A SPARSE RECONSTRUCTION BASED ALGORITHM FOR IMAGE AND VIDEO CLASSIFICATION

Tanaya Guha and Rabab Ward

Electrical and Computer Engineering
University of British Columbia, Vancouver, Canada

ABSTRACT

The success of sparse reconstruction based classification algorithms largely depends on the choice of overcomplete bases (dictionary). Existing methods either use the training samples as the dictionary elements or learn a dictionary by optimizing a cost function with an additional discriminating component. While the former method requires a good number of training samples per class and is not suitable for video signals, the later adds instability and more computational load. This paper presents a sparse reconstruction based classification algorithm that mitigates the above difficulties. We argue that learning class-specific dictionaries, one per class, is a natural approach to discrimination. We describe each training signal by an error vector consisting of the reconstruction errors the signal produces w.r.t each dictionary. This representation is robust to noise, occlusion and is also highly discriminative. The efficacy of the proposed method is demonstrated in terms of high accuracy for image-based Species and Face recognition and video-based Action recognition.

Index Terms— classification, class-specific dictionary, overcomplete, reconstruction error, sparse representation.

1. INTRODUCTION

Sparse decomposition of signals has emerged as a popular research front in recent years. The basic idea is to represent a signal with a linear combination of a small number of basis functions. It is observed that signals such as audio, images and videos admit sparse representation w.r.t some properly chosen basis functions. For example, music signals are sparse in Fourier bases and many natural images have sparse representation in wavelets bases.

In practice, signals are often found to contain mixed structures that can not be efficiently captured by sinusoids or wavelets alone. This leads to the idea of combining multiple bases to create an overcomplete bases - where the number of basis vectors is greater than the dimensionality of the input signal. A set of overcomplete bases (dictionary) offers greater flexibility in representing the essential structure in a signal which results in higher sparsity in the transform domain. Other advantages of such representation are robustness to additive noise, occlusion and translation of the input signal [1]. An overcomplete dictionary however is redundant i.e. for a given signal there can be many different representations; this redundancy can be removed by imposing proper sparsity constraints. To address this, stable algorithms based on convex optimization [2] or greedy pursuits [3] exist in literature.

A critical task is to choose the appropriate sparsifying bases. One can choose from pre-defined bases (curvelets, variants of wavelets etc. or build a dictionary by concatenating all the training examples in columns [4]. A more generalized approach is to learn the dictionary from a set of examples so that the dictionary elements are better-adapted to the given data. Such dictionaries have been shown to achieve better results compared to off-the-shelf ones [5].

The theory of sparse representation is primarily suitable for classical signal processing problems like denoising, compression etc. Recently a work on face recognition [4] showed that sparse representation is naturally discriminative - it selects only those basis vectors among many, that most compactly represent a signal and therefore is useful for classification. The idea of sparse decomposition has also been extended to other classification problems [6, 7, 8]. Among these [4] and [6] are purely reconstructive i.e. the label of a test image is determined by the label of the dictionary elements producing the lowest reconstruction error. Such reconstruction errors, though robust against noise and occlusion, are not discriminative enough. To increase the discriminative power of the reconstruction errors, [7] builds a dictionary by jointly optimizing an energy formula containing both sparse reconstruction and class discrimination components. However, this approach introduces further difficulty to the already complicated optimization task.

In this paper, we propose a new idea for increasing the discriminative power of the sparse reconstruction errors, for the purpose of classification. The proposed method relies on a set of class-specific dictionaries - each learnt to represent a single class. By learning a dictionary per class, we expect that a dictionary tailored to represent a particular class will give rise to a lower reconstruction error while approximating signals of that class; simultaneously, it will produce larger

Thanks to Prof. Eamonn Keogh, UC Riverside, for sharing the Nematodes dataset and for his insightful comments.

reconstruction errors for signals belonging to a different class. We exploit this inherent discriminating nature of class-specific dictionaries for classification. To describe each input signal a feature vector is constructed which consists of all the reconstruction errors the signal produces w.r.t each dictionary. We also note that the signals of the same class follow a grossly similar pattern of reconstruction errors. This is shown in fig. 1. This joint representation of reconstruction errors is robust to noise, distortions and missing data/occlusion and is also discriminative. Moreover, the error vectors can be used with any standard, reliable classifiers like Support Vector Machine (SVM) or Nearest Neighbor (NN), that can further improve the classification accuracy. Our proposed method is applicable to both image and video classification. Experimental evaluations are presented for three classification tasks - species identification (image based), face recognition (image based) and action recognition (video based). For all applications our results are rather encouraging.

