Beyond SVD: Sparse Projections Onto Exemplar Orthonormal Bases for Compact Image Representation

Karthik S. Gurumoorthy, Ajit Rajwade, Arunava Banerjee and Anand Rangarajan,
Dept. of CISE, University of Florida, Gainesville, USA.
{ksg,avr,arunava,anand}@cise.ufl.edu

Abstract

We present a new method for compact representation of large image datasets. Our method is based on treating small patches from an image as matrices as opposed to the conventional vectorial representation, and encoding those patches as sparse *projections onto a set of exemplar orthonormal bases, which are learned* a priori *from a training set. The end result is a low-error, highly compact image/patch representation that has significant theoretical merits and compares favorably with existing techniques on experiments involving the compression of ORL and Yale face databases.*

1 Introduction

Most conventional techniques of image analysis treat images as elements of a vector space. Lately, there has been a steady growth of literature which regards images as matrices, e.g. [7], [11], [10], [4]. As compared to a vectorial method, the matrix-based representation helps to better exploit the spatial relationships between image pixels. In this paper, we regard an image as a set of matrices (one per image patch) instead of using a single matrix for the entire image as in [7], [11]. There usually exists a great deal of similarity between a large number of patches in one image or across several images of a similar kind. We exploit this fact to learn a small number of full-sized orthonormal bases (as opposed to a single set of low-rank bases learned in [7], [11] and [2], or a single set of projection vectors learned in [10]) to reconstruct a set of patches from a training set by means of *sparse* projections with least possible error.

There exist several research papers on sparse image representation, such as [5]. Some recent contributions include the work in [1], which encodes image patches as a sparse linear combination of a set of overcomplete dictionary vectors (learned from a training set). However all these are again *vector-based techniques*, unlike our matrix-based approach. The matrix-based algorithm presented in [4] may cursorily appear to be similar to the one we present here. However, that technique is based on learning a *single* set of non-orthonormal bases for producing a *neighborhood-preserving* projection of the original training samples onto a lower-dimensional space for a classification application, as opposed to optimizing for good reconstruction, which is one of our main aims here.

Our paper is organized as follows. We describe the theory and the main algorithm in sections (2.1) to (2.3). Section (3) presents experimental results and comparisons with existing techniques. Further discussion is presented in section (4).

2 Theory

Consider a set of images of a particular kind (say face images), each of size $M_1 \times M_2$. We divide each image into non-overlapping patches of size $m_1 \times m_2, m_1 \ll M_1, m_2 \ll M_2$, and treat each patch as a separate matrix. Exploiting the similarity inherent in these patches, we effectively represent them by means of sparse projections onto (appropriately created) orthonormal bases, which we term 'exemplar bases'. We learn these exemplars *a priori* from a set of training image patches. Before describing the learning procedure, we first explain the mathematical structure of the exemplars.

2.1 Exemplar Bases and Sparse Projections

Let $P \in R^{m_1 \times m_2}$ be an image patch. Using singular value decomposition (SVD), we can represent P as a combination of orthonormal bases $U \in R^{m_1 \times m_1}$ and $V \in R^{m_2 \times m_2}$ in the form $P = USV^T$, where $S \in R^{m_1 \times m_2}$ is a diagonal matrix of singular values. However P can also be represented as a combination of *any* set of orthonormal bases $\bar{U}$ and $\bar{V}$, different from those obtained from the SVD of P. In this case, we have $P = \bar{U}S\bar{V}^T$ where S turns out to be a *non-diagonal* matrix [1]. Contemporary SVD-based compression methods leverage the fact that the SVD provides the best *low-rank* approximation to a matrix [9], [6]. We choose to depart from this notion, and instead answer the following question: What *sparse* matrix $W \in R^{m_1 \times m_2}$ will reconstruct P from a pair of orthonormal bases $\bar{U}$ and $\bar{V}$ with the least error $\|P - \bar{U}W\bar{V}^T\|^2$? Sparsity is quantified by an upper bound T on the L_0 norm of W, i.e. on the number of non-zero elements in W (denoted as $\|W\|_0$)[2]. We prove that the *optimal* W with this sparsity constraint is obtained by nullifying the least (in absolute value) $m_1m_2 - T$ elements of the estimated projection matrix $S = \bar{U}^T P \bar{V}$. Due to the ortho-normality of $\bar{U}$ and $\bar{V}$, *this simple greedy algorithm turns out to be optimal* (see Theorem 1). This is quite unlike the approximation algorithms such as orthogonal matching pursuit (OMP), employed in [1] for least-error representation of a given vector as a sparse linear combination of an overcomplete set of unit vectors, which is known to be an NP-hard problem. Moreover, the quality of the approximation in OMP is dependent on T, with an upper bound on the reconstruction error that is $\sqrt{1+6T}$ times the optimal error under some conditions ([8], Theorem C). Our algorithm does not have such dependencies.

