Subspace Projection Methods for Large Scale Image Data Analysis

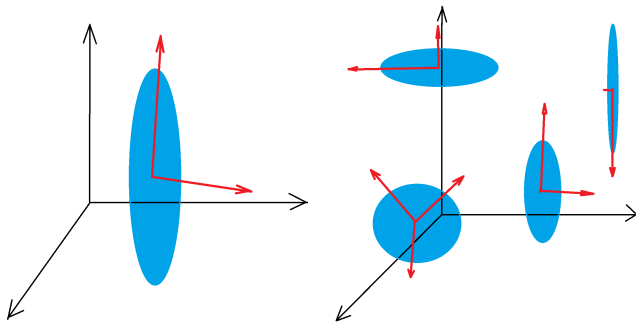
Ashish Gupta and Alper Yilmaz
College of Engineering
Ohio State University
Columbus, Ohio 43210
Email: {gupta.637, yilmaz.15}@osu.edu

***Abstract*—Images have become the most popular type of multimedia in the Big Data era. Widely used applications like CBIR underscore the importance of image understanding, especially in terms of semantically meaningful information. Typically, high dimensional image descriptors are embedded to a subspace using a simple linear projection. However, semantic information has a complex distribution in feature space that requires a non-linear projection. We first estimate an intrinsic dimensionality of image data. Next we build a measure of visual information in embedded subspace. We compare several linear and non-linear projection methods. We use multiple image databases towards a comprehensive evaluation in terms of information content, consequent recognition rates, and computational cost. This paper is relevant for researchers interested in dimensionality reduction for large scale image understanding that preserves semantically relevant information.**

I. Introduction

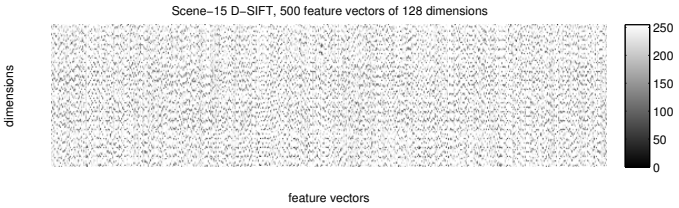
The volume of image data stored and shared across networks has grown exponentially in the Big Data era. This trend is expected to continue and also extends to other types of multimedia like videos. Similar to textual information, the visual content of an image is also being used as both structured query and retrieved information. These conditions require both efficient data management and quick analysis of large scale image data. Towards data management, the MapReduce framework designed by Google is very simple to implement and very flexible in that it can be extended for various large-scale data processing functions. This framework is a powerful tool to develop scalable parallel applications to process big data on large clusters of computers.

Towards image analysis, much effort has been made in the information retrieval (IR) field to find lower-dimensional representation of the original high-dimensional data, which enables efficient processing of a massive data set while preserving essential features [1]. Note that, image content needs to model complex visual concepts, and therefore IR methods need to be semantically relevant. Probabilistic topic models proved to be an effective way to organize large volumes of text documents [2]. In natural language processing, a topic model refers to a type of statistical model for representing a collection of documents by discovering abstract topics. This paradigm translates to modeling image data in terms of abstract visual categories. Each of the plethora of visual objects in any image can be associated with an appropriate visual category. A category itself is a node in a taxonomy of visual concepts. A popular example is ImageNet [4], that is inspired by WordNet for textual data. Unlike textual data, the principal caveat here is the large variation in feature description of visual objects associated with any category. The main reason is that any visual object is in turn composed of multiple parts [5], which are associated with different set of visual categories. In essence, every visual concept has a very complex feature description. Correspondingly, the set of descriptors spatially sampled from that image have a very complex distribution in feature space. Based on the work in: [6] that describes subspace clustering which is related to the concept of computing a visual model on multiple subspaces; and [7], that introduces a probabilistic subspace clustering approach that represents an image as a sparse combination of cluster exemplars embedded in a union of subspaces, our hypothesis is that descriptors from the same image occur in different regions of feature space and are embedded in different subspaces, depicted in Figure 1b. In order for IR methods to be effective, the lower-dimensional representation should preserve the nature of the complex distribution of an image. In other words, the subspace projection method should attempt to model visual concepts in feature space with

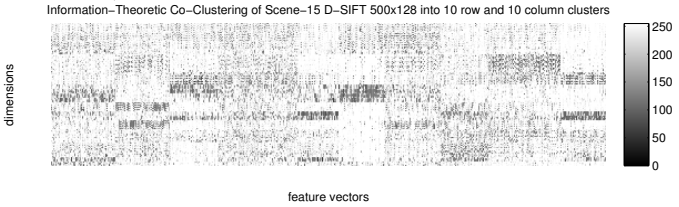


(a) (b)

Fig. 1: Typical subspace projection methods assume relevant features have global extent and compute an optimal subspace for them with a linear transformation between descriptors in high dimensional space to the subspace. We pursue a multiple subspace model for visual categories, which have limited extent and different embedded subspaces.



(a) Sample of 500 D-SIFT image feature descriptors, each of 128 dimensions, from Scene-15 dataset.



(b) Reordered rows and columns of the 500×128 feature vector matrix of Scene-15 data shown in 2a

Fig. 2: Visual concepts are embedded in multiple subspaces. We used Information-theoretic co-clustering [3] to compute 10 block clusters. The columns are instances of the feature descriptors and rows the dimensions. In each block, rows correspond to subspaces, whereas and columns correspond to spatial clusters. The color-bar shows the SIFT feature value ranging from 0 to 255. These values here are inverted for the sake of visibility.

different limited spatial extent and different subspaces. This insight motivates our work in this paper.

