TRANSFORMATION-INVARIANT DICTIONARY LEARNING FOR CLASSIFICATION WITH 1-SPARSE REPRESENTATIONS

*Ahmet Caner Yüzügüler**

Middle East Technical University
Dept. of Electrical and Electronics Engineering
06800, Ankara, Turkey

Elif Vural and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne
Signal Processing Laboratory (LTS4)
Lausanne, 1015 - Switzerland

ABSTRACT

Sparse representations of images in well-designed dictionaries can be used for effective classification. Meanwhile, training data available in most realistic settings are likely to be exposed to geometric transformations, which poses a challenge for the design of good dictionaries. In this work, we study the problem of learning class-representative dictionaries from geometrically transformed image sets. In order to efficiently take account of arbitrary geometric transformations in the learning, we adopt a representation of the dictionaries in an analytic basis. Then, the proposed algorithm learns atoms that are attracted to the samples of their own class while being repelled from the samples of other classes so that the discrimination between different classes is promoted. The dictionary learning objective is formulated such that it enhances the class-discrimination capabilities of individual atoms rather than the ones of the subspaces they generate, which renders the designed dictionaries especially suitable for fast classification of query images with very sparse approximations. Experimental results demonstrate the performance of the proposed method in handwritten digit recognition applications.

Index Terms— Dictionary learning, image classification, transformation-invariance.

1. INTRODUCTION

Dictionaries adapted to the characteristics of the signals of interest generally facilitate their processing, analysis or coding. Many image processing problems such as image compression, inpainting, and denoising [1], [2], or image classification [3], [4], [5], benefit from sparse representations in well-designed dictionaries. Meanwhile, in real applications, the image data at hand are seldom perfectly aligned. Therefore, the learning of dictionaries in a way that is invariant to the geometric transformations of available training data is critical in a variety of practical scenarios. Transformation-invariance in dictionary learning has been addressed in several previous works, which however only target invariance to specific geometric transformations; e.g., translations [6], [7], scale changes [8], [9], or rotations and scalings [10].

In this work, we study the particular problem of transformation-invariant dictionary learning for image classification. Given a set of training images with known class labels, we learn a dictionary for each class by taking into account the geometric transformations undergone by the training images as well. The atoms in the learned dictionaries approximate well the data samples of their own class while they also contain features that are discriminative between different classes, leading to a good classification performance. The dictionaries are learned in an analytic form by computing their representations in an analytic basis. This is especially useful for handling arbitrary types of geometric transformations of the training data, since common geometric transformations are very often representable in an analytic form and can be integrated directly into the formulation of the dictionary learning objective. Furthermore, the representation of dictionaries in an analytic basis is also desirable from storage and coding perspectives, as the number of basis vectors that is sufficient to compute a good atom is usually much smaller than the dimensionality of the image space. The studies in [16], [17], and [18] are some other works focusing on the representation of signals with analytic or parametric atoms.

Besides achieving invariance to geometric transformations, an important difference of our method from classification-based dictionary learning algorithms such as [4] is that atoms are individually computed such that each atom has the purpose of providing a good representation of a particular region of the image space where the risk of misclassification is high. Hence, our dictionary learning algorithm leads to accurate representations of images even with their 1-sparse approximations in the dictionary. This makes our method especially suitable for applications where a high-speed estimation of class labels of query images is desired, since the 1-sparse approximation of an image in a dictionary can be computed much faster than its approximation with several atoms. In this sense, our approach contrasts with most dictionary learning algorithms for classification such as [3], [4], in which the focus is on subspaces generated by atoms rather than the individual characteristics of the atoms, or, several recent studies such as [14], [15], which propose to classify data based on subspace or union-of-subspace models. Finally, as far as the test stage is concerned, where each image is represented with a single atom, the proposed method bears some resemblance to vector quantization algorithms for classification [11], [12], [13]. However, the main difference between our dictionary learning method and vector quantization is in the training phase. In vector quantization, the mapping between the training data samples and the learned exemplars (codewords) is many-to-one, whereas in dictionary learning a training data sample can be used in the learning of more than one atom.