Theorem 1: The optimal sparse projection matrix W with $\|W\|_0 = T$ is obtained by setting to zero $m_1m_2 - T$ elements of the matrix $S = \bar{U}^T P \bar{V}$ having least absolute value.

Proof: We have $P = \bar{U}S\bar{V}^T$. The least-squares error (Frobenius norm) in reconstructing a patch P using a matrix W other than S is $e = \|\bar{U}(S-W)\bar{V}^T\|^2 = \|S - W\|^2$ as $\bar{U}$ and $\bar{V}$ are orthonormal. Let $I_1 = \{(i,j) | W_{ij} = 0\}$ and $I_2 = \{(i,j) | W_{ij} \neq 0\}$. Then $e = \sum_{(i,j)\in I_1} S_{ij}^2 + \sum_{(i,j)\in I_2} (S_{ij} - W_{ij})^2$. This error will be minimized when $S_{ij} = W_{ij}$ in all locations where $W_{ij} \neq 0$ and $W_{ij} = 0$ at those indices where the corresponding values in S are as small as possible. Thus if we want $\|W\|_0 = T$, then W is the matrix obtained by nullifying $m_1m_2 - T$ entries from S that have the least absolute value and leaving the remaining intact.□

2.2 Learning the Bases

The essence of this paper lies in a learning method to produce K exemplar orthonormal bases $\{(U_a, V_a)\}$, $1 \leq a \leq K$, to encode a training set of N image patches $P_i \in R^{m_1 \times m_2}$ $(1 \leq i \leq N)$ with least possible error (in the sense of the L_2 norm of the difference between the original and reconstructed patches). Note that $K \ll N$. In addition, we impose a sparsity constraint that every S_{ia} (the matrix used to reconstruct P_i from (U_a, V_a)) has at most T non-zero elements. The main objective function to be minimized is

$$E(\{U_a, V_a, S_{ia}, M_{ia}\}) = \sum_{i=1}^{N}\sum_{a=1}^{K} M_{ia}\|P_i - U_a S_{ia} V_a^T\|^2 \quad (1)$$

subject to the constraints that $U_a^T U_a = V_a^T V_a = I, \forall a$, $\|S_{ia}\|_0 \leq T, \forall (i,a)$ and $\sum_a M_{ia} = 1, \forall i$. Here M_{ia} is a binary matrix of size $N \times K$ which indicates whether the i^{th} patch belongs to the space defined by (U_a, V_a). This is a difficult optimization problem as M_{ia} is binary, so we cast it in an expectation-maximization (EM) framework and relax the binary membership constraint so that now $M_{ia} \in (0,1), \forall (i,a)$, subject to $\sum_{a=1}^{K} M_{ia} = 1, \forall i$ [3]. Using Lagrange parameters $\{\mu_i\}$ and a temperature parameter β, we rewrite the objective function as follows (without the orthonormality constraint on U_a and V_a):

$$E(\{U_a, V_a, S_{ia}, M_{ia}\}) = \sum_{ia} M_{ia}\|P_i - U_a S_{ia} V_a^T\|^2 + \frac{1}{\beta}\sum_{ia} M_{ia}\log M_{ia} + \sum_i \mu_i(\sum_a (M_{ia}) - 1). \quad (2)$$

