EVALUATING DIMENSIONALITY REDUCTION TECHNIQUES FOR VISUAL CATEGORY RECOGNITION USING RÉNYI ENTROPY

Ashish Gupta, Richard Bowden

C.V.S.S.P., University of Surrey
Guildford, GU2 7XH, Surrey, United Kingdom
phone: + (44) 1483 682260, fax: + (44) 1483 686031, email: a.gupta@surrey.ac.uk
web: http://www.ee.surrey.ac.uk/cvssp

ABSTRACT

Visual category recognition is a difficult task of significant interest to the machine learning and vision community. One of the principal hurdles is the high dimensional feature space. This paper evaluates several linear and non-linear dimensionality reduction techniques. A novel evaluation metric, the rényi entropy of the inter-vector euclidean distance distribution, is introduced. This information theoretic measure judges the techniques on their preservation of structure in lower-dimensional sub-space. The popular dataset, Caltech-101 is utilized in the experiments. The results indicate that the techniques which preserve local neighborhood structure performed best amongst the techniques evaluated in this paper.

1. INTRODUCTION

The topic of visual category recognition is of considerable interest to researchers in machine learning, computer vision and pattern recognition. It is a challenging task due to the complexity of data utilized to learn efficient classifiers [21]. There are numerous approaches employed from state-of-the-art algorithms in machine learning [28]. The objective is recognition of visual objects in images in terms of their visual category. Consequently, in addition to the issues of pose and illumination variance, background clutter, and partial occlusion, successful approaches must contend with significant intra-category appearance variation. This requires large datasets, which are generally poorly annotated, leading to noisy training data. Significant progress towards recognition has been achieved in the past few years by researchers working to improve the components of a typical visual categorization approach like feature descriptors, clustering methods, contextual information, visual models, and classifiers. However, the topic of dimensionality reduction of feature descriptors for that task has not received due attention. Researchers have, as yet continued to utilize the simplest linear dimensionality methods like PCA [12]. This is perhaps due to the computational efficiency and popularity of PCA and its demonstrated success [13]. However, independently there has been considerable progress in the development of dimensionality reduction, especially non-linear techniques, which has not been sufficiently explored for improving visual categorization. Consequently, the focus of this paper is quantitative exploration of such techniques to aid in visual categorization. The principal issues are: the curse of dimensionality faced by the clustering method used to compute the visual codebook model [1]; considerably noisy feature data due to lack of sufficient annotated training images [22]; and inherently different lower dimensional sub-spaces for different visual categories.

The issues with visual datasets and the nature of the visual categorization task together presents a challenge which can benefit from the properties of dimensionality reduction. The motivation for dimensionality reduction is to:

- project the data to a lower dimensional space to ameliorate the problem of high dimensionality faced by a distance metric.
- reduce the effects of noise by exclusion of feater descriptor dimensions irrelevant to the visual category.
- discover a visual category specific lower dimensional sub-space.

The challenge of visual categorization is:

- noise in data due to lack of annotation and background clutter.
- use of affine invariant local image patch descriptors like SIFT[16] which results in a high dimensional feature space not amenable to a Euclidean distance metric utilized by popular clustering techniques.
- significant intra-category appearance variation and use of low-level image descriptors which results in semantically equivalent descriptor vectors being scattered in feature space rather than being clustered together.