The incorporation of Principal Component Analysis (PCA) by [8] helped improve image classification and retreival results. However, researchers like [9], used an arbitrary choice of lower dimensionality. Their motivation was restricted to reducing processing time and removing dimensions associated with very low Eigenvalues. In practise researchers used an arbitrarily selected lower dimension. Instead, we estimate the intrinsic lower-dimensionality of image data. A popular survey on dimensionality reduction techniques was made in [10], but it is now dated and it was written well before the arrival of Big Data problems. Non-negative Matrix Factorization (NMF) was used for information retrieval in [11]. Note that PCA and SVD are special cases of NMF. The work in [12] analyzed a few subspace projection methods, but their focus was limited to evaluating method of discriminative projections versus principal components. In addition, it does not report computational cost, which is important for practitioners in Big Data and a major part the evaluation criteria in this paper. While simple and fast linear methods like PCA are traditional, subsequently developed non-linear projection methods like ISOMAP [13], LLE [14] have demonstrated better performance, but their proof was initially restricted to small toy problems in machine learning. Nevertheless, these methods suggested that relevant information regarding a visual concept exists in spatially local structures in the distribution of feature descriptors [15] and the importance of the preservation of these structures when embedded in a sub-manifold [16]. In [17], the authors offer a theoretical discussion, but are focused on only linear methods, whereas we consider both linear and non-linear methods with a thorough empirical evaluation. The authors in [18] and [19] write an explanatory survey which provides an overview of some linear and non-linear methods. In comparison we explore more recent and effective non-linear methods as well as a focus on visual concepts in image data.

Our contributions in this paper are as follows.

a) Subspace intrinsic dimensionality: We note that choice lower-dimensionality by practitioners, regardless of projection method, is typically ad-hoc and large since their rationale is reduce dimensionality by as little as possible and consider descriptor distribution in the entire feature space to have a single source. We adopt the rationale of descriptor distribution to have multi-region multi-subspace sources, where the intrinsic dimensionality of any one visual category is relatively small. Accordingly, we utilize multiple intrinsic dimensionality estimation techniques to validate our hypothesis. This work is described in Section II.

b) Subspace information measure: In Section III, we develop an entropy based measure for information in embedded subspace, inspired by work in [20] that explores methods for discovering 'hidden' structure in high-dimensional data. The quality of a projection method is related to the information it succeeds in preserving. We translate the mutual separation between all descriptors into a probability distribution and compute its Rényi entropy as information measure.

c) Subspace projection methods: In Section IV, we consider several linear and non-linear subspace projection methods. We evaluate these methods based on a criteria of information content, classification performance, and computational time in Section V. Our objective is a subspace projection method effective as preserving visual concepts, capable for being scaled to Big Data applications using images.

The rest of the paper is organized as follows. We discuss estimation of intrinsic dimensionality in Section II. In Section III, we describe our measure of information in embedded subspace. We discuss subspace projection methods in Section IV. The experiments and results are in Section V-D. We conclude in Section VI.

II. Subspace Dimensionality

In [12], the researchers use linear discriminant projections to evaluate sub-manifolds ranging from 128 to 30 dimensions; and report best comparative performance for 30 dimensions. Besides seeming incomplete, their estimate is not based on an inherent dimensionality leading to issues of training set bias. Since, we are interested in Big Data, we adopt a principled approach to the estimation of intrinsic dimensionality, that can be scaled to arbitrarily large dataset size. We consider the geometry of descriptor distributions by four different methods to understand variance in estimated dimensionality across visual concepts to finally compute the most likely estimate across all datasets.

a) Correlation Dimension: The intuition behind the correlation dimension estimator is that, the number of feature descriptors in a hyper-sphere of radius r is proportional to r^p, where p is the estimated dimensionality. Following from [21], the relative number of feature vectors $C(r)$ that lie within a hyper-sphere of radius r:

$$C(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1+1}^{n} \mathbb{1}_r \ , \ \mathbb{1}_r = \begin{cases} 1 & \text{if } \| \mathbf{z}_i - \mathbf{z}_j \| \leq r \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Since, $C(r)$ is proportional to r^p, it is used to estimate dimensionality p as $p = \lim_{r \to 0} \frac{C(r)}{\log r}$. Since, the limit cannot be explicitly solved, its value is approximated by computing $C(r)$ for two values of r. Approximated dimensionality $\hat{p}_{CorrDim}$ is the ratio of these two values give as:

$$\hat{p}_{CorrDim} = \frac{\log(C(r_2) - C(r_1))}{\log(r_2 - r_1)} \tag{2}$$

b) Maximum Likelihood Estimate: The Maximum Likelihood Estimate computes the number of feature descriptors $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ covered by a hyper-sphere with a growing radius r. This number is modelled as a Poisson process. Based on [22], the relation between the rate of the Poisson process $\lambda(t)$ and intrinsic dimensionality p is

$$\lambda(t) = \frac{f(\mathbf{z}) \pi^{p/2} p t^{p-1}}{\Gamma(1 + \frac{p}{2})} \tag{3}$$

where $f(\mathbf{z})$ is the sampling density. Intrinsic dimensionality p around $\mathbf{z}_i$ given k nearest neighbors is

$$\hat{p}_k(\mathbf{z}_i) = \Big(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(\mathbf{z}_i)}{T_j(\mathbf{z}_i)} \Big)^{-1} \tag{4}$$

where $T_k(\mathbf{z}_i)$ represents the radius of a hyper-sphere with center $\mathbf{z}_i$ that includes exactly k neighboring data points.