We first initialize $\{U_a\}$ and $\{V_a\}, \forall a$ to random orthonormal matrices, and $M_{ia} = \frac{1}{K}$, $\forall (i,a)$. Secondly, as $\{U_a\}$ and $\{V_a\}$ are orthonormal, the projection matrix we have $S_{ia} = U_a^T P_i V_a, \forall (i,a)$. Then $m_1m_2 - T$ elements in S_{ia} with least absolute value are nullified. Thereafter, U_a and V_a are updated using the following equations:

$$U_a = Z_{1a}(Z_{1a}^T Z_{1a})^{-\frac{1}{2}} = \Gamma_{1a}\Upsilon_{1a}^T \quad (3)$$

$$V_a = Z_{2a}(Z_{2a}^T Z_{2a})^{-\frac{1}{2}} = \Gamma_{2a}\Upsilon_{2a}^T \quad (4)$$

where $Z_{1a} = \sum_{i=1}^{K} M_{ia} P_i V_a S_{ia}^T$, $Z_{2a} = \sum_{i=1}^{K} M_{ia} P_i^T U_a S_{ia}$, and $(\Gamma_{1a}, \Upsilon_{1a})$ and $(\Gamma_{2a}, \Upsilon_{2a})$ are orthonormal matrix pairs from the SVD of Z_{1a} and Z_{2a} respectively. The membership values are obtained by the following update:

$$M_{ia} = \frac{e^{-\beta\|P_i - U_a S_{ia} V_a^T\|^2}}{\sum_{a=1}^{K} e^{-\beta\|P_i - U_a S_{ia} V_a^T\|^2}}. \quad (5)$$

The matrices S, U, V, M are then updated sequentially following one another for a fixed β value, until convergence. The value of β is then increased and the sequential updates are repeated. The entire process is repeated until an integrality condition is met.

[1] The decomposition $P = \bar{U}S\bar{V}^T$ exists for any P even if $\bar{U}$ and $\bar{V}$ are not orthonormal. We still follow ortho-normality constraints to facilitate optimization and coding. See section 2.3 and 3.1.

[2] See section 3 for the merits of our sparsity-based approach over the low-rank approach.

2.3 Application to Compact Image Representation

Our framework is geared towards compact but *low-error* patch reconstruction. We are not concerned with the *discriminating* assignment of a *specific kind* of patches to a *specific* exemplar, quite unlike in a clustering or classification application. In our method, after the optimization, each training patch P_i $(1 \leq i \leq N)$ gets represented as a projection onto one out of the K exemplar orthonormal bases, which produces the least reconstruction error, i.e. the k^{th} exemplar is chosen if $\|P_i - U_k S_{ik} V_k^T\|^2 \leq \|P_i - U_a S_{ia} V_a^T\|^2, \forall a \in \{1, 2, ..., K\}, 1 \leq k \leq K$. For patch P_i, we denote the corresponding 'optimal' projection matrix as $S_i^\star = S_{ik}$, and the corresponding exemplar as $(U_i^\star, V_i^\star) = (U_k, V_k)$. Thus the entire training set is approximated by (1) the *common* set of basis-pairs $\{(U_a, V_a)\}, 1 \leq a \leq K$ $(K \ll N)$, and (2) the optimal sparse projection matrices $\{S_i^\star\}$ for each patch, with at most T non-zero elements each. The overall storage per image is thus greatly reduced (see also section 3.1). Furthermore, these bases $\{(U_a, V_a)\}$ can now be used to encode patches from a new set of images that are somewhat similar to the ones existing in the training set. However, a practical application demands that the reconstruction meet a specific error threshold on unseen patches, and hence the L_0 norm of the projection matrix of the patch is adjusted dynamically in order to meet the error. Experimental results using such a scheme are described in the next section.

