Learning a Semantically Relevant Multiple Sub-Space Visual Dictionary for Object Recognition

Ashish Gupta A.GUPTA@SURREY.AC.UK
University of Surrey, Guildford, GU2 7XH, United Kingdom

Abstract

This paper presents a novel approach to learning a visual dictionary from sub-manifolds, using co-clustering, where each sub-manifold is associated with a semantically relevant part of a visual category. The standard dictionary learning technique, called 'Bag-of-Features' is limited by problems of high-dimensionality, sparsity, and noise associated with affine invariant feature descriptors. Our approach draws inspiration from the relation between object part-based models; semantic topic models; non-negative matrix factorization of multivariate data; and sub-spaces in feature space, to resolve these issues in learning a dictionary. We use co-clustering, which performs simultaneous clustering and dimensionality reduction in an optimal way, to discover multiple semantically relevant sub-spaces. We use an information-theoretic and Euclidean divergence based co-clustering. Our approach is comprehensively evaluated on several popular datasets. This work constitutes a principled first step towards a semantically meaningful dictionary, with regards to correspondence between object parts and multiple sub-manifolds, and is not intended to compete with state-of-the-art methods like sparse coding. It is specially pertinent for the future for learning a dictionary with increasing complexity of visual categories.

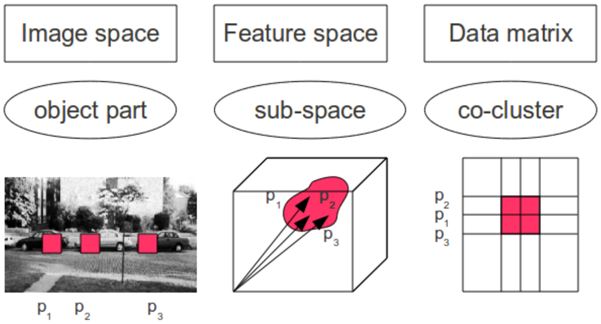

Figure 1. Conceptual association between object part, sub-space, and data matrix block. Feature vectors $\{p_1, p_2, p_3\}$ describing a car part 'wheel' in the image space lie in a sub-space in the feature space, which is associated with 'wheel'. They have similar subset of low level feature attributes and so are co-located in a joint distribution of feature instances and attributes in a data matrix.

1. Introduction

The classification of image visual category is a very difficult task. This paper addresses two principal challenges associated with this task. The first is the affine invariant local image patch feature descriptor commonly utilized, which provides a high-dimensional, sparse and noisy feature space. It is a large vector of low-level image texture features. The second is the visual category, which is not a well defined entity. Rather than a concrete combination of syntactic parts, it is an abstract ensemble of semantic parts. The standard approach currently employed is building a 'Bag-of-Features' (BoF) model. It uses learning vector quantization to learn a visual dictionary and represents each category as an occurrence histogram of the elements of this dictionary. Despite the popularity of this approach, it has certain drawbacks: clustering of feature vectors is based on syntactic similarity rather than semantic similarity; dictionary elements have the same dimensionality as the original feature space. Due to significant intra-category appearance variation, feature vectors belonging to the same object part are generally scattered in feature space and a hard partitioning algorithm like learning vector quantization (LVQ) can not create appropriate clusters. The in-

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

trinsic dimensionality of any visual category is much smaller than the dimensionality of the feature descriptor, which implies that feature vectors pertaining to a visual category are actually embedded in a lower dimensional sub-manifold. The BoF model is unable to address sub-spaces. The use of dimensionality reduction techniques like PCA have been utilized to discover the optimal sub-manifold for the data. This is not a satisfactory solution since a visual category consists of several parts, each of which exist in their own sub-space and consequently, feature vectors of a visual category are actually embedded in an ensemble of multiple sub-manifolds. For example, a 'person' category consists of head, hand, legs, etc. These parts are, in terms of appearance, cohesive and distinct. They have a different set of low-level features associated with each of them. In other words, each part has its own sub-space. We propose a dictionary learning approach which is based on these different sub-spaces. Such a dictionary is expected to perform better because: object parts suffer less intra-category appearance variation and therefore serve as a useful basis to build a model; each sub-space has lower dimensionality; each sub-space occupies a small part of feature space and thereby ameliorates the issue of sparsity; a sub-space has less noise since noisy feature vectors exist in a different sub-space from all the object part sub-spaces.