The lack of visual category specificity of these descriptors is solved by computing their occurrence histograms, an idea inspired by the 'bag-of-words' model used successfully in text-document retrieval, which is the basis of the popular 'bag-of-features' model (BoF) [6]. The histogram of a visual category is called a visual codebook. The mathematical formulation of the BoF model is provided in 2.1. In brief, it is computed by a LVQ method, popularly k-means clustering, which encodes the feature space. The classifier is trained on the histogram feature rather than the original low-level descriptor. Therefore, the performance of the classifier depends on the information content in the visual codebook after compression of feature space. A hurdle is the 'curse of dimensionality' which renders the Euclidean distance metric ineffective for high dimensions (the SIFT descriptor for example is of 128 dimensions). Therefore, the dimensionality reduction method must project feature space to a lower dimensional sub-space with the aim of maximizing information content in the sub-space. Consequently, an entropic measure of information is utilized in this paper to evaluate the efficacy of various dimensionality reduction techniques. Rényi entropy is selected as the entropy measure, since it: is a generalization of Shannon entropy, which is a family of functionals for quantifying the diversity in a data distribution; admits a generic closed form expression for distributions belonging to the same exponential family [20] (exponential family is a unifying framework for many common distributions like Gaussian, multinomial, Beta); and varies monotonically with information and so can be used interchangeably with it. The key contributions of this paper are:

- quantitative evaluation of linear and non-linear dimensionality reduction methods for visual category recognition.
- use of an entropic measure from information theory to evaluate the efficacy of dimensionality reduction for computing visual codebooks.
- introduction of rényi entropy of inter-vectors distance distribution as evaluation measure.

A background to visual category recognition using BoF model, dimensionality reduction techniques evaluated in this paper, rényi entropy, the dataset, and the feature descriptor utilized is provided in section 2. The experiment for estimation of intrinsic dimensionality of categories is described in section 3. The motivation for the evaluation metric and the experimental setup is described in section

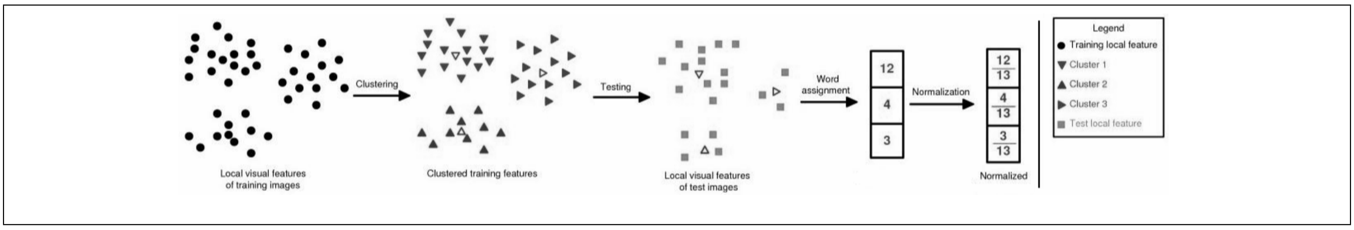

Figure 1: Bag-of-Features technique for computing visual codebook

4. The results for the experiments comparing the techniques is presented in section 5. An analysis of the results and future work is discussed in section 6, along with a summary of the paper.

2. BACKGROUND

This section discusses the BoF model, the dimensionality reduction techniques evaluated in this paper, and rényi entropy.

2.1 Bag-of-Features Model

The BoF model is a basic component of most of the approaches towards visual categorization, and consequently the focus of this paper. It seeks to encode feature data in a high-dimensional manifold using a set of reference or codebook vectors, which are most representative of the training data. An illustrative example is shown in figure 1. The training data is a set of K images, each represented by a set of feature points. Each such set can be denoted by $\{X^1, X^2, \ldots, X^K\}$, where $X^k = \{f_1^k, f_2^k, \ldots, f_{N_k}^k\}$. Each feature point is defined by its appearance descriptor $f_i^k \in \mathbb{R}^D$, where D is the dimensionality of the appearance descriptor. The total number of feature points in a dataset are $N = \sum_{k=1}^{K} N_k$, where N_k are the number of points in the k^{th} image. The feature vectors f lie in a sub-manifold $V \subseteq \mathbb{R}^D$. BoF utilizes a vector quantization technique to encode V using a finite set $C = \{c_1, \ldots, c_M\}$ of codebook vectors (cluster-centers or visual words), $c_i \in \mathbb{R}^D, i = 1, \ldots, N$. A data-vector $f \in V$ is described by the 'winning' reference vector c_i of C for which the distortion error $d(f, c_i)$ is minimum. The euclidean metric is generally selected as a distortion measure, so $d(f, c_i) = \| f - c_i \|^2$. This technique splits the manifold V into hyper-cells $V_i = f \in V \,||\, \| f - c_i \| \leq \| f - c_j \| \, \forall j$, in which each data vector f is described by its corresponding reference vector c_i. If the probability distribution of data vectors over the manifold V is described by $P(f)$, then the average distortion error is determined by: $E = \int d^D f P(f)(f - c_i)^2$, which is minimized by a search for an optimal set of reference vectors c_i. A histogram of data vector association to reference vectors is computed. The histogram, $H = \{h_1, h_2, \ldots, h_M\}$, where $h_i = |V_i|$. This histogram is often normalized and is characteristic of the data X^K. It is utilized as a feature vector for training a classifier like a SVM.