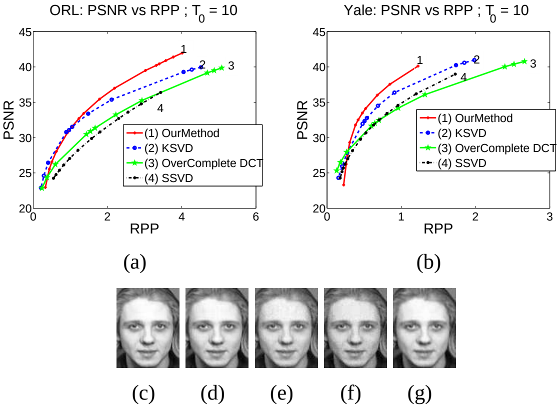

Figure 1. ROC curves on (a) ORL and (b) Yale Databases. Legend- Red (1): Our Method, Blue (2): KSVD, Green (3): Overcomplete DCT, Black (4): SSVD. (c) Original image from ORL database. Sample reconstructions with $\delta = 3 \times 10^{-4}$ of (c) by using (d) Our Method [RPP: 1.785, PSNR: 35.44], (e) KSVD [RPP: 2.112, PSNR: 35.37], (f) Overcomplete DCT [RPP: 2.929, PSNR: 35.256] and (g) SSVD [RPP: 2.69, PSNR: 34.578].

3 Experiments

We tested our algorithm for compression of the ORL database[3] and a subset of the Yale database[4]. For the ORL database, we created a training set of patches of size 12×12 from images of 10 different people, with 10 images per person. Patches from images of the remaining 30 people (10 images per person) were treated as the test set. From the training set, a total of 50 pairs of orthonormal bases were learned using the algorithm described in Section (2). The T value for sparsity of the projection matrices was set to 10 during training. Then, we projected each test patch P_i onto that exemplar $(U_i^\star, V_i^\star)$ which produced the *sparsest* projection matrix $S_i^\star$ that yielded an average per-pixel reconstruction error $\frac{\|P_i - U_i^\star S_i^\star V_i^{\star T}\|^2}{m_1 m_2}$ of no more than some chosen δ. Note that different test patches required different T values, depending upon their inherent 'complexity'. We varied the sparsity of the projection matrix (but keeping its size fixed to 12×12), by greedily nullifying the smallest elements in the matrix, without letting the reconstruction error go above δ. This gave us the flexibility to adjust to patches of different complexities, without altering the rank of the exemplar bases $(U_i^\star, V_i^\star)$. As any patch P_i is projected onto exemplar orthonormal bases that are different from those produced by its own SVD, the projection matrices turn out to be non-diagonal. Hence, there is no such thing as a hierarchy of 'singular values' as in ordinary SVD. As a result, we cannot resort to restricting the rank of the projection matrix (and thereby the rank of $(U_i^\star, V_i^\star)$) to adjust for patches of different complexity. *This highlights an advantage of our approach over that of algorithms that adjust the rank of the projection matrices.* Next, the fractional parts of the values in $S_i^\star$ were quantized using the coding scheme described below, so as to yield a new matrix $\hat{S}_i^\star$. After quantization, the PSNR for each image was measured as $10\log_{10}\frac{N m_1 m_2}{\sum_{i=1}^{N}\|P_i - U_i^\star \hat{S}_i^\star V_i^{\star T}\|^2}$, and then averaged over the entire test set. The average number of bits per pixel (RPP) was calculated as in equation (6) below for each image, and then averaged over the whole test set. We repeated this procedure for different δ values from 8×10^{-5} to 8×10^{-3} (range of image intensity values was $[0, 1]$) and plotted an ROC curve of average PSNR vs. average RPP. We pitted our method against three existing approaches: (1) KSVD