2. PROBLEM FORMULATION

Let $U = \{U^m\}_{m=1}^M$ be a collection of geometrically transformed images from M classes, where the set $U^m = \{u_i^m\}$ consists of labeled image samples $u_i^m \in \mathbb{R}^n$ from class m. We consider that the images in U have been transformed according to some geometric transformation model parametrizable with vectors $\lambda \in \Lambda$ in a trans-

*Most of the work has been performed while the first author was at EPFL.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

formation parameter domain Λ. The vector λ typically represents a geometric transformation such as a rotation, scale change, affine transformation, etc., or a combination of such transformations.

We then would like to learn a dictionary D^m for each class such that $D^m = \{d_l^m\}_{l=1}^L \subset \mathbb{R}^n$ consists of L atoms. The design of a particular dictionary D^m for each class m, rather than learning a common dictionary for all classes, has the purpose of allowing a simple estimation of the class labels of query images based on their reconstruction errors yielded by their 1-sparse approximations in the learned dictionaries. Note that in a setting where the spatial complexities of images vary very much between different classes, one can select a different number of atoms L_1, ..., L_M for each class. Given a query image u, which can be assumed to be aligned or non-transformed for the simplicity of formulation, we consider that the class label $\hat{m}$ of u is estimated with respect to its reconstruction error in the learned dictionaries as

$$\hat{m} = \arg \min_{m=1,\ldots,M} \|u - \sum_{l=1}^{L} c_l^m \, d_l^m\|. \tag{1}$$

Here $c^m = [c_1^m \ldots c_L^m]$ is a coefficient vector representing a sparse approximation of u in D^m. Since we aim at a fast classification of u, we focus on the case where c^m is 1-sparse, i.e., $\|c^m\|_0 = 1$. Our purpose is then to design the dictionaries $\{D^m\}$ such that the classification rule in (1) gives an accurate estimation.

While attaining a good discrimination power between different classes with 1-sparse representations, we also would like to learn the dictionaries in a transformation-invariant manner in order to reduce their dependence on the geometric transformations of the training images in U. In order to handle geometric transformations in a convenient way, we represent each atom $d_l^m \in \mathbb{R}^n$ as the discretization[1] of a two-dimensional analytic function $\phi_l^m(x,y) \in L^2(\mathbb{R}^2)$. Let $\phi \in L^2(\mathbb{R}^2)$ be the analytic representation of an atom $d \in \mathbb{R}^n$ and ϕ_λ denote the geometrically transformed version of ϕ by λ. We consider geometric transformation models where the two atoms can be related as $\phi_\lambda(x,y) = \phi(a_\lambda(x,y))$ with a bijection $a_\lambda : \mathbb{R}^2 \to \mathbb{R}^2$ representing the coordinate mapping defined by the transformation λ. We then denote by d_λ the discretized version of ϕ_λ in $\mathbb{R}^n$.

We propose to learn the atoms $d_l^m \in D^m$ of the m-th class by minimizing an objective function of the form

$$\begin{aligned} f(d) = & \sum_{i \in I_l^{mm}} \|u_i^m - d_{\lambda_i^m}\|^2 \\ & - \sum_{j \in \{1,\ldots,M\} \setminus \{m\}} \eta_j \sum_{i \in I_l^{mj}} \|u_i^j - d_{\lambda_i^j}\|^2. \end{aligned} \tag{2}$$

Here I_l^{mj} is the set consisting of the indices of the images from the j-th class U^j that are used in the computation of the atom d_l^m. The scalars η_j's adjust the weights of different classes in the optimization of the atoms from class m. The parameter λ_i^j is the geometric transformation applied to the atom d in order to compensate for the transformation undergone by the training image u_i^j. Provided that the index sets I_l^{mj} are chosen suitably, the objective function in (2) encourages the selection of atoms that are good representatives of the images of their own class that are concentrated in a particular region of the image space, while they are pushed away from the nearby samples from other classes in order to reduce misclassifications. In Section 3, we discuss the selection of the parameters of the objective function (2) and its minimization.

[1] We discretize a function simply by sampling it on a regular grid within a rectangular support that captures a substantial part of its energy.