$$\hat{T}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{T}_k(\mathbf{z}_i) \tag{5}$$

The estimation of the intrinsic dimensionality $\hat{p}_{MLE}$ of data $\mathbf{Z}$ is obtained by averaging over the n local estimates $\hat{p}_k(\mathbf{z}_i)$.

c) Normalized Eigenvalue: Eigenvalue based intrinsic dimensionality estimation was proposed by [23]. After computing PCA on data $\mathbf{Z}$, we look at the normalized eigenvalues. The intrinsic dimension is determined by the number of these eigenvalues that are greater than a given small threshold value. This threshold value is determined empirically.

d) GMST Estimate: The Geodesic Minimum Spanning Tree (GMST) estimate is based on the observation that the length function of a geodesic minimum spanning tree is dependent on the intrinsic dimensionality p [24]. Similar to Isomap, the GMST estimator constructs a neighborhood graph $\mathcal{G}$ on the data $\mathbf{Z}$, in which every data point $\mathbf{z}_i$ is connected with its k nearest neighbors $\mathbf{z}_{ij}$. The GMST T is defined as the minimal graph over $\mathbf{Z}$, which has length $L(\mathbf{Z}) = \min_{T \in \mathcal{T}} \sum_{e \in T} g_e$, where $\mathcal{T}$ is the set of all sub-trees of graph $\mathcal{G}$, where e is an edge, and g_e is euclidean distance corresponding to edge e. We construct subsets $A \subset \mathbf{Z}$ of $\mathbf{Z}$ with various sizes m, and compute their lengths $L(A)$. The ratio $\log L(A) / \log m$ is linear and approximated by a function of the form $y = ax + b$. The estimated intrinsic dimensionality of $\mathbf{Z}$ is $\hat{p}_{GMST} = \frac{1}{1-a}$.

III. Subspace Information Measure

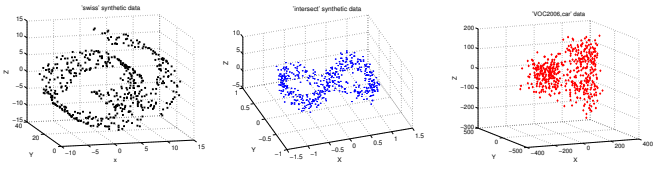
In order to measure information in embedded descriptors, we compute an entropy based measure on the mutual separation between these descriptors. First we compute a pair-wise distance all the descriptors. A histogram of these distances is normalized and posed as a probability distribution. Next we compute the Rényi entropy of this probability distribution. Rényi entropy was introduced as a generalization of Shannon entropy. It was developed to be the most general type of information measure, which preserves additivity of statistically independent systems and compatible with Kolmogorov's axioms of probability, [25]. For a discrete probability distribution $P = \{p_1, p_2, \ldots, p_n\}$ satisfying the conditions of $\sum_{i=1}^{n} p_i = 1$, and $p_i \geq 0$, $\forall 1 \leq i \leq n$, the Rényi entropy of order α is defined as

$$H_\alpha(x) = \frac{1}{1-\alpha} \log \Big(\sum_{i=1}^{n} p_i^\alpha \Big) \tag{6}$$

In the limit $\alpha \to 1$, Rényi entropy reduces to Shannon entropy. We motivate this measure in our work, using an illustrative example, shown in Figure 3. Consider an isometric distribution. It arguably lacks 'structure', in the sense that any two randomly selected subsets of the data will exhibit identical mutual separation statistics. A distribution with 'structure' deviates from such an isometric distribution. This property can be estimated by measuring the diversity in the set of all pair-wise distances. We consider the measure of diversity to correlate to a measure of 'structure' in the descriptor distribution. For example, the 'swiss-roll' in 3a and the 'intersect-loop' in 3b, wherein the structure is readily apparent. The third data distribution is a 1000 random sample of PCA embedded D-SIFT feature vectors of images labeled 'car' in the VOC-2006 database in 3c. A set of $\frac{n(n-1)}{2}$ pair-wise distances are computed for each of the $n = 500$ data points using the Euclidean distance metric. Subsequently, Rényi entropy with $\alpha = 2$ is computed for the normalized histogram (with 100 bins) of each distribution. In this experiment, the 'intersect'-loop distribution with $H_\alpha = -19.31$ is estimated to have more structure than the 'swiss-roll' distribution with $H_\alpha = -25.33$; and both more than the PCA embedded 'VOC-2006:car' distribution with $H_\alpha = -33.03$. Rényi entropy appears to successfully encode 'structure' in descriptor distribution.