The algorithm utilized to discover sub-spaces is co-clustering, which is more commonly used by researchers in bio-informatics for gene sequence clustering, described in (Cheng & Church, 2000), where is referred to as bi-clustering. Consider a matrix of visual descriptor data, where the rows are the feature vectors or instances and the columns are dimensions or low-level descriptors. Note figure 1: each feature vector describes a local patch in image space which is uniquely associated with an object part; object part has a unique sub-space in feature space or a unique subset of feature vectors and a unique subset of dimensions. Thus, an object part can be viewed as a block in the data matrix. Co-clustering performs simultaneous row-column clustering in an optimal way to discover these blocks (Dhillon et al., 2003). We build the visual dictionary of an object category using the centroids of these blocks.

To visualize the functioning of multiple sub-manifold dictionary consider figure 2. We shall be training binary classifiers so there are two classes of descriptors for each category. The positive label (blue) descriptors represent the category under consideration and the negative label (red) descriptors represent not the category (descriptors sampled from all other categories in the dataset). The distance between interleaved descriptors with different labels is increased as they are projected to different sub-manifolds. We utilize a k-NN classifier to assess this increase in grouping of descriptors with the same label. The result are shown in section 3.

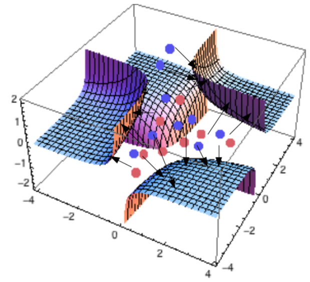

Figure 2. Multiple Sub-Manifold Dictionary: Descriptors from positive and negative categories are projected to respective sub-manifolds. The consequence is that interleaved descriptors of different categories are separated. This should improve classification performance.

2. Co-clustering

Co-clustering analyses similarity between feature vectors in terms of their distribution in feature space and its sub-spaces. The measure of similarity (or its converse the divergence) leads to various schemes of co-clustering. We focus on the information theoretic scheme and discuss the theory and formulation of it. Details of these can be found in (Cho et al., 2004). Consider data matrix $\mathbf{Z}_{m\times n}$ where rows are the feature vectors and the columns are the dimensions of the feature vector. We consider $\mathbf{Z}$ as a joint probability distribution on the random variable z_{uv}. Let U be a random variable that takes values in $1,\ldots,m$ and V takes values in $1,\ldots,n$. Let (U,V) be distributed according to probability distribution μ, where $\mu = \{\mu_{uv} : [u]_1^m, [v]_1^n\}$. We define co-clustering approach as a pair of maps from rows to row-clusters and from columns to column-clusters. Clearly, these maps induce clustered random variables:

$$\begin{aligned} \rho &: \{1,\ldots,m\} \mapsto \{1,\ldots,k\} \\ \gamma &: \{1,\ldots,n\} \mapsto \{1,\ldots,l\} \end{aligned} \tag{1}$$

Let $\hat{U}$ and $\hat{V}$ be random variables in $\{1,\ldots,k\}$ and $\{1,\ldots,l\}$ respectively. So, $\hat{U} = \rho(U)$ and $\hat{V} = \gamma(V)$. Let $\hat{\mathbf{Z}} = [\hat{z}_{uv}]$ be an approximation of data matrix $\mathbf{Z}$ such that $\hat{\mathbf{Z}}$ depends only upon co-clustering (ρ,γ) and statistics derived from the co-clustering. The quality of the co-clustering can be measured by the expected distortion between $\mathbf{Z}$ and $\hat{\mathbf{Z}}$, given by:

$$E[d_\phi(\mathbf{Z}, \hat{\mathbf{Z}})] = \sum_{u=1}^{m} \sum_{v=1}^{n} \mu_{uv} d_\phi(z_{uv}, \hat{z}_{uv}) \tag{2}$$

The co-clustering task is to find (ρ, γ) such that $E[d_\phi(\mathbf{Z}, \hat{\mathbf{Z}})]$ is minimized. The expected divergence is equal to the loss in Mutual information:

$$E[d_\phi(\mathbf{Z}, \hat{\mathbf{Z}})] = I_\phi(\mathbf{Z}) - I_\phi(\hat{\mathbf{Z}}) \tag{3}$$

So, information theoretic co-clustering is defined as: Given k, l; KL divergence d_ϕ; a random variable Z; we wish to find co-clustering (ρ^*, γ^*) that minimizes:

$$\begin{aligned} (\rho^*, \gamma^*) &= \arg\min_{(\rho,\gamma)} E[d_\phi(\mathbf{Z}, \hat{\mathbf{Z}})] \\ &= \arg\max_{(\rho,\gamma)} I_\phi(\hat{\mathbf{Z}}) \end{aligned} \tag{4}$$