3. DICTIONARY LEARNING ALGORITHM

We first initialize the dictionaries $\{D^m\}$ as follows. We compute the centroid of each set U^m, $m = 1, \ldots, M$, and then initialize the atoms of one class with the training images of other classes that are the most distant to the centroids of their own classes. This initialization strategy has the purpose of providing an initial bias to sample effectively the regions of the image space where images from different classes get critically close to each other. It is usually possible to obtain a better initialization from roughly aligned versions of training images. An important design parameter in the dictionary learning is the dictionary size L, which should be chosen according to the trade-off between the accuracy and complexity constraints in the targeted classification application.

We then update the atoms of the dictionaries $\{D^m\}$ in a sequential way, by minimizing the objective in (2) individually for each atom. We select the index sets I_l^{mj} by identifying a predefined number of images from each class that have the highest correlation with the atom d_l^m. We prefer such an approach instead of the classical sparse coding step in dictionary learning algorithms for the following reasons. First, since we have multiple dictionaries and classes, it is quite costly to compute the sparse coding of all training images in all dictionaries, whereas the above approach is much faster. Second, and more importantly, this strategy of choosing the index sets I_l^{mj} together with the form of the cost function in (2) mimics a sparse coding stage with only one atom, which is consistent with our purpose of accurate classification with 1-sparse representations in the learned dictionaries. Next, λ_i^j's can be set according to an estimation of the geometric transformations undergone by the training images, which can be obtained by aligning u_i^j's with a reference class-representative image for example. However, one may explore more sophisticated strategies to fine tune them, e.g., with an alternating optimization of the atoms and the transformation parameters.

We now discuss the minimization of (2). First, we adopt a representation of the atoms in terms of Hermite 2D functions [19], which provides an efficient way to compute analytic atoms as they form an orthonormal basis for the space of square-integrable functions. Let $\{h_k(x,y)\}_{k=0}^{\infty}$ denote the basis of Hermite 2D functions ordered with respect to increasing degrees of the Hermite polynomials used in their construction [19]. We approximate atoms ϕ with a finite number s of elements from the Hermite basis as

$$\phi(x,y) = \sum_{k=1}^{s} \alpha_k \, h_k(x,y) \tag{3}$$

where α_k are the coefficients of the basis vectors. The parameter s can typically be chosen according to the resolution of the discrete representation in $\mathbb{R}^n$. The transformed version ϕ_λ of ϕ is given by

$$\phi_\lambda(x,y) = \phi(a_\lambda(x,y)) = \sum_{k=1}^{s} \alpha_k \, h_k(a_\lambda(x,y)) \tag{4}$$

where $h_k(a_\lambda(x,y))$ are geometrically transformed Hermite functions. Now let $H_\lambda \in \mathbb{R}^{n \times s}$ be a matrix such that the k-th column of H_λ is obtained by discretizing $h_k(a_\lambda(x,y))$. We can then represent the discrete transformed atom d_λ as

$$d_\lambda = H_\lambda \, \alpha \tag{5}$$

where $\alpha \in \mathbb{R}^{s \times 1}$ is the vector whose k-th entry is α_k. Due to the linearity of geometric transformations, the untransformed version d of the atom d_λ can be represented with the same coefficients α as

$$d = H \, \alpha$$

where H is the matrix constructed from the untransformed Hermite functions $\{h_k(x,y)\}_{k=1}^{s}$. This shows that the coefficients α provide a transformation-invariant representation of atoms.

We can now reformulate the objective function (2) as a function of the coefficients α of the atom d in the Hermite basis as follows.

$$f(\alpha) = \sum_{i \in I_l^{mm}} \|u_i^m - H_{\lambda_i^m}\,\alpha\|^2 - \sum_{j \in \{1,\ldots,M\}\setminus\{m\}} \eta_j \sum_{i \in I_l^{mj}} \|u_i^j - H_{\lambda_i^j}\,\alpha\|^2 \tag{6}$$

Rearranging $f(\alpha)$, we get

$$f(\alpha) = \alpha^T A \alpha - 2b^T \alpha + c \tag{7}$$

where

$$\begin{aligned} A &= \sum_{i \in I_l^{mm}} H_{\lambda_i^m}^T H_{\lambda_i^m} - \sum_{j \in \{1,\ldots,M\}\setminus\{m\}} \eta_j \sum_{i \in I_l^{mj}} H_{\lambda_i^j}^T H_{\lambda_i^j} \\ b &= \sum_{i \in I_l^{mm}} H_{\lambda_i^m}^T u_i^m - \sum_{j \in \{1,\ldots,M\}\setminus\{m\}} \eta_j \sum_{i \in I_l^{mj}} H_{\lambda_i^j}^T u_i^j \\ c &= \sum_{i \in I_l^{mm}} (u_i^m)^T u_i^m - \sum_{j \in \{1,\ldots,M\}\setminus\{m\}} \eta_j \sum_{i \in I_l^{mj}} (u_i^j)^T u_i^j. \end{aligned} \tag{8}$$