IV. Subspace Projection Methods

Canonical dimensionality reduction seeks to project a p-dimensional feature descriptor vector $\mathbf{z} = [z_1, \ldots, z_p]^T$ to a lower dimensional representation of it, $\mathbf{s} = [s_1, \ldots, s_k]^T$ with $k \leq p$, based on some criterion. Typically, the criterion is preservation of some geometric structure, which can be global or localized to neighborhood of $\mathbf{z}$. The technique could utilize a linear or a non-linear projection function to compute $\mathbf{s}$. Based



(a) 'Swiss' (b) 'Intersect' (c) 'VOC-2006:car'

Fig. 3: Rényi entropy reflects 'structure' in descriptor distribution.

on these distinctions, the set of techniques considered here are grouped to traditional linear, global non-linear, local non-linear, and variants of local non-linear techniques.

A. Global Linear Techniques

1) Principal Component Analysis: PCA [8] constructs a low-dimensional representation of descriptors that describes as much of the variance in that descriptors as possible. It projects the data onto computed Eigenvectors with the greatest variance. The key benefits of PCA are its ease of implementation and speed of computation.

B. Global Non-linear Techniques

These attempt to preserve global properties of the descriptor distribution, while constructing non-linear projections.

1) Multi Dimensional Scaling: MDS [26] seeks to find an embedding to lower dimensional space such that distances between pairs of feature descriptors is preserved. The quality of the mapping is expressed in a 'stress function' - a measure of the error between the pairwise distances in the low- and high-dimensional representation. Given n data vectors, let the distance between $\mathbf{z}_i$ and $\mathbf{z}_j \in \mathbb{R}^p$ be $\delta_{i,j}$. The goal of MDS is to find vectors $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in $\mathbb{R}^k$, given Δ, such that $\parallel \mathbf{s}_i - \mathbf{s}_j \parallel^k \approx \delta_{i,j} \forall \mathbf{z}_i, \mathbf{z}_j \in \mathbf{Z}$.

2) Stochastic Proximity Embedding: SPE is an iterative algorithm that generates low-dimensional euclidean embeddings [27]. It attempts to preserve similarities between 'related' descriptors, by minimizing a 'stress function' used in MDS IV-B1. It uses an iterative pair-wise refinement strategy, which attempts to optimize the trade-off between preservation of local geometry and minimization of separation between distant vectors.

3) Isomap: Isomap [13] attempts to preserve pair-wise geodesic distance - the distance between two points on a manifold, which distinguishes it from methods which preserve pair-wise Euclidean distances. The geodesic distances between feature vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ are computed by constructing a neighborhood graph $\mathcal{G}$, in which every feature vector $\mathbf{z}_i$ is connected with its k nearest neighbors $\mathbf{z}_{ij}$. Dijkstra's algorithm is utilized to compute the shortest distance between two nodes on $\mathcal{G}$, which is a satisfactory approximation to the geodesic distance between corresponding feature vectors.

4) Diffusion Maps: Diffusion maps are based on defining a Markov random walk on the graph of the descriptors. The proximity of descriptors is inferred from aggregate results of random walk trials on the graph between nodes corresponding to these descriptors. The objective is retention of pair-wise diffusion distances between embedded descriptors [28]. Since diffusion distance is based on multiple paths through the graph, it is comparatively more robust to noise, than the geodesic distance.

5) Kernel PCA: Kernel PCA computes the principal eigenvectors of the kernel matrix, instead of the covariance matrix as in traditional PCA [29]. The application of PCA in kernel space provides Kernel PCA the property of constructing non-linear mappings.

C. Local Non-linear Techniques

Local non-linear embedding techniques aim to preserve distribution properties of local neighbourhoods around feature vectors.

1) Locally Linear Embedding: LLE is similar to Isomap, in that it constructs a graph, however unlike Isomap it embeds to a non-convex manifold [14]. Local distribution properties of the manifold are expressed as a linear combination of the nearest neighbors of each feature vector. A criterion of LLE is retention of 'reconstruction' weights in these linear combinations. The objective of LLE is to find low-dimensional global co-ordinates for feature vectors that lie on or near a manifold in high dimensional space.

2) Laplacian Eigenmaps: Laplacian Eigenmaps is a geometrically inspire embedding technique by [30]. The objective is minimization of distance between each embedded descriptor and its k nearest neighbors, mapped to a graph structure by solving an Eigenvector problem set up using the graph's Laplacian.

D. Variants of Local Non-linear Techniques

1) Locality Preserving Projection: Locality Preserving Projection (LPP) combines aspects of global linear and local non-linear embedding techniques [16]; computing a linear mapping that minimizes the cost function of the non-linear Laplacian Eigenmaps technique. It builds a graph $\mathcal{G}$ using neighborhood information of feature vectors. The edge weights of $\mathcal{G}$ are

$$w_{ij} = exp\Big(- \frac{\parallel \mathbf{z}_i - \mathbf{z}_j \parallel^2}{2\sigma^2} \Big) \tag{7}$$

Next, LPP solves the generalized Eigen problem

$$(\mathbf{Z} - \bar{\mathbf{Z}})^T L (\mathbf{Z} - \bar{\mathbf{Z}}) \upsilon = \lambda (\mathbf{Z} - \bar{\mathbf{Z}})^T M (\mathbf{Z} - \bar{\mathbf{Z}}) \upsilon \tag{8}$$

where L is the graph Laplacian, and M is the degree matrix. The eigenvectors υ_i corresponding to the k smallest non-zero eigenvalues form the columns of a linear mapping T that minimizes the Laplacian Eigenmap cost function [16]. The embedded vectors are expressed as $\mathbf{Y} = (\mathbf{Z} - \bar{\mathbf{Z}})^T$.

