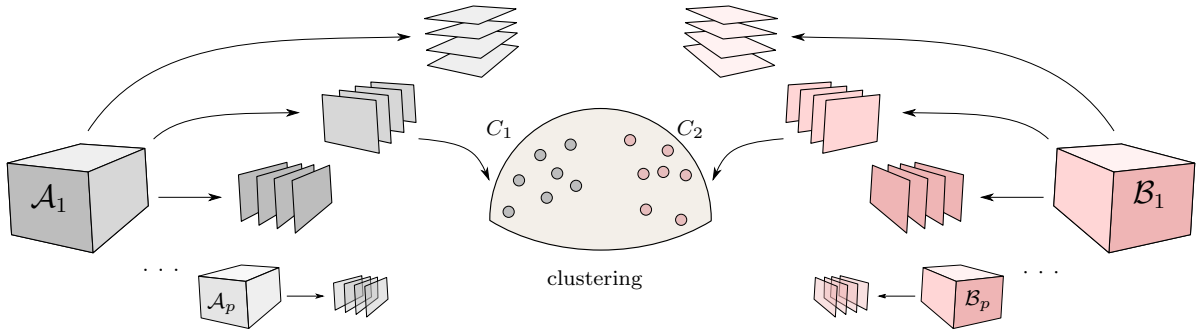


Figure 11: Conceptual figure of the proposed framework. First, the input tensors are unfolded and the n-mode SVD is applied on the unfolded patterns. The n-mode subspaces provided by the n-mode SVD are then projected on the n-mode TFKT, where discriminative features are extracted and the Fisher regularization process is performed. The n-mode Karcher mean is developed on the product of manifolds to support the k-means clustering.

The proposed method [38] is evaluated on datasets containing gestures and actions in videos. We also compare it with commonly used tensor clustering approaches for tensor data and subspace-based methods adjusted to handle tensor data. The obtained results demonstrate that the proposed clustering method presents advantages in terms of accuracy compared with conventional subspace-based methods for the tensor clustering task. Besides, the employed approach is efficient when handling scarce training data, which is beneficial in many applications.

We believe that other types of clustering algorithms could also be employed for future directions. For instance, spectral clustering, k-medoids, and DBSCAN could also benefit from a tensor representation, widening the range of applications. A straightforward application would be applying the proposed framework by replacing the k-means with k-medoids. This would improve the interpretability and explainability of the framework since, in some applications, the data average may not accurately reflect the physical properties of the observed object.

Other applications can benefit from TFKT flexibility, such as acoustic data. In this direction, pre-trained convolutional neural networks may also provide features to improve the TFKT performance further. We can also employ the TFKT projection matrices as an initialization scheme for neural networks, which may speed up the convergence while still refining the results.

7.5 Experimental Results Summary

We have evaluated the proposed approach on five video datasets containing human actions and compared its results with the results achieved by other state-of-the-art approaches. The experimental results have shown that the n-mode GDS outperforms conventional subspace-based methods on action recognition in terms of accuracy. Moreover, the proposed n-mode GDS does not require pre-training, which is an advantage in several applications where pre-trained models are scarce.

8 Conclusion and future work

Among other contributions, we have investigated the invariant properties of pattern-sets by their representations through subspaces. These invariances reflect the pattern sets physical properties and may be applied to represent and analyze many practical problems. Our studies have adopted a geometric framework in which the pattern sets statistical behavior is parametrized by subspaces. We handle pattern-sets as points in a metric space under this framework and analyze them using Grassmann geometry theories. The introduced methods present low computational complexity, simple implementation, and strong theoretical background.

We have examined the invariances of pattern-sets in terms of their subspaces in section 3. By this study, we concluded that the traditional subspaces (e.g., MSM and GDS) could not efficiently represent two-dimensional patterns without loss of information. Also, they cannot express temporal information, which is mandatory in gesture and action recognition from videos. To solve these issues, we introduced variants of subspace-based methods in which two-dimensional structures are well preserved. We also present the concept of Hankel subspace to express ordered sets of images. Our results present a recognition rate advantage compared to related methods, which are obtained in reduced processing time.

We presented a discriminative learning approach for shallow networks called Fukunaga-Koontz Network (FKNet). We employed a discriminative space developed using the Fukunaga-Koontz Transform to train shallow networks' weights. The results provided by FKNet were competitive with modern neural networks in image classification tasks. Its performance is also superior to the state-of-the-art shallow networks when the number of training samples is reduced. Encouraged by these results, we developed a semi-supervised shallow network. This semi-supervised approach is composed of a supervised and an unsupervised subspace, which shares the task of learning through datasets containing semi-supervised data. The results suggest that the proposed network is efficient even when both supervised and unsupervised data is limited.