2.2 Dimensionality Reduction

Dimensionality reduction can be mathematically formulated as: consider a p-dimensional feature descriptor vector $\bar{x} = \{x_1, \ldots, x_p\}^T$. For $\bar{x}$ compute a lower dimensional representation of it, $\bar{s} = \{s_1, \ldots, s_k\}^T$ with $k \leq p$, that captures the content in the original data, based on some criterion. In brief, the methods considered in this paper are:

1. Principal Component Analysis (PCA)[12]: It projects the data onto the eigenvectors with the greatest variance.
2. Linear Discriminant Analysis (LDA)[27]: It preserves as much of the class discriminatory information as possible.
3. Multi-Dimensional Scaling (MDS)[5]: It seeks to find an embedding to lower dimensional space such that distances between data vectors is preserved. Mathematically, given M data vectors, the distance between x_i and $x_j \in \mathbb{R}^N$ is $\delta_{i,j}$. The goal of MDS is, given Δ, to find vectors $s_1, \ldots, s_M$ in $\mathbb{R}^P$ such that $\| s_i - s_j \|^P \approx \delta_{i,j} \forall x_i, x_j \in M$.
4. Probabilistic Principal Component Analyzers [26]: It 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.
5. Factor Analysis [8]: It originates from the field of psychology. It assumes that the data distribution is sourced from underlying 'factors'. The methods attempts to estimate this factors and thereby reduces dimensions of the data.
6. Isomap [25]: It overcomes the issues with traditional linear scaling methods which attempt to conserve pair-wise euclidean distances. Instead it attempts to preserve pair-wise geodesic distance, which is the distance between two points on a manifold.
7. Landmark Isomap (L-Isomap) [24]: It is a method for approximating a large global computation in Isomap by a much smaller set of calculations. The method focuses on a small subset of the data, called the landmark points.
8. Locally Linear Embedding (LLE) [23]: It aims to find low-dimensional global co-ordinates for data that lie on or near a manifold embedded in high dimensional space $\mathbb{R}^N$. The three parts to LLE: find the k nearest neighbors for each $x_i \in \mathbb{R}^N$; find matrix w which minimizes the residual sum of squares for reconstructing each x_i from its neighbors, where $RSS(w) \equiv \sum_{i=1}^{n} \| x_i - \sum_{j \neq i} w_{ij} x_j \|^2$; and compute low-dimensional coordinates Y best reconstructed to minimize cost function $\omega(Y) = \sum_i (y_i - \sum_{j=1}^{k} w_{ij} y_i)^2$.
9. Diffusion Maps [14]: It builds a map between data points and computes all possible paths between point through the graph. The lower dimensional projection aims to retain the best possible pair wise diffusion distance.
10. Kernel PCA [11]: It computes the principal eigenvectors of the kernel matrix, instead of the covariance matrix as in traditional PCA.
11. Symmetric Stochastic Neighbor Embedding (SymSNE) [18]: It is a variation of Stochastic Neighbor Embedding.
12. t-distributed Stochastic Neighbor Embedding (tSNE) [17]: It is an improvement on Stochastic Neighbor Embedding. It is particularly important for high-dimensional data that lie on several different, but related, low-dimensional manifolds.
13. Neighborhood Preserving Embedding (NPE) [10]: It aims to preserve local neighborhood structure and is less sensitive to outliers compared to PCA.
14. Locality Preserving Projection (LPP) [9]: It builds a graph G using neighborhood information of the data vectors. Using the Laplacian of this graph $L(G)$, transformation matrix A is computed, which maps the data points to a subspace, $x_i \rightarrow y_i = A^T x_i$. This linear transformation optimally preserves local neighborhood information in a certain sense.
15. Stochastic Proximity Embedding (SPE) [2]: It generates low-dimensional euclidean embeddings which attempt to preserve the similarities between related feature vectors. The embedding is carried out using an iterative pair-wise refinement strategy that attempts to preserve local geometry while trying to maintain minimum separation between distant vectors.