2) Neighborhood Preserving Embedding: Similar to LPP, Neighbourhood Preserving Embedding (NPE) [31] minimizes a cost function typical of a non-linear technique but using a linear mapping. NPE can be considered as a linear approximation to LLE.

(a) Pascal VOC 2007 (b) Scene-15 (c) Caltech-101

Fig. 4: Samples of visual categories in popular image databases used in this paper.

3) Probabilistic Principal Component Analysis: Probabilistic Principal Component Analysis [32] is a probabilistic model for PCA which combines local PCA models within the framework of a probabilistic mixture in which all the parameters are determined from maximum-likelihood using an EM algorithm.

4) Landmark Isomap: Landmark Isomap [33] is an extension of Isomap. An issue with Isomap is the costly global computation utilizing all feature vectors $\mathbf{Z}$. Landmark Isomap approximates this global computation, by a much smaller set of computations, by using a small subset of the data called landmark points.

5) Stochastic Neighborhood Embedding: Stochastic Neighborhood Embedding (SNE) [34] approximates a probability distribution of descriptors in high-dimensional space, with a probability distribution in the embedded space.

6) Symmetric Stochastic Neighbor Embedding: SymSNE [35] is a variation of Stochastic Neighbor Embedding.

7) t-SNE: An improvement over SNE, t-SNE [36] avoids 'crowding' issue, where moderately dissimilar feature vectors are huddled together in the embedding, by use of heavy-tailed Student t-distribution that allows moderate distances in the high-dimensional space to be projected to much larger distances in the embedding.

V. Experiments and Results

In this section we describe the datasets utilized in this paper, empirical results on estimation of intrinsic dimensionality of subspaces, and evaluation of subspace projection methods in terms of embedded information content, classification performance and finally computational time cost. The visual descriptor we used is densely sampled SIFT descriptor[37]. All our experiments are cross-validated using a training and test set split based on literature.

A. Image databases

We selected several popular databases that are frequently referred to in literature. They vary in number of available labeled visual categories and number of training and validation images. In our experiments we use the protocol associated with these datasets in terms of train, validation and test image partitioning. The databases utilized in this paper are Caltech-101 [38], Caltech-256 [39], Scene-15 [40], VOC-2006, VOC-2007, and VOC-2010 [41]. Details on these databases can be found in the cited references. An illustrative sample of visual categories in Scene-15 and VOC-2007 is shown in Figure 4. We computed Dense-SIFT on 8×8 pixel patches with step size of 4 pixels. For Scene-15, in accordance with [40], 100 images from each category are used as training set and the rest are used for testing. The train, validate, and test image sets are provided by the publishers of the VOC-2006, VOC-2007, and VOC-2010 datasets [42]. For Caltech-101 and Caltech-256, the training set was 30 randomly selected images, according to [39] and the remaining used as test images.

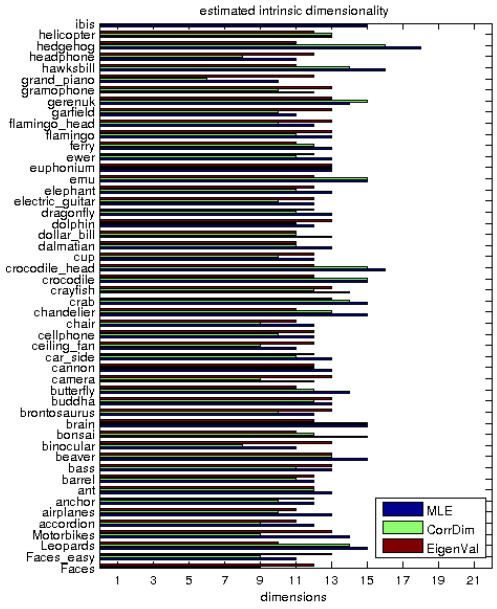


Fig. 5: Intrinsic dimensionality of visual categories in Caltech-101 image database. Typically, 'simpler' categories have a lower dimensionality.

B. Empirical intrinsic dimensionality

To compute intrinsic dimensionality using each of the 4 methods, we extracted a random sample of 10000 descriptors from the images in the training set and used a $10-fold$ cross-validation routine. The results are shown in 6, where the mean estimated intrinsic dimensionality for all databases is in the neighborhood of 14. The variance in dimensionality depicted