2. THE PROPOSED METHOD

Consider a labeled dataset of images or videos having N different classes. The available training samples per class is m. The training samples are represented as $\mathbf{I}_{ij}$, $i = 1, 2, ..., N$ and $j = 1, 2, ..., m$. First, we intend to learn N different dictionaries, one per class, such that each dictionary is adapted to only one class. The dictionaries can be trained using a good number of raw image/video patches or any other important features extracted from the data. We use randomly extracted overlapping patches to train the dictionaries for images and local spatio-temporal features for videos. Let the set of p feature vectors extracted from a training sample of a certain class be $\{\mathbf{x}_i\}_{i=1}^{p}$, $\mathbf{x} \in \mathbb{R}^d$. Thus for this case, the set of all features obtained from the m available training samples $\hat{\mathbf{X}} = \{\mathbf{x}_i\}_{i=1}^{mp}$ are used to train the dictionary of that class.

2.1. Random Projection

The features are usually high dimensional i.e. d is typically large. Recall that an overcomplete dictionary has more basis vectors than the dimensionality of the input signal. The high value of d thus seriously limits the speed and applicability of our method. A natural solution is to reduce the dimensionality using Principal Component Analysis (PCA), Linear Discriminant Analysis etc. But these traditional methods are slow and data-dependent i.e. the process has to be repeated every time a new datapoint is added. Recently, Random Projection (RP) has emerged as a powerful tool for dimensionality reduction [9]. Theoretical results show that the projections of a high dimensional vectors on a random lower-dimensional subspace can preserve the distances between vectors quite reliably. The original d-dimensional features are projected onto an n dimensional subspace ($n << d$) by premultiplying $\hat{\mathbf{X}}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{n \times d}$. In practice, any normally distributed $\mathbf{R}$ with zero mean and unit variance serves the purpose. The dimensionality reduction step then simplifies to a simple matrix multiplication.

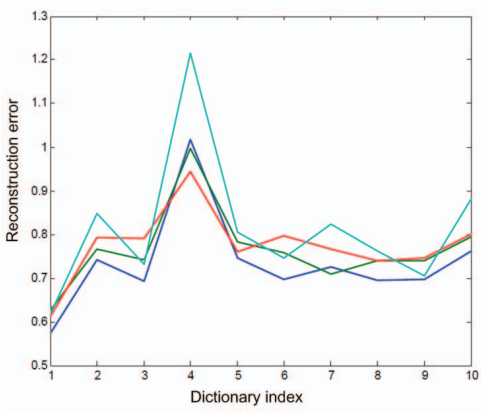

Fig. 1. Shown are the reconstruction errors generated by 4 training samples of same class w.r.t. different dictionaries. The error vectors have their minima at dictionary 1 indicating the dictionary they are closest to; also the maxima for each shown vector corresponds to the same dictionary. Note that, the errors follow a similar pattern for all samples.

$$\mathbf{X} = \mathbf{R}\hat{\mathbf{X}} \tag{1}$$

The reduced data matrix $\mathbf{X} \in \mathbb{R}^{n \times mp}$ contains projections (not true projections, because the vectors are not orthogonal) of $\hat{\mathbf{X}}$ on some random n dimensional subspace.

2.2. Dictionary Learning

After projecting the set of mp features, the reduced data-matrix $\mathbf{X} \in \mathbb{R}^{n \times mp}$ is obtained. The next step is to learn an overcomplete dictionary $\mathbf{\Phi} \in \mathbb{R}^{n \times k}$ having k ($k > n$) atoms over which $\mathbf{X}$ has a sparse representation $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{mp}$, $\mathbf{y}_i \in \mathbb{R}^k$. This optimization problem can be framed as

$$\min_{\mathbf{\Phi}, \mathbf{Y}} \ \|\mathbf{X} - \mathbf{\Phi}\mathbf{Y}\|_F^2, \text{s.t. } \|\mathbf{y}_i\|_0 \leq \tau \tag{2}$$