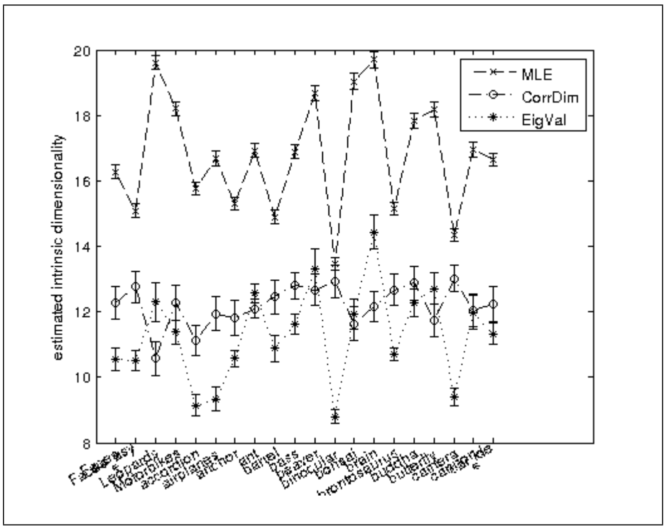

Figure 2: Estimation of intrinsic dimensionality of visual categories in the Caltech-101 dataset, using (i)MLE,(ii)Correlation dimensional, and (iii) Eigenvalue, based methods.

16. Linear Local Tangent Space Alignment (LLTSA) [29]: It uses the tangent space in the neighborhood of a feature vector to represent the local geometry. It then aligns these local tangent spaces in the lower dimensional space. The feature space is projected linearly to the lower dimensional space.

2.3 Rényi Entropy

Rényi entropy is an extension of the commonly known Shannon entropy, which is a measure of information in a system. Shannon defined entropy H of a distribution with probability $p_i[i]_1^N$ as:

$$H(p) = -K \sum_{i=1}^{N} p_i ln p_i$$

where K is a positive constant. Rényi extended Shannon entropy to a continuous family of entropy measures that obey:

$$H_\alpha(p) = \frac{1}{1-\alpha} ln \sum_{i=1}^{N} p_i^\alpha$$

where Rényi entropy tends to Shannon entropy as $\alpha \to 1$. Rényi entropy is a monotonic function of the information, which implies that they can be used interchangeably in any practical application.

2.4 Dataset

The dataset used in this paper is one of the most popular and benchmark datasets Caltech-101 [7]. There are a total of 9146 images, split between 101 different object categories, as well as an additional background/clutter category. Some other popular datasets that could alternatively have been used for this paper are Caltech-256 (256 categories), Pascal VOC 2007-10 (20 categories),and Graz (3 categories). Caltech-101 was selected for its popularity and appropriate number of categories for this task.

2.5 Feature Descriptor

The feature descriptor utilized in this work is the Scale Invariant Feature Transform (SIFT)[16]. It belongs to the class of affine invariant feature descriptors which were designed for the task of wide-baseline stereo matching. It performs remarkably well for image matching under significant affine variation, partial occlusion and background clutter. A comparative performance analysis can in found in [19].