[3] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[4] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

algorithm from [1] using 441 unit norm dictionary vectors of size 144 with the same T value for training, (2) Overcomplete DCT dictionary with 441 unit norm vectors of size 144 created by sampling cosine waves of various frequencies, (3) the SSVD method from [6]. In the former two methods, we computed the optimal sparse projections of each patch onto the dictionary elements using the OMP algorithm [8]. While testing the KSVD and overcomplete DCT methods, the T values for each patch were adjusted dynamically so as to meet the error threshold δ, in the same way as in our method. The same experiment was run on a subset of the Yale database with a value of $T = 10$ (with 12×12 patches from 58 images for training) and 248 images for testing. For our method, we learned 50 pairs of exemplar bases, and the dictionary size for KSVD and overcomplete DCT was 441. The ROC curves for our method were superior to those of other methods over a significant range of δ, for the ORL as well as the Yale database, as seen in Figure 3(a) and (b). Sample reconstructions for an image from the ORL database are shown in Figure 3(c) to (g) for $\delta = 3 \times 10^{-4}$. For this image, our method produced a better PSNR to RPP ratio than others.

3.1 Image Coding

We obtain $S_i^\star$ by sparsifying ${U_i^\star}^T P_i V_i^\star$. As $U_i^\star$ and $V_i^\star$ are orthonormal, we can show that the values in $S_i^\star$ will always lie in the range $[-12, 12]$ for 12×12 patches, if the values of P_i lie in $[0, 1]$. Similar bounds exist for KSVD and overcomplete DCT based approaches as well. We Huffman-encoded the integer parts of the values in the $\{S_i^\star\}$ matrices over the whole image (giving us an average of some Q_1 bits per entry) and quantized the fractional parts with Q_2 bits per entry. Thus, we needed to store the following information per test-patch to create the compressed image: (1) the index of the best exemplar, using a_1 bits, (2) the index and value of each non-zero element in its $S_i^\star$ matrix, using a_2 bits per index and $Q_1 + Q_2$ bits for the value, and (3) the number of non-zero elements per patch encoded using a_3 bits. Hence the total number of bits per pixel for the whole image is given by:

$$RPP = \frac{N(a_1 + a_3) + T^{whole}(a_2 + Q_1 + Q_2)}{M_1 M_2} \quad (6)$$

where $T^{whole} = \sum_{i=1}^{N} \|S_i^\star\|_0$. The values of a_1, a_2 and a_3 were obtained by Huffman encoding. For KSVD and overcomplete DCT, the RPP value was computed using the formula in [1], equation (27), with the modification that the integer parts of the coefficients were Huffman-encoded and the fractional parts separately quantized (as it gave a better ROC curve for KSVD). For the SSVD method, the RPP was calculated as in [6], section (5).

4 Conclusions and Discussion

We have presented a new matrix-based learning-based method for image compression using sparse projections onto exemplar bases. Our approach is radically different from typical SVD-based compression algorithms, in that we replace a low-rank reconstruction by a sparse reconstruction. This sparsity-based approach has two major advantages: (1) optimal reconstructions even with a simple greedy algorithm, unlike approximation algorithms required for computing the optimal sparse combination of an overcomplete set of unit vectors, (2) elegant, principled adjustment to the varying complexity of different image patches. Our algorithm has been tested with good results on two major face databases.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

[2] C. Ding and J. Ye. Two-dimensional singular value decomposition (2DSVD) for 2D maps and images. In *SIAM Intl. Conf. Data Mining*, pages 32–43, 2005.

[3] R. Hathaway. Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, 4:5356, 1986.

[4] X. He, D. Cai, and P. Niyogi. Tensor subspace analysis. In *Neural Information Processing Systems*, pages 499–506, 2005.

[5] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, pages 1457–1469, 2004.

[6] A. Ranade, S. Mahabalarao, and S. Kale. A variation on SVD based image compression. *Image and Vision Computing*, 25(6):771–777, 2007.

[7] A. Rangarajan. Learning matrix space image representations. In *EMMCVPR*, volume 2134, pages 153–168. Springer-Verlag, 2001.

[8] J. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[9] J. Yang and C. Lu. Combined techniques of singular value decomposition and vector quantization for image coding. *IEEE Transactions on Image Processing*, 4(8):1141–1146, 1995.

[10] J. Yang, D. Zhang, A. F. Frangi, and J. yu Yang. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1), 2004.

[11] J. Ye. Generalized low rank approximations of matrices. *Machine Learning*, 61:167–191, 2005.