We have also considered scenarios where data are not only in the form of matrices but also in the form of tensors. We developed a method named n-mode GDS to represent tensor data through discriminative subspaces, revealing information that was not available previously. We also introduced an optimization strategy to infer weights for the tensor modes to improve their efficiency in classification tasks.

Through the n-mode GDS results, we learned that the discriminative spaces produced by both GDS and FKT could be regarded as a discriminative unsupervised model. Inspired by these findings, we developed an unsupervised model. The goal here is to confirm whether the discriminative subspace provided by FKT is efficient for clustering tensor data. In this unsupervised learning scenario, k-means clustering is constructed on a product manifold, allowing the computation of distances between tensors.

It is worth noticing that our contributions are part of a compilation of efforts in the research community to design intelligent solutions to learn from a wide range of datasets even when high computational resources are not available. The introduced methods offer competitive results even under small sample size conditions and without the use of backpropagation. Also, we show that subspace methods are adjustable to operate on pre-trained models. In practice, these solutions may be seen as environmental-friendly since they reuse existing models, avoiding unnecessary energy consumption.

Future work: (1) Investigate new subspace representations to express text [39, 40], sound [41, 42], tables, trees, and graphs [43], to name a few. These variants can be readily applied to the subspaces framework and employ their benefits. (2) Develop new shallow networks using the theory of Lie groups which present a simple model for continuous symmetry, such as rotational symmetry found in three dimensions. (3) Introduce a deterministic neural network initialization by applying the convolutional kernels produced by FKT or GDS as an alternative to the random initialization process. (4) Propose an information fusion strategy by subspaces that may handle both sound subspaces [44, 45] and video subspaces in a unified framework.

References

[1] K.-i. Maeda, "From the subspace methods to the mutual subspace method," in *Computer Vision*. Springer, 2010, pp. 135–156.

[2] K. Fukunaga and W. L. Koontz, "Application of the karhunen-loeve expansion to feature selection and ordering," *IEEE Transactions on computers*, vol. 100, no. 4, pp. 311–318, 1970.

[3] L. S. Souza, N. Sogi, B. B. Gatto, T. Kobayashi, and K. Fukui, "Grassmannian learning mutual subspace method for image set recognition," *arXiv preprint arXiv:2111.04352*, 2021.

[4] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*. IEEE Computer Society, 1991, pp. 586–587.

[5] H. Akçay, "Spectral estimation in frequency-domain by subspace techniques," *Signal Processing*, vol. 101, pp. 204–217, 2014.

[6] S. Zhang, D. Wei, W. Yan, and Q. Sun, "Probabilistic collaborative representation on grassmann manifold for image set classification," *Neural Computing and Applications*, pp. 1–14, 2020.

[7] L. S. Souza, N. Sogi, B. B. Gatto, T. Kobayashi, and K. Fukui, "An interface between grassmann manifolds and vector spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 846–847.

[8] N. Sogi, L. S. Souza, B. B. Gatto, and K. Fukui, "Metric learning with a-based scalar product for image-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 850–851.

[9] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[10] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. Springer, 1992, pp. 162–190.

[11] K. Fukui and O. Yamaguchi, "The kernel orthogonal mutual subspace method and its application to 3d object recognition," in *Asian Conference on Computer Vision*. Springer, 2007, pp. 467–476.

[12] B. B. Gatto, H. Hino, and K. Fukui, "Block-based komsm for hand shape recognition with occlusion," in *proceedings of the Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV). Okinawa, Japan: IEEE*, 2014, pp. 1–8.

[13] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.

[14] B. B. Gatto, W. S. da Silva, and E. M. dos Santos, "Kernel two dimensional subspace for image set classification," in *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2016, pp. 1004–1011.

[15] B. B. Gatto and E. M. Dos Santos, "Image-set matching by two dimensional generalized mutual subspace method," in *Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 2016, pp. 133–138.

[16] B. B. Gatto, J. G. Colonna, E. M. dos Santos, and E. F. Nakamura, "Mutual singular spectrum analysis for bioacoustics classification," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

[17] L. S. Souza, B. B. Gatto, J.-H. Xue, and K. Fukui, "Enhanced grassmann discriminant analysis with randomized time warping for motion recognition," *Pattern Recognition*, vol. 97, p. 107028, 2020.