3. INTRINSIC DIMENSIONALITY ESTIMATION

The choice of dimension of the lower dimensional sub-space is based on the intrinsic dimensionality of the visual category. There are numerous estimation methods in the literature. Three methods utilized in this paper are:

- eigenvalue[4]: It is based on global or local PCA. The intrinsic dimension is determined by the number of eigenvalues greater than a given threshold.
- correlation dimension [3]: It is a fractal based technique to estimate the attractor dimension of the underlying dynamic system. It is defined as: let $\Omega = \{x_1, x_2, \ldots, x_N\}$ be a set of points in $\mathbb{R}^n$ of cardinality N . If the correlation integral $C_n(r)$ is defined as: $C_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{1} \| x_i - x_j \| \leq r$, then the correlation dimension D of Ω is: $D = lim_{r \to 0} \frac{ln(C_m(r))}{ln(r)}$.
- maximum likelihood estimate [15]: It estimates the intrinsic dimensions $\hat{m}$ for $\{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^d, m \leq d$. Based on the distance of each data point x_i from its k neighbors, the estimated dimension is: $\hat{m}_k(x_i) = [\frac{1}{k-2} \sum_{j=1}^{k-1} log(\frac{T_k(x_i)}{T_j(x_i)})]^{-1}$, where $T_k(x_i)$ is the distance of x_i to it's k nearest neighbors. The intrinsic dimensionality is the average over all observations: $\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(x_i)$.

In the experiment, a random set of 2000 data-vectors was sampled from each category in Caltech-101. This experiment was repeated for 100 trials. The average and standard deviation of the estimated dimensionality by each of the methods for each of the categories in shown in figure 2 (a representative set of 25 categories is shown due to lack of space). The data marker denotes the average dimensionality and the error bars denote the standard deviation in the estimated value. Although each method estimates a different absolute value for a category, they can be seen to follow the same trend. The relatively small error bars indicates that the sample size of 2000 vectors is a sufficient sample to estimate the nature of distribution of the feature vectors. This result highlights the benefit that can be found in category specific dimensionality reduction. Since, the objective of this work is a comparative evaluation, an average is computed across all categories. The average and standard deviation of dimensionality across all methods and categories is 13.0786 and 0.0817 respectively. Accordingly, rounding the average intrinsic dimensionality, the lower dimensionality used in this paper will be 13 dimensions.

4. EXPERIMENT

The section explains the evaluation metric used for comparison of the techniques and the experimental setup.

4.1 Evaluation Metric

The information content in the distribution of feature vectors is measured using rényi entropy of the pair-wise distance distribution of these feature vectors. Pair-wise distance is important because: all the dimensionality reduction methods try to preserve this quantity in different ways; it determines the visual codebook created by clustering. The relevance of this metric is illustrated using two synthetic data distributions shown in figure 3. The first (A) is a helix where the distribution has a self-evident structure, and the second (B) is a random distribution with no discernible structure. The histogram of pair-wise distances for data vectors for both these distributions is shown in (C). The unimodal histogram of the distribution (B) has a lower rényi entropy of -21.8477, compared to the helix which has a entropy of -16.8132.

4.2 Setup

The feature descriptor data for all the training images for every category in Caltech-101 is computed using SIFT [16]. The dataset has different number of sample images for each category. To introduce uniformity of comparison, 2000 feature vectors are randomly