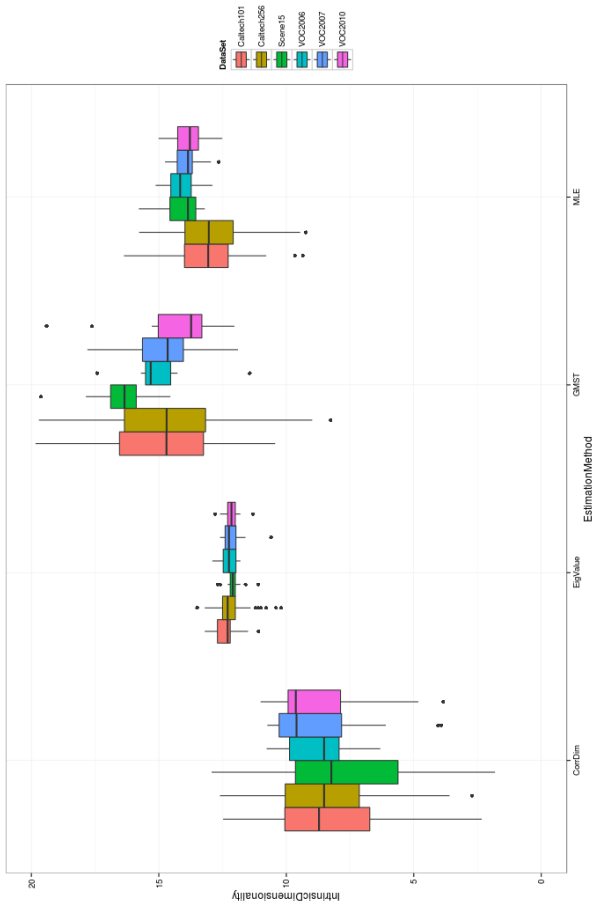


Fig. 6: Intrinsic dimensionality estimated using CorrDim, EigenValue, MLE, and GMST methods. The box-plot shows the mean and variance across visual classes in each dataset. The variance here reflects the difference in the structure of feature descriptor distribution across classes.

by the box-plot, is indicative of the difference between visual categories.

C. Subspace information measure

We use a random sample of 500 embedded descriptors of each image. Pair-wise distances are computed using the Minkowski metric[1], which is a generalization of the Euclidean metric, selected in part due to concern over the reliability of the Euclidean metric in higher dimensions, discussed by [43]. The parameter m, in this experiment, is assigned the value of the estimated intrinsic dimensionality of the visual category to which the image belongs. A normalized histogram of 100 bins is computed followed by Rényi entropy with $\alpha = 2$. The average entropic value of all visual categories for each database are listed in Table I. To compare the subspace methods, an aggregate entropic measure for all datasets is listed in the rightmost column. The variance in entropy is indicative that visual categories have an inherently different local descriptor distribution structure.

Note that LPP has a mean entropic measure of -4.54, which is promising in comparison to PCA, which is -24.43. Though both methods use linear embedding, LPP uses local structure, instead of global structure as used by PCA. Evidently, the global embedding of PCA loses more relevant information than the local embedding of LPP in comparison.

Method	Scene15	VOC2006	VOC2007	VOC2010	Average
Diff. Maps	-14.55	-14.56	-14.49	-14.52	*-14.55*
Factor Anal.	-33.53	-33.39	-33.44	-33.47	*-33.47*
Isomap	-26.92	-27.25	-27.24	-27.26	*-27.11*
L. Isomap	-26.11	-26.40	-26.41	-26.42	*-26.29*
LLE	-9.35	-8.16	-8.05	-8.06	*-8.69*
LPP	-5.23	-3.69	-3.69	-3.69	***-4.54***
MDS	-23.82	-23.69	-23.76	-23.77	*-23.79*
NPE	-9.35	-8.32	-8.32	-8.33	*-9.10*
PCA	-24.45	-24.40	-24.44	-24.44	*-24.43*
ProbPCA	-15.33	-15.38	-15.37	-15.37	*-15.33*
SPE	-11.45	-11.41	-11.48	-11.49	*-11.51*
SymSNE	-16.12	-16.04	-16.07	-16.08	*-16.10*
tSNE	-19.53	-19.37	-19.46	-19.44	*-19.45*

TABLE I: Rényi entropy of normalized distribution density of pair-wise distances in embedded space by different subspace methods for VOC- and Scene-15 databases.

D. Classification Performance Analysis

In this experiment we consider the Scene-15, VOC-2006, VOC-2007, and VOC-2010 databases. We first learn a visual model using random sampled images from training set. The model is computed using sparse coding with ℓ_1-norm and regularization parameter of $\lambda = 10$. The training routine is run for a maximum of $36,000$ iterations. The appropriate regularization parameter was determined empirically after a grid search for $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$. The maximum iterations is set sufficiently high to allow the optimization routine to converge within satisfactory tolerance. A similar parameter set is used in image encoding, with $\lambda = 10$. We use a SVM with RBF kernel and the performance is reported in II, shows the mAP for each database averaged across all the visual categories in it. The best performance is highlighted by bold text. The computational time for all the techniques is shown in figure 7. Some of the non-linear techniques utilized time of several orders of magnitude higher than the linear methods. For effective visualization the y-axis is on a log scale. The error bars denote the variation across categories, showing the minimum and maximum time utilized. The traditional linear methods PCA, LDA, and MDS are the fastest, as expected. The important outcome is that LPP which performed the best is also faster than other non-linear methods. This makes LPP stand out as a valid candidate to replace PCA in future.

[1]Minkowski metric $d_{\mathcal{M}}(\cdot,\cdot)$, with parameter m, between two vectors $k, l \in \mathbb{R}^p$, is $d_{\mathcal{M}}(k,l) = (\sum_{j=1}^{p} |k_j - l_j|^m)^{\frac{1}{m}}$

Method	Scene15	VOC-2006	VOC-2007	VOC-2010	Average
Diff. Maps	61.79	69.18	68.91	69.05	67.35
Factor Anal	62.68	70.38	68.86	68.44	67.54
Isomap	64.04	71.84	68.95	64.09	66.77
LLE	58.04	70.21	68.15	68.13	66.13
MDS	59.77	72.58	71.49	**70.32**	68.59
NPE	64.94	72.68	**73.15**	68.85	69.86
PCA	64.80	70.97	70.21	69.88	68.98
ProbPCA	62.77	70.91	71.80	69.514	68.87
LPP	**68.21**	**73.68**	72.75	69.14	**70.74**
SPE	63.13	64.59	70.1501	69.70	67.53
SymSNE	63.14	69.91	69.94	68.83	68.03
tSNE	56.58	63.91	69.1634	69.10	65.43

TABLE II: Comparative analysis of classification performance, measured in mAP, for the subspace methods aggregated over all the visual categories in each of the databases considered here.