The function $f(\alpha)$ is strictly convex and has a unique minimum if A is a positive definite matrix. We thus select the class weights η_j sufficiently small to make A positive definite[2]. One can observe from (2) that selecting the weights in this manner in fact causes the attraction of an atom to the samples of its own class to be stronger than its repulsion from the samples of other classes. This enables the selected atoms to be good representatives of their own class, while they are also encouraged to have features that are distinctive between different classes. Once the weights are set, the coefficients α are then easily computed by solving $\nabla f = 0$, which yields

$$\alpha = A^{-1} b. \tag{9}$$

This gives the untransformed version of the computed atom as $d_l^m = H\alpha$, which is then updated in the dictionary D^m. We continue the updates on the atoms in this manner until some stopping criterion is met. In our implementation, we terminate the algorithm based on the number of iterations. In particular, due to the proposed strategy of initializing the dictionaries in a specific way by prioritizing class-separation boundaries, we have experimentally observed that terminating the algorithm after a single iteration yields a good classification performance as it is useful for retaining the diversity of the atoms within a particular class. The proposed method is summarized in Algorithm 1.

4. EXPERIMENTAL RESULTS

We now study the performance of the proposed method in image classification. We evaluate our method on a data set of handwritten "2, 3, 5, 8, 9" digit images generated from the MNIST database [20] by applying geometric transformations. Each digit is considered as a different class and the training images in each class are obtained with geometric transformations composed of a rotation and an anisotropic scale change. The transformation parameters of training images are randomly selected from the parameter ranges $\theta \in [-\pi/3, \pi/3]$ for the rotation angle; and $s_x, s_y \in [0.7, 1.2]$ for the scaling factors in the x and y-directions.

We first examine the influence of the accuracy of the initial estimation of the transformation parameters $\{\lambda_i^m\}$ on the performance of classification. We experiment on 200 training and 200 test images in each class, and learn a dictionary of 100 atoms for each class with the training images. The sizes of the index sets are chosen as $|I_l^{mm}| = 7$, and $|I_l^{mj}| = 3$ for $j \neq m$ in the algorithm. The dictionaries are first learned with the correct values of the transformation parameters (used in the generation of the data sets) and then by corrupting the transformation parameters with an error in order to simulate parameter estimation errors. The parameter estimation errors are selected uniformly at random from an interval, whose range is increased gradually throughout the experiment. Test images are then classified with respect to the reconstruction errors of their 1-sparse approximations in the learned dictionaries as in (1). We compare the proposed Dictionary Learning for Fast Classification - DLFC method with the reference method of learning a dictionary for each class with the KSVD algorithm. KSVD is applied on aligned

Algorithm 1 Dictionary Learning for Fast Classification (DLFC)

Input: Training images $U = \{U^m\}$
Initialization:
Estimate transformation parameters $\{\lambda_i^m\}$ of training images
Initialize dictionaries $\{D^m\}$ with training images most distant to the centroids of $\{U^m\}$
repeat
  for $m = 1, \ldots, M$ **do**
    for $l = 1, \ldots, L$ **do**
      Determine index sets $\{I_l^{mj}\}$ for atom d_l^m
      Compute Hermite coefficients α of d_l^m according to (8) and (9)
      Update $d_l^m = H\alpha$
    end for
  end for
until stopping criterion
Output: Class-representative dictionaries $\{D^m\}$

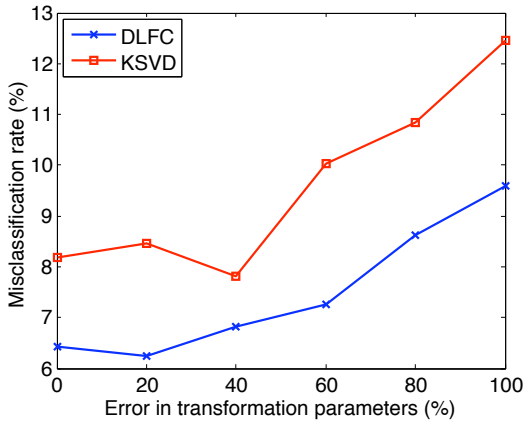

Fig. 1. Dependence of the misclassification error on the accuracy of the estimation of transformation parameters. In the parameter estimation error, 100% corresponds to an error of $\pi/6$ in the rotation angle θ, and an error of 0.1 in the scale factors s_x and s_y.