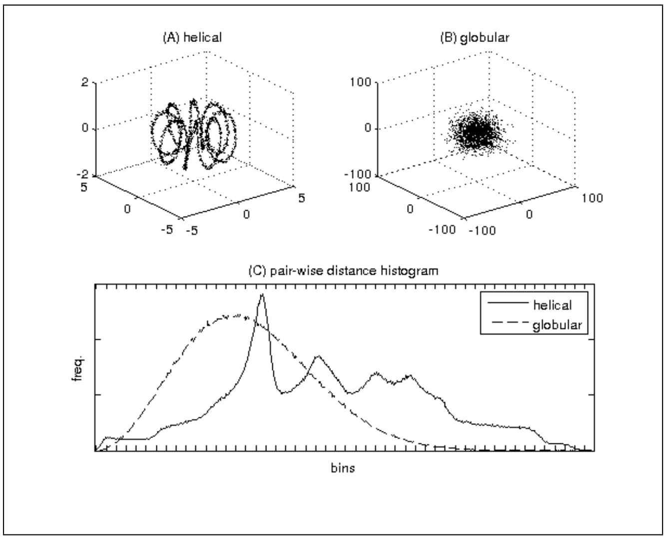

Figure 3: Illustrative example of relation between inter-vector distance distribution histogram, entropic measure, structure in visual category data.

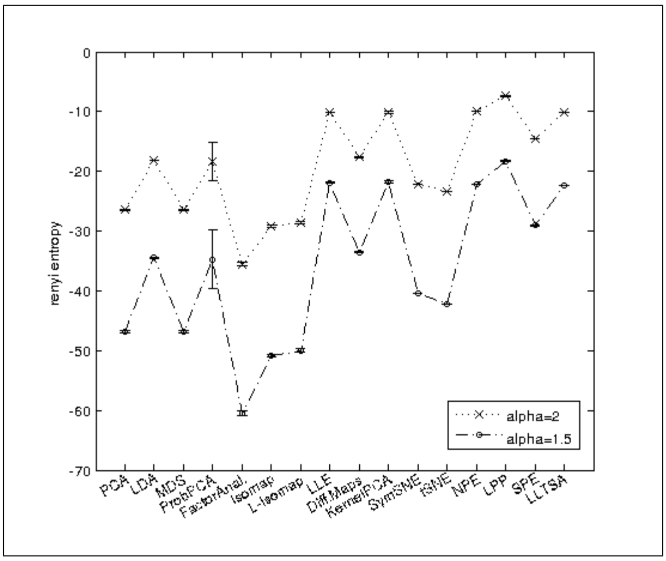

Figure 4: Rényi entropy of various techniques for $\alpha = 1.5$ and $\alpha = 2$.

selected from the available set for each category. The pair-wise distance between these feature vectors is computed using the euclidean distance. Next rényi entropy is computed for different values of the parameter $\alpha \in \{1.5, 2\}$. Each of the dimensionality reduction methods being evaluated projects the data in each category to its sub-space of 13 dimensions, which is the average intrinsic dimensionality of all the categories. The rényi entropy for the pair-wise distance distribution is computed for the same set of values of α for the data in the projected sub-space. Besides the entropy value, the computational time complexity of each dimensionality reduction method is also important. Accordingly, the time taken by each method is recorded for comparison. The experiment is run using *Matlab*® (R2010a) on a 3.0 GHz Xeon processor on a Linux system.

5. RESULTS

The rényi entropy for various techniques for parameter values of $\alpha = 1.5$ and $\alpha = 2$ is shown in figure 4. The best performance is achieved by LPP, with NPE, LLE, and kernel-PCA amongst the

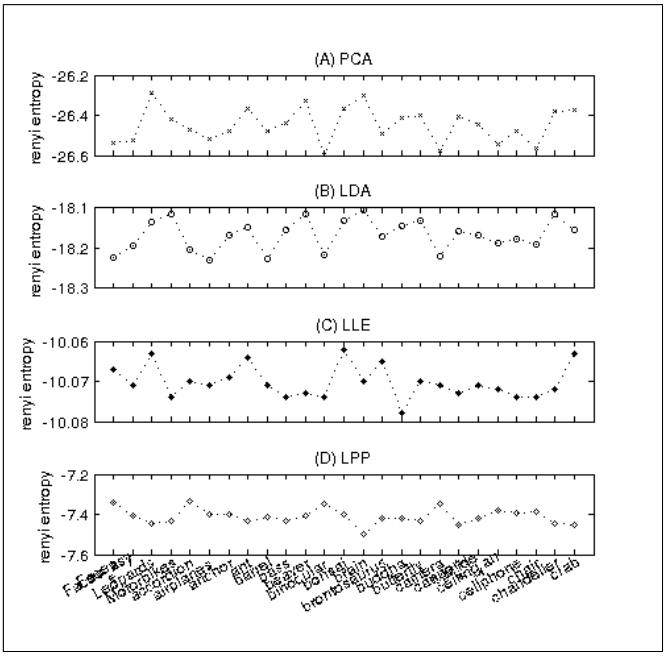

Figure 5: Variation in rényi entropy for different visual categories by different techniques.