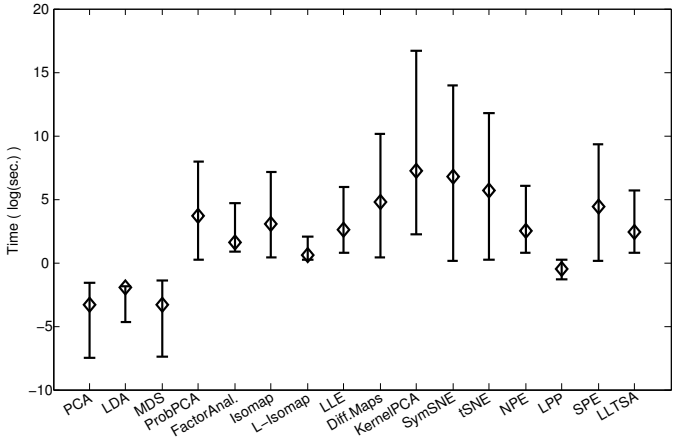


Fig. 7: Comparison of computation time of various techniques.

E. Computational Time Cost

In this experiment, we report results on the VOC-2006, VOC-2007, and VOC-2010 databases. The time taken to learn a subspace projection function for 10000 randomly selected feature vectors from training images of each dataset and subspace method is reported in Table III. The experiment was run using *Matlab* scripts [35] on *3.0 GHz Intel* processors with 48 GB of RAM. The average time across all databases is shown in the Aggregate column. As expected, linear methods, MDS and PCA, are found to be the fastest. Of interest here, is the cost of LPP, which is the well below other methods of its class, that preserve local structure. Although it is more complex than PCA, it is candidate for offline systems and more efficient than newer methods like SPE, t-SNE, which are orders of magnitude more costly. In addition, LPP uses a linear projective matrix, which makes it very fast during operation, once it has been trained.

VI. Summary

In this paper, we considered subspace projection methods for high-dimensional and large scale image data. We focused on developing a method that best preserved semantically relevant information, which we argued was embedded in different subspaces with varying local spatial extent. The hierarchical composition of a visual category in terms of other visual categories along with the properties of local patch image descriptor support our interpretation of the semantic information in images. Instead of an arbitrary lower-dimensionality, we computed an intrinsic dimensionality using multiple approaches for several databases. The mean and variance of these results indicate that image data has much lower dimensionality than had been used in literature. We developed an information measure for descriptor distribution in embedded subspace and used it to compare several subspace projection methods. Since, researchers and practitioners overwhelming use simple linear methods, our main objective was to build upon our intuition of multi-subspace, local-extent model of semantic information distribution in feature space. Consequently, we focused on non-linear methods that preserve mutual separation of descriptors in close proximity while ignoring the mutual separation of descriptors far from each other. Their comparison with linear methods like PCA was interesting since these methods consider the global distribution.

Subspace Method	VOC2006	VOC2007	VOC2010	Average
PCA	0.444	0.357	0.224	***0.233***
LDA	0.990	0.948	0.917	*0.643*
MDS	0.211	0.215	0.214	*0.148*
ProbPCA	861	875	880	*611*
FactorAnalysis	113	116	111	*79.8*
Isomap	7827	7987	7799	*5834*
LandmarkIsomap	217	222	223	*208*
LLE	813	890	757	*617*
DiffusionMaps	4926	5130	4881	*3783*
KernelPCA	52729	50741	68566	*39736*
SymSNE	126050	188320	179230	*142233*
tSNE	6397	28810	27941	*15407*
NPE	123	997	948	*505*
LPP	31.5	268	213	***132***
SPE	13142	22524	23141	*16038*

TABLE III: Relative computational time (seconds) of subspace projection methods on VOC- databases.

Since, this work is intended for Big Data, the computational complexity of each method and its ability to scale to large data was also important. Accordingly, we also analyzed the relative computational time of all the methods. Global linear method like PCA took 0.23 seconds in comparison to a global non-linear method like Kernel-PCA which took 39736.6 seconds. Regardless of their performance benefits, when applied to large scale learning problems, non-linear embedding methods are prohibitively slow. However, LPP took 132.5 seconds, which though higher than PCA, is comparatively much lower than other embedding techniques with comparable performance.

In conclusion, we have shown that semantically meaningful large scale image analysis requires a subspace projection method that emphasizes preservation of geometry of distribution of feature descriptors in close mutual proximity in the training phase to learn a linear projection that is quick

in the testing phase. While this paper focused on images, the inferences from this work can also be extended to other multimedia like videos. Since videos are often summarized in terms of key-frames, which is merely a set of images relevant to the video. Dimensionality reduction of video is equivalent to dimensionality reduction of the visual features of the set of key-frame images.

REFERENCES

[1] D. Markonis, R. Schaer, I. Eggel, H. Muller, and A. Depeursinge, "Using MapReduce for Large-Scale Medical Image Analysis," *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, no. August 2009, pp. 1–1, 2012.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, may 2003.