or, alternatively as

$$\min_{\mathbf{\Phi}, \mathbf{Y}} \ \|\mathbf{X} - \mathbf{\Phi}\mathbf{Y}\|_F^2, \text{s.t. } \|\mathbf{y}_i\|_1 \leq \tau \tag{3}$$

where $\|.\|_F$ is the Frobenious matrix norm and $\|.\|_0$ is the ℓ_0 seminorm that counts the number of non-zero elements in a vector. We employ a fast dictionary learning algorithm called K-SVD that solves (2). It performs two steps at each iteration: (i) sparse coding and (ii) dictionary update. In the first step, $\mathbf{\Phi}$ is kept fixed and $\mathbf{Y}$ is computed. Next, the atoms of a dictionary are updated sequentially, allowing the relevant coefficients in $\mathbf{Y}$ to change as well. For the details of this algorithm please refer to [5].

(a) AT&T face dataset

(b) Nematodes dataset

Fig. 2. Sample images from the image datasets used.

2.3. Reconstruction based classification

Using K-SVD, N class-specific dictionaries $\mathbf{\Phi}^1, \mathbf{\Phi}^2 ... \mathbf{\Phi}^N$ are learnt. Then each training sample is approximated by the N dictionaries with some constant sparsity τ which generates N different reconstruction errors. Let ϵ^γ denote the reconstruction error corresponding to the dictionary $\mathbf{\Phi}^\gamma$. Define ϵ^γ as

$$\epsilon^\gamma = \sqrt{\frac{1}{p}\sum_{i=1}^{p} \|\mathbf{x}_i - \mathbf{\Phi}^\gamma \mathbf{y}_i^\gamma\|_2^2} \tag{4}$$

For every training sample $\mathbf{I}_{ij}$, $i = 1, 2, ..., N$ and $j = 1, 2, ..., m$ an error vector $\mathbf{E}_{ij}$ is computed as

$$\mathbf{E}_{ij} = \left[\epsilon_{ij}^1,\ \epsilon_{ij}^2,\ \ ... \ \epsilon_{ij}^N\right]^T \tag{5}$$

Given a query signal, its corresponding error vector $\mathbf{E}_q$ is computed. The classification is performed in an NN framework where the distance between $\mathbf{E}_q$ and $\mathbf{E}_{ij}$ is computed as

$$d\left(\mathbf{E}_q, \mathbf{E}_{ij}\right) = \sqrt{\left(\mathbf{E}_q - \mathbf{E}_{ij}\right)^T \mathbf{L} \left(\mathbf{E}_q - \mathbf{E}_{ij}\right)} \tag{6}$$

In (6), $\mathbf{L}$ is the Mahalanobis distance metric learnt using an optimization algorithm presented in [10].

3. PERFORMANCE EVALUATION

Experimental evidence for the proposed classification approach is provided for three classification problems: species identification, face recognition and action recognition.

3.1. Image based classification

We use two different image datasets - AT&T dataset for face recognition and the Nematodes dataset for biological species identification. For both datasets, 1000 random patches of size 24×24 are extracted from each image. The patch vectors are projected onto a 64-dimensional subspace. The patch and reduced feature dimensions are found empirically. The class-specific dictionaries are learnt using $n = 64$, $k = 128$, $\tau = 8$ and 20 K-SVD iterations. We compare our results with the convex optimization based image classification approach proposed in [4] which we have implemented using a standard ℓ_1 solver. We also provide comparison with other well-known methods for each application.

Approach	Recognition rate(%)
ℓ_1 optimization [4]	94.3
Eigenface [12]	92.6
Our method (1NN)	**96.5**
Our method (3NN)	**95.6**

Table 1. Results on the AT&T Face dataset.

Approach	Recognition rate (%)
ℓ_1 optimization [4]	54.0
Compression based [11]	56.0
Our method (1NN)	**64.0**
Our method (3NN)	**64.0**

Table 2. Results on the Nematodes dataset.

The *AT&T face dataset*: This benchmark dataset contains 400 grayscale images of 40 individuals in 10 poses. The images were taken at different times, with varying illumination, facial expressions and details. Each image is downsampled by a factor of 2. The training set is constructed by randomly selecting 7 images per class and the rest is used for testing. The results shown in Table 1 are the mean accuracy computed over 10 runs. Our results show improvements over both existing methods which use the same image and feature dimensions. The ℓ_1 approach uses 64 random projections and the Eigenface method uses 64 principal components.