The Kullback-Leibler (KL) divergence d_ϕ is given as:

$$d_\phi(\mathbf{Z}, \hat{\mathbf{Z}}) = \sum p(\mathbf{Z}) \log \frac{p(\mathbf{Z})}{p(\hat{\mathbf{Z}})} \tag{5}$$

Co-clustering differs from ordinary one-sided clustering in that at all stages the row cluster prototypes incorporate column clustering information, and vice versa. It intertwines both row and column clustering at all stages. Row clustering is done by assessing closeness of each row distribution, in relative entropy, to certain 'row cluster prototypes'. Column clustering is done similarly, and this process is iterated till it converges to a local minimum.

3. Experiments

In this section we describe the experiments for evaluating the performance of our approach. The objective is to ascertain if projecting the descriptors to multiple sub-manifolds has succeeded in reducing the distance between equivalently labeled descriptors and increased the distance between differently labeled descriptors. To verify this we utilized the k-NN classifier. It may be argued that k-NN is a weak classifier, compared to SVM. But the SVM does not serve the objective here. Further justifications for using this weak classifier can be found in (Boiman et al., 2008). In order to understand the performance of our approach with dictionary size, we select the following set dictionary sizes: { 100, 500, 1000, 5000 }. We utilized six popular datasets for a comprehensive evaluation of our approach: Caltech-101; Caltech-256; Pascal VOC 2006,2007,2010; and Scene-15. The feature detector and descriptor SIFT detects on average a thousand interest points per image. In the implementation of BoF model, PCA is used to project the feature space to 13 dimensions, see (Gupta & Bowden, 2011). K-means clustering is utilized with Euclidean distance metric; randomly generated initial cluster centroids; and upper limit of 100 iterations. The k-NN classifier had a neighborhood size of 10. Classification performance was measured using F_1 score, which is the harmonic mean of precision and recall. It is commonly used for measuring document retrieval and classification performance. In the results of the experiments, the dashed line denotes the BoF model and the solid line denotes our approach.

Expt.1 We compared the BoF model to our approach for all combinations of: datasets = {VOC2006, VOC2007, VOC2010, Scene15, Caltech101, Caltech256 }; co-clustering schemes = { information theoretic, euclidean }; and dictionary sizes = { 100, 500, 1000, 5000 }. The results are collated in table 1.

Expt.2 We analyze the effect of dictionary size { 100, 500, 1000, 5000 } on classification performance. The results are shown in table 2. An interesting result is that for dictionary size of 5000. The category dependency of performance is much higher than in other case. In addition, the relative difference between the two approaches has diminished. The performance of dictionaries of 500 and 1000 are roughly equal. These results indicate that an appropriate dictionary size would likely exist somewhere in that range.

4. Summary

The visual dictionary is a compact representation of the feature descriptor corpus, which is utilized by the encoding method to construct an image representation used to train a classifier. Recent research activity shows a predominant emphasis on encoding techniques like sparse coding and locality constrained coding, whose remarkable classification performance benefit has caused a paradigm shift in the research community. The importance of learning a good dictionary has diminished to the extent that state-of-the-art encoding techniques can yield good classification results with a dictionary of randomly selected elements. This paper argued that the core issues with learning a dictionary is, the lack of semantic relevance in the feature space utilized to build the dictionary; inability to handle the large intra-category appearance variation; inter-mixed descriptors from different categories in a high-dimensional feature space. We viewed a visual category as a combination of multiple visually distinct parts, each of which would have their own sub-set of low-level features. The intuition was that co-clustering would be able to estimate the correlation