[2]In practice, we have obtained good results by choosing all η_j equally and assigning them one quarter of the smallest η value that makes the smallest eigenvalue of A vanish.

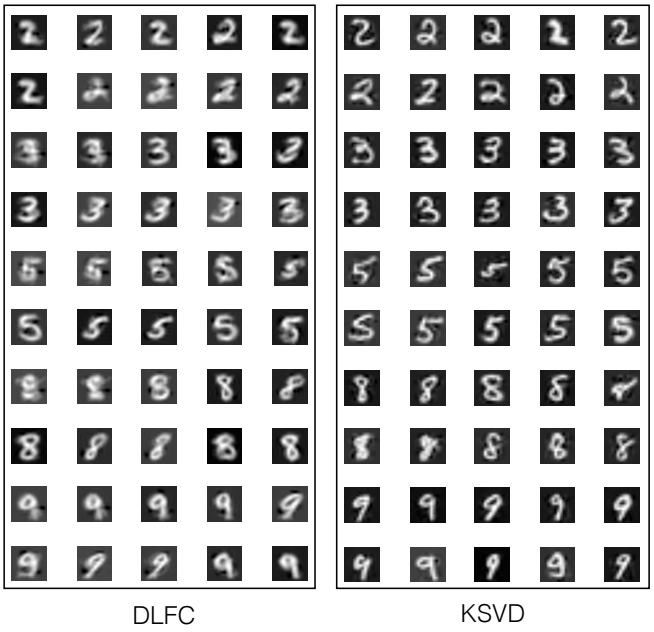

Fig. 2. Atoms learned with DLFC and KSVD

versions of the training images using the same transformation parameters as DLFC. The sparsity parameter of KSVD is set to 40, which gives the best results. In Figure 1, the misclassification rate of test images in percentage is plotted with respect to the transformation parameter estimation error, which is obtained by averaging the results of four instances of the experiment with different training and test sets. The results show that the proposed DLFC method performs better than KSVD in fast image classification with 1-sparse approximations. While the classification performance of the learned dictionaries is seen to have some sensitivity to the accuracy of the transformation parameter estimates, the evolution of the misclassification rate with the parameter estimation error is seen to be similar with the reference KSVD method. In a further experiment where we initialized both algorithms with estimates of the transformation parameters computed by aligning the training images with respect to a reference image from each class, we have also observed a very similar difference between the performances of DLFC and KSVD. The sensitivity of the learned dictionaries to parameter estimation errors can possibly be overcome by integrating the alignment in the dictionary learning, which remains as a future direction of research. Figure 2 provides a visual comparison between the outputs of DLFC and KSVD respectively, where 10 sample atoms are shown from the dictionary of each class. While the atoms learned with KSVD are only approximative of their own class, the atoms learned with DLFC are seen to also include components that increase their distance to the samples of other classes, improving thus the distinction between different classes.

We then study the relation between the classification error and the number of atoms L. We learn dictionaries of different sizes with DLFC and KSVD with the correct transformation parameters and use them to classify test images with 1-sparse approximations as in the previous setting. Figure 3 shows the variation of the misclassification rate with the number of atoms. It is observed that the classification performance of the dictionaries learned with KSVD approaches that of DLFC as the dictionary size increases; however,

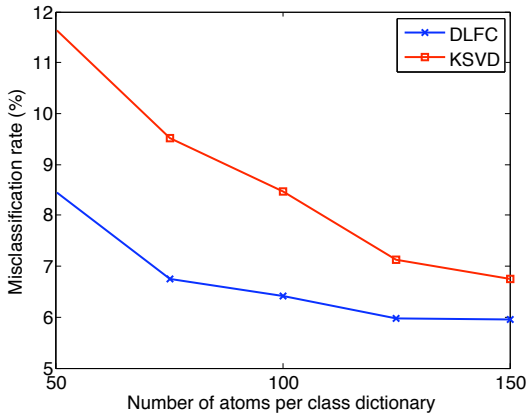

Fig. 3. Variation of the classification error with dictionary size.