[18] F. Chatelin, *Eigenvalues of Matrices: Revised Edition*. SIAM, 2012, vol. 71.

[19] B. B. Gatto, E. M. dos Santos, and W. S. da Silva, "Orthogonal hankel subspaces for applications in gesture recognition," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2017, pp. 429–435.

[20] B. B. Gatto, A. Bogdanova, L. S. Souza, and E. M. dos Santos, "Hankel subspace method for efficient gesture representation," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

[21] K. Hotta, "Local co-occurrence features in subspace obtained by kpca of local blob visual words for scene classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3687–3694, 2012.

[22] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[23] A. I. Maqueda, C. R. del Blanco, F. Jaureguizar, and N. García, "Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns," *Computer Vision and Image Understanding*, vol. 141, pp. 126–137, 2015.

[24] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[25] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013.

[26] B. B. Gatto, E. M. dos Santos, K. Fukui, W. S. Júnior, and K. V. dos Santos, "Fukunaga–koontz convolutional network with applications on character classification," *Neural Processing Letters*, vol. 52, pp. 443–465, 2020.

[27] B. B. Gatto, L. S. de Souza, and E. M. dos Santos, "A deep network model based on subspaces: A novel approach for image classification," in *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*. IEEE, 2017, pp. 436–439.

[28] B. B. Gatto, L. S. Souza, E. M. dos Santos, K. Fukui, W. S. Júnior, and K. V. dos Santos, "A semi-supervised convolutional neural network based on subspace representation for image classification," *EURASIP Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 1–21, 2020.

[29] B. B. Gatto and E. M. dos Santos, "Discriminative canonical correlation analysis network for image classification," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4487–4491.

[30] B. B. Gatto, E. M. dos Santos, and K. Fukui, "Subspace-based convolutional network for handwritten character recognition," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1044–1049.

[31] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 833–839.

[32] Y. M. Lui, "Human gesture recognition on product manifolds," *Journal of Machine Learning Research*, vol. 13, no. Nov, pp. 3297–3321, 2012.

[33] G. H. Golub *et al.*, "Cf van loan, matrix computations," *The Johns Hopkins*, 1996.

[34] G. Stewart and J. Sun, "Computer science and scientific computing. matrix perturbation theory," 1990.

[35] B. B. Gatto, E. M. dos Santos, A. L. Koerich, K. Fukui, and W. S. Junior, "Tensor analysis with n-mode generalized difference subspace," *Expert Systems with Applications*, vol. 171, p. 114559, 2021.

[36] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2164–2177, 2015.

[37] B. B. Gatto, M. A. Molinetti, E. M. dos Santos, and K. Fukui, "Tensor fukunaga-koontz transform for hierarchical clustering," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 2019, pp. 150–155.

[38] B. B. Gatto, E. M. dos Santos, M. A. Molinetti, and K. Fukui, "Multilinear clustering via tensor fukunaga–koontz transform with fisher eigenspectrum regularization," *Applied Soft Computing*, vol. 113, p. 107899, 2021.

[39] E. K. Shimomoto, L. S. Souza, B. B. Gatto, and K. Fukui, "News2meme: An automatic content generator from news based on word subspaces from text and image," in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.

[40] ——, "Text classification based on word subspace with term-frequency," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[41] B. B. Gatto, J. G. Colonna, E. M. d. Santos, A. L. Koerich, and K. Fukui, "Discriminative singular spectrum classifier with applications on bioacoustic signal recognition," *arXiv preprint arXiv:2103.10166*, 2021.

[42] L. S. Souza, B. B. Gatto, and K. Fukui, "Grassmann singular spectrum analysis for bioacoustics classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 256–260.

[43] R. Mendoncaneto, D. Fenyo, Z. Li, E. F. Nakamura, F. G. Nakamura, and C. T. Silva, "A gene selection method based on outliers for breast cancer subtype classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[44] L. S. Souza, B. B. Gatto, and K. Fukui, "Classification of bioacoustic signals with tangent singular spectrum analysis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 351–355.

[45] B. B. Gatto, E. M. dos Santos, J. G. Colonna, N. Sogi, L. S. Souza, and K. Fukui, "Discriminative singular spectrum analysis for bioacoustic classification," in *INTERSPEECH 2020*. International Speech Communication Association (ISCA), 2020.