The *Nematodes dataset* [11] is a collection of 50 color images (converted to grayscale) of 5 nematode species [11]. Nematodes are a diverse phylum of wormlike animals, with great commercial and medical importance. Nematodes, because of their diversity, are known to be extremely difficult to be classify correctly. Images are downsampled by a factor of 4 to be consistent with the image size and other parameters. We have adopted a leave-one-out scheme for the evaluation of this dataset to allow direct comparisons to the results obtained by the original authors. The classification results shown in Table 2 show 8% improvement over the state-of-the-art.

3.2. Video based classification

The *UCF sports dataset* [13] is one of the more challenging datasets for action recognition. It contains about 200 real action sequences collected from various sports videos featured on broadcast television channels. The dataset exhibits a wide range of scenes and viewpoints, occlusion, cluttered background, variations in illumination, scale and motion discontinuity. The action classes are: diving, golf swinging, kicking, lifting, horse riding, running, skating, swinging and walking.

Action sequences are commonly described as a collection of local spatio-temporal features. We choose the very popular cuboids feature detector [14] and concatenated gradient

Fig. 3. Sample frames from the UCF sports video dataset.

Approach	Accuracy (%)
Hough transform [15]	86.6
Local trinary patterns [16]	79.2
Our method (1NN)	**81.4**
Our method (3NN)	**82.8**

Table 3. Results on the UCF sports dataset.

descriptors [14] for motion description. The features are obtained at 2 spatial and 2 temporal scales. The high dimensional descriptors are reduced to 128 dimensional subspace. The dictionaries are learnt using $k = 256$, $\tau = 12$ and 20 K-SVD iterations. Like [15, 16], a leave-one-out scheme is adopted for evaluation. Sparse representation based classification of video signals is rare. We have thus compared our results with the state-of-the-art achieved on this dataset. Our classification accuracy (Table 3) is lower than that of [15] which is obtained with more sophisticated features and a computationally expensive classification algorithm.

4. CONCLUSION

We proposed a sparse classification algorithm based on learnt class-specific dictionaries. The proposed algorithm is generic - it is applicable to a variety of classification tasks involving images as well as videos. The novelty of this work lies in the representation of each datapoint in terms of the reconstruction errors produced by each of the class-specific dictionaries. The advantages of the proposed framework are: (i) each class-specific dictionary can be learnt independently; thus the learning process does not have to be repeated when a new class of data is added, (ii) RP reduces computational load significantly, (iii) the error based representation is discriminative and robust to noise and occlusion, (iv) the error vectors can be used as features in any traditional classifier. Superior recognition accuracies achieved on a variety of classification tasks confirm the strength of the proposed method. We have experimented with datasets having 5, 10 and 40 classes. An important question is if this framework is suitable for datasets with larger number of classes. This can be addressed in a future work.

5. REFERENCES

[1] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[3] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Signals, Systems and Computers*, 1993.

[4] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. PAMI*, vol. 31, pp. 210–227, 2008.

[5] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. SP*, vol. 54, pp. 4311–4322, 2006.

[6] G. Peyré, "Sparse modeling of textures," *J. Math. Imaging Vis.*, vol. 34, no. 1, pp. 17–31, 2009.

[7] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. CVPR*, 2008, pp. 1–8.

[8] T. Guha and R.K. Ward, "Action recognition by learnt class-specific overcomplete dictionaries," in *Proc. IEEE FG*, 2011, pp. 143 –148.

[9] R. Baraniuk and M. Wakin, "Random projections of smooth manifolds," *Foundations of Computational Mathematics*, vol. 9, pp. 51–77, 2009.

[10] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2006.

[11] B. J. L. Campana and E. J. Keogh, "A compression-based distance measure for texture," *Statistical Analysis and Data Mining*, vol. 3, no. 6, 2010.

[12] M. Turk and A. Pentland, "Eigen faces for recognition," *J. of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[13] M.D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. CVPR*, 2008.

[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. ICCV VSPETS Workshop*, 2005, pp. 65–72.

[15] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Proc. CVPR*, 2010, pp. 2061 –2068.

[16] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Proc. ICCV*, 2009, pp. 492 – 497.