DLFC yields considerably better results at small dictionary sizes. This shows the benefit of the proposed method for applications demanding a high-speed classification of query patterns, where limitations on the dictionary size would be of particular concern.

In this study, we have evaluated our method on handwritten digit data. Meanwhile, the proposed method is expected to perform similarly with other data types as well, provided that the data samples of different classes belong to separable regions of the image space, or finite unions of separable regions. Such a hypothesis on the data model, as opposed to alternative data models such as linear or union-of-subspace models, is needed in order to be able to sample the image space sufficiently with individual atoms so that a good classification performance can be obtained with only 1-sparse approximations of data with the learned dictionaries. Finally, although we have demonstrated our method on images undergoing a global geometric transformation, our approach can possibly be extended to learn dictionaries for images containing local geometric transformations with the use of patch-based representations.

5. CONCLUSIONS

We have proposed a method to learn dictionaries in a transformation-invariant way for fast image classification with 1-sparse approximations. Dictionaries are computed in an analytic basis in order to easily handle arbitrary geometric transformations of training data in the learning. The proposed method is based on the idea of optimizing the class-discrimination capability of individual atoms and gives better classification results than purely approximative dictionary learning. A future direction is to improve the sensitivity of the method to the initial estimation of transformation parameters, possibly by performing the dictionary learning and image alignment in a joint manner.

6. ACKNOWLEDGEMENT

The authors would like to thank Ehsaneddin Asgari for providing an implementation of the Hermite 2D function basis.

7. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse repre-

sentation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[3] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," *IMA Preprint Series #2213*, 2008.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[5] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Adv. in Neur. Inf. Proc. Sys.*, 2006.

[6] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Shift-invariant dictionary learning for sparse representations: Extending K-SVD," in *Proc. Eur. Sig. Proc. Conf.*, 2008.

[7] P. Jost, S. Lesage, P. Vandergheynst, and R. Gribonval, "MoTIF: An efficient algorithm for learning translation-invariant dictionaries," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, 2006, vol. 5, pp. 857–860.

[8] J. Mairal, G. Sapiro, and M. Elad, "Multiscale sparse image representation with learned dictionaries," in *Proc. IEEE Int. Conf. Image Proc.*, Sep 2007, vol. 3, pp. 105–108.

[9] P. Sallee and B. Olshausen, "Learning sparse multiscale image representations," in *Adv. in Neur. Inf. Proc. Sys.* 2002, MIT Press.

[10] L. Bar and G. Sapiro, "Hierarchical dictionary learning for invariant classification.," in *IEEE Int. Conf. Acous. Speech Signal Proc.*, 2010, pp. 3578–3581.

[11] C. W. Yen, C. N. Young, and M. L. Nagurka, "A vector quantization method for nearest neighbor classifier design.," *Pattern Recognition Letters*, vol. 25, no. 6, pp. 725–731, 2004.

[12] J. Li, R. M. Gray, and R. A. Olshen, "Joint image compression and classification with vector quantization and a two dimensional hidden markov model.," in *Data Compression Conference*. 1999, pp. 23–32, IEEE Computer Society.

[13] P. Somervuo and T. Kohonen, "Self-organizing maps and learning vector quantization for feature sequences.," *Neural Processing Letters*, vol. 10, no. 2, pp. 151–159, 1999.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[16] E. Kokiopoulou and P. Frossard, "Minimum distance between pattern transformation manifolds: Algorithm and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1225–1238, Jul. 2009.

[17] L. Jacques and C. De Vleeschouwer, "A geometrical study of matching pursuit parametrization," *IEEE Transactions on Signal Processing*, vol. 56, no. 7-1, pp. 2835–2848, 2008.

[18] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4800–4810, Dec. 2009.

[19] A. Wünsche, "Hermite and Laguerre 2D polynomials," *Jour. Comp. Appl. Math.*, pp. 665–678, 2001.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.