[3] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation," in *ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, vol. 8. New York, New York, USA: ACM Press, 2004, pp. 1–6.

[4] a.C. Berg, M. Maire, J. Malik, A. Berg, M. Maire, J. Malik, and R. Socher, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 248–255, jun 2009.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models." *IEEE Transacttions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–45, sep 2010.

[6] R. Vidal, "Subspace Clustering," *IEEE Signal Processing Magazine*, no. MARCH, pp. 52–68, mar 2011.

[7] A. Adler and M. Elad, "Probabilistic Subspace Clustering via Sparse Representations," in *ICML 2012 Workshop Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing*, Endinburgh, 2012, pp. 1–4.

[8] I. T. Jolliffe, *Principal Component Analysis*, second edi ed. New York: Springer, 2002.

[9] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE Comput. Soc, 2004, pp. 506–513.

[10] I. K. Fodor, "A survey of dimension reduction techniques," *Center for Applied Scientific Computing Lawrence Livermore National Laboratory*, vol. 9, no. 1, pp. 1–18, 2002.

[11] S. Tsuge and M. Shishibori, "Dimensionality Reduction using Non-Negative Matrix Factorization for Information Retrieval," *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 2, pp. 960–965, 2001.

[12] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 338–52, feb 2011.

[13] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction." *Science (New York, N.Y.)*, vol. 290, no. 5500, pp. 2319–23, dec 2000.

[14] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding." *Science (New York, N.Y.)*, vol. 290, no. 5500, pp. 2323–6, dec 2000.

[15] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis." *IEEE Transacttions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1610–26, sep 2010.

[16] X. He, S. Yan, Y. Hu, and H.-j. Zhang, "Learning a locality preserving subspace for visual recognition," *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 385–392 vol.1, 2003.

[17] J. P. Cunningham and Z. Ghahramani, "Linear Dimensionality Reduction: Survey, Insights, and Generalizations," *Journal of Machine Learning Research*, vol. 16, pp. 2859–2900, 2015.

[18] C. O. S. Sorzano, J. Vargas, and a. P. Montano, "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*, pp. 1–35, 2014.

[19] A. Sarveniazi, "An Actual Survey of Dimensionality Reduction," *American Journal of Computational Mathematics*, vol. 04, no. 02, pp. 55–72, 2014.

[20] D. Scott and S. Sain, "Multidimensional Density Estimation," *Handbook of Statistics*, vol. 24, no. August 2004, pp. 229–261, 2005.

[21] F. Camastra and A. Vinciarelli, "Estimating the Intrinsic Dimension of Data with a Fractal-Based Method," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, 2002.

[22] E. Levina and P. J. Bickel, "Maximum Likelihood Estimation of Intrinsic Dimension," in *Neural Information Processing Systems*, 2005, pp. 777–784.

[23] F. Camastra, "Data dimensionality estimation methods: a survey," *Pattern Recognition*, vol. 36, no. 12, pp. 2945–2954, dec 2003.

[24] B. Poczos and A. Lorincz, "Independent Subspace Analysis Using Geodesic Spanning Trees," in *International Conference on Machine Learning*, Bonn, 2005, pp. 1–8.

[25] J. C. Principe and D. Xu, "Information-theoretic learning using renyi's quadratic entropy," in *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation*, 1999, pp. 407–412.

[26] T. F. Cox and M. A. A. Cox, *Multidimensional scaling*, 2nd ed. Chapman and Hall/CRC, Sep. 2001.

[27] D. K. Agrafiotis, "Stochastic Proximity Embedding," *Journal of Computational Chemistry*, vol. 24, pp. 1215–1221, 2003.

[28] S. Lafon and A. B. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization." *IEEE Transacttions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1393–403, sep 2006.

[29] H. Hoffmann, "Kernel pca for novelty detection," *Pattern Recognition*, vol. 40, no. 3, pp. 863–874, mar 2007.

[30] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," in *Neural Information Processing Systems*. MIT Press, 2002, pp. 585—-591.

[31] X. He, S. Yan, and H.-j. Zhang, "Neighborhood Preserving Embedding," in *IEEE International Conference on Computer Vision*. IEEE Computer Society, 2005, pp. 1–6.

[32] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers." *Neural computation*, vol. 11, no. 2, pp. 443–82, feb 1999.

[33] V. D. Silva and J. B. Tenenbaum, "Global Versus Local Methods in Nonlinear Dimensionality Reduction," in *NIPS*, 2003, pp. 721–728.

[34] G. Hinton and S. Roweis, "Stochastic Neighbor Embedding," in *Neural Information Processing Systems*, 2002, pp. 833–840.

[35] L. J. P. van der Maaten and L. van der Maaten, "An Introduction to Dimensionality Reduction Using Matlab," Universiteit Maastricht, Maastricht, Netherlands, Tech. Rep. July, 2007.

[36] L. V. D. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[37] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[38] L. Fei-fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, apr 2007.

[39] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007.

[40] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *IEEE Computer Vision and Pattern Recognition*. Ieee, 2006, pp. 2169–2178.

[41] M. Everingham, L. V. Gool, C. K. I. Williams, and J. Winn, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results," http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

[43] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in *Database Theory - ICDT 2001*, vol. 1973. London: Springer-Verlag Berlin, Heidelberg, 2001, pp. 420–434.