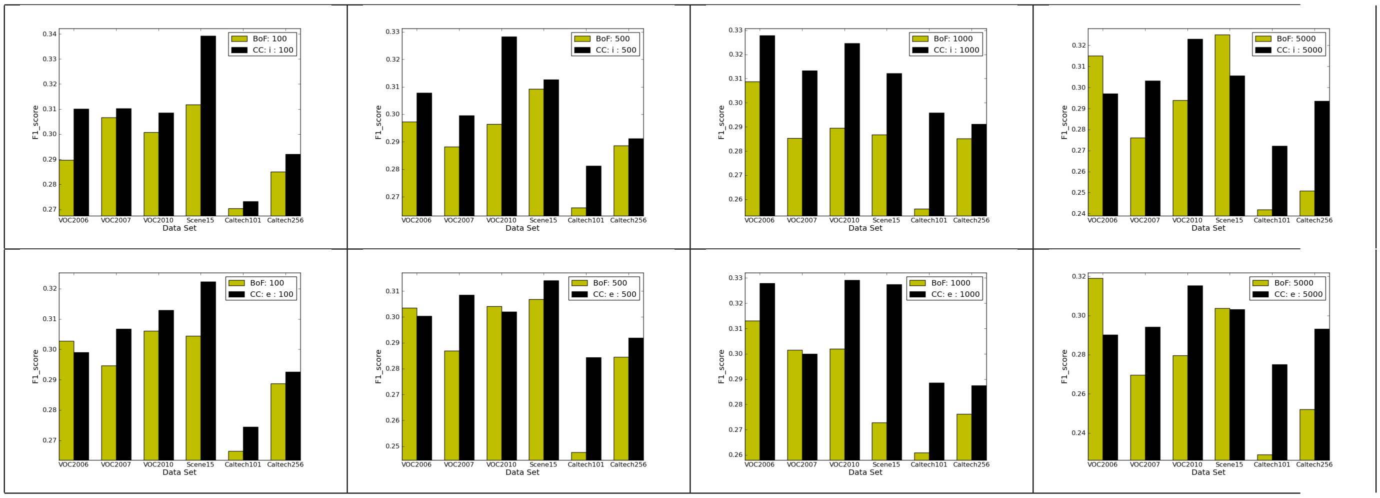

Table 1. Expt. 1: Comparison of classification performance (F1-score) of co-clustering schemes against the BoF model. The co-clustering schemes information-theoretic (i), euclidean (e). The performance is compared for dictionary sizes of 100, 500, 1000, and 5000

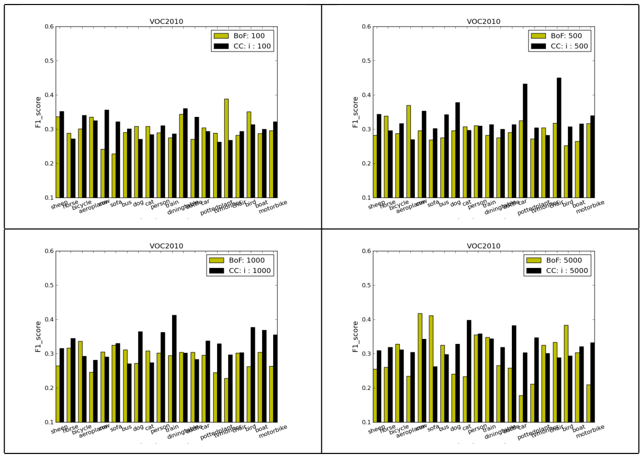

Table 2. Expt. 2: Significance of dictionary size on comparative classification performance of information-theoretic co-clustering against the BoF model. Graphs show results for dictionary sizes of 100, 500, 1000, and 5000 top to bottom. The dataset used this experiment is VOC2010.

between parts of the category and sub-spaces in feature space and thereby learn a dictionary where each dictionary element is embedded in a sub-manifold. The core idea was that the visual category being learned would have a sufficiently different set of sub-manifolds associated with it as compared to other categories and consequently descriptors from two different categories would be projected to different sub-manifolds. This would increase the distance between dissimilar descriptors and subsequently increase classification performance, which was tested using a k-NN classifier. Based on the results across all datasets, we found our approach to provide a marginal but consistently better dictionary than the standard BoF approach. This is a promising result and the margin of better performance could be improved in future. Amongst the various dictionary sizes we experimented with, those in the neighborhood of 1000 dictionary elements provided the best results. The performance of the information theoretic scheme was comparatively better to the other schemes.

References

Boiman, O., Shechtman, E., and Irani, M. In defense of nearest-neighbor based image classification. In *CVPR08*, pp. 1–8, 2008.

Cheng, Yizong and Church, George M. Biclustering of expression data, 2000.

Cho, Hyuk, Y, Inderjit S. Dhillon, Y, Yuqiang Guan, and Y, Suvrit Sra. Minimum sum-squared residue co-clustering of gene expression data, 2004.

Dhillon, Inderjit S., Mallela, Subramanyam, and

Modha, Dharmendra S. Information-theoretic co-clustering. In *In KDD*, pp. 89–98. ACM Press, 2003.

Gupta, Ashish and Bowden, Richard. Evaluating dimensionality reduction techniques for visual category recognition using renyi entropy. In *In Proc. of European Signal Processing Conf.*, pp. 913–917, sept. 2011.