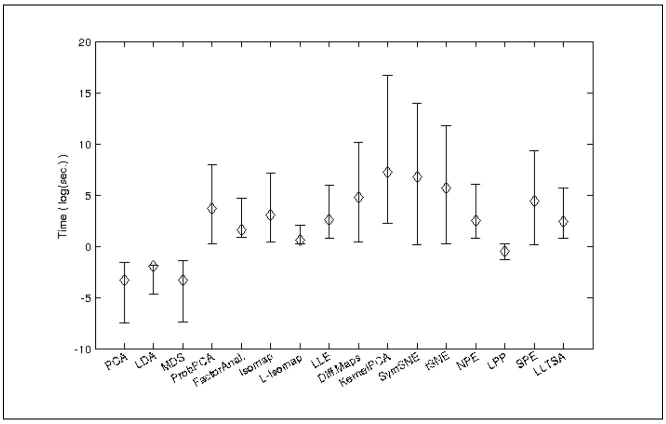

Figure 6: Comparative look at computation time of various techniques.

next best. The error bars show the standard deviation in the entropy measure across all the categories. It is interesting to note that, with the exception of Probabilistic-PCA, all techniques have small variance in entropy measured across categories. This demonstrates that the choice of category is independent of the choice of dimensionality reduction technique and need not be considered further for any more categories. Conversely, from the perspective of the visual categories, no category had a consistently better or worse performance for all techniques. This is shown by a representative set of techniques and categories, in figure 5. The rényi entropy for PCA,LDA,LLE,and LPP for the first 25 categories of Caltech-101 shows no consistent correlation between technique and visual category. The computational time for all the techniques is shown in figure 6. 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.

6. SUMMARY & DISCUSSION

The paper evaluates several dimensionality reduction techniques for the task of visual category recognition. The techniques include linear methods, which are popular and computationally fast, and comparatively recent non-linear methods, which are not commonly utilized and have greater computational time complexity. The issues with visual categorization using BoF model are presented and the features of dimensionality reduction that can help resolve some of these issues is discussed. A novel evaluation metric using rényi entropy of inter-vector distance distribution, is introduced. Rényi entropy is monotonically related to information content and is a family of functionals of the diversity in a distribution. The rényi entropy of the dimensionality reduction methods is compared for multiple values of the parameter α, with the time taken by each of these methods to project the data in each visual category to it's lower dimensional representation. The value of the lower dimension is based on estimation of intrinsic dimensionality using three methods for all categories.

The technique which performed best in terms of the evaluation metric and computation time is the Locality Preserving Projection (LPP). The next best performance was achieved by Neighborhood Preserving Embedding (NPE). The results could be explained by the ability of these techniques to preserve local clusters of visually semantically related feature vectors. Overall, the results highlight the efficacy of non-linear methods over the prevalent linear methods, at a cost of higher computational time.

While the entropic measure indicates performance of the dimensionality reduction technique in terms of preservation of structure in the data, it is unable to ascertain whether semantically equivalent feature vectors are project to the same clusters, so that they count towards the same visual codebook element. Future work would include repeating this experiment on other public datasets to ascertain if LPP continues to perform consistently superior to other techniques, and assessing classification performance along with the entropic measure as evaluation criteria.

Acknowledgement: This work is supported by the EU project Dicta-Sign (FP7/2007-2013) under Grant No. 231135 and the EPSRC project "Making Sense".

REFERENCES

[1] C.C. Aggarwal, A. Hinneburg, and D.A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces", in Proc. 8^{th} Int'l Conf. Databaset Theory, London, UK, pp.420-434, Springer, 2001.

[2] D.K. Agrafiotis, "Stochastic proximity embedding", Journal of Computational Chemistry, 24(10):1215-1221, 2003.

[3] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based approach", in IEEE Trans. on PAMI, 24(10):14041407, 2002.

[4] F. Camastra,"Data dimensionality estimation methods: a survey", in Pattern Recognition, vol. 36(12):2945-2954, 2003.

[5] T.F. Cox and M.A.A. Cox, Multidimensional Scaling, second ed., Chapman and Hall, 2001.

[6] G. Csurka, et. al., "Visual categorization with bags of keypoints", in Proc. of ECCV Int'l Workshop on Statistical Learning in Computer Vision, 2004.

[7] 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", in CVPR'04, Workshop on Generative-Model Based Vision. 2004.

[8] I.K. Fodor, "A Survey of Dimension Reduction Techniques", Technical Report UCRL-ID-148494, Lawrence Livermore Nat'l Laboratory, June 2002.

[9] X. He and P. Niyogi, "Locality preserving projections", In Advances in Neural Information Processing Systems, vol 16, Cambridge, MA, USA, 2004.

[10] X.F. He, D. Cai, S.C. Yan, and H.J. Zhang, "Neighborhood Preserving Embedding", in Proc. of ICCV05,pp. 1208-1213, 2005.

[11] H. Hoffmann, "Kernel PCA for novelty detection", in Pattern Recognition, 40(3):863-874, 2007.

[12] I.T. Jolliffe, Principal Component Analysis, Springer Verlang, pp.487, 1986.

[13] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors", in Proc. of CVPR'04, Vol. 2, pp. 506-513, 2004.

[14] S. Lafon and A.B. Lee. "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization", in IEEE Trans. PAMI, 28(9):1393-1403, 2006.

[15] E. Levina and P.J. Bickel, "Maximum likelihood estimation of intrinsic dimension", In Advances in Neural Information Processing Systems, volume 17, Cambridge, MA, USA, 2004.

[16] D.G. Lowe, "Distinctive Image Features from Scale Invariant Keypoints", Int'l Journal Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[17] L.J.P. van der Maaten and G.E. Hinton, "Visualizing High-Dimensional Data Using t-SNE", Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[18] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik, "Dimensionality Reduction: A Comparative Review", Tilburg University Technical Report, TiCC-TR 2009-005, 2009.

[19] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", in Proc. CVPR'03, Wisconsin, USA, 2003.

[20] F. Nielsen and R. Nock, "On Rényi and Tsallis entropies and divergences for exponential families", in Journal ArXiv e-prints, 1105.3259, 2011.

[21] N. Pinto, D.D. Cox, and J.J. DiCarlo, "Why is Real-World Visual Object Recognition Hard?", in PLoS Comput Biol 4(1): e27, 2008.

[22] J. Ponce et. al., "Dataset issues in object recognition, in Toward Category-Level Object Recognition", volume 4170 of LNCS, pp. 29-48, Springer, 2006.

[23] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by Locally Linear Embedding", Science, 290(5500), pp. 2323-2326, 2000.

[24] V. de Silva and J.B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction", In Advances in Neural Information Processing Systems, vol. 15, pp. 721-728, Cambridge, MA, USA, 2003.

[25] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction", Science, 290(5500):2319-2323, 2000.

[26] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analysers", Neural Computation, 11(2):443-482, 1999.

[27] K. Torkkola, "Linear discriminant analysis in document classification", In IEEE ICDM 2001 Workshop on Text Mining, pp. 800-806, 2001.

[28] J. Zhang, et. al., "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study", Intl Journal Computer Vision, vol. 73, no. 2, pp. 213-238, 2007.

[29] T.H. Zhang, J. Yang, D.L. Zhao and X.L. Ge, "Linear local tangent space alignment and application to face recognition", in Neurocomputing 70, pp. 1547-153, 2007.