ANALYSIS SPARSE CODING MODELS FOR IMAGE-BASED CLASSIFICATION

Sumit Shekhar, Vishal M. Patel and Rama Chellappa

Center for Automation Research, University of Maryland, College Park, MD 20742

{sshekha, pvishalm, rama}@umiacs.umd.edu

ABSTRACT

Data-driven sparse models have been shown to give superior performance for image classification tasks. Most of these works depend on learning a synthesis dictionary and the corresponding sparse code for recognition. However in recent years, an alternate analysis coding based framework (also known as co-sparse model) has been proposed for learning sparse models. In this paper, we study this framework for image classification. We demonstrate that the proposed approach is robust and efficient, while giving a comparable or better recognition performance than the traditional synthesis-based models.

***Index Terms*—** analysis sparse coding models, efficient sparse coding, image classification

1. INTRODUCTION

Sparse representation-based data-driven models have become popular in vision and image processing communities. Olshausen and Field [1] in their seminal work introduced the idea of learning representation based on data itself rather than off-the-shelf bases. Since then sparse representation-based dictionaries have been widely used for image restoration and classification [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Given a data matrix $\mathbf{Y} \in \mathbb{R}^{d \times N}$, whose columns represent d-dimensional signals, the basic formulation underlying these methods involves learning a K-atom synthesis dictionary $\mathbf{D}^* \in \mathbb{R}^{d \times K}$ and sparse code $\mathbf{X}^* \in \mathbb{R}^{K \times N}$, obtained as:

$$\{\mathbf{D}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \text{ s.t. } \|\mathbf{X}\|_0 \leq T_0$$

where, T_0 is the sparsity level. This is a non-convex problem and different schemes have been proposed for optimization, notably, K-SVD [2], matrix factorization [12] and gradient descent [7] techniques.

In recent years, an alternate analysis sparse coding (or co-sparse) model has also been examined [13]. Figure 1 presents a brief comparison of the two models. Previous works have shown that analysis model can yield richer feature representations and better results for image restoration [13]. However, to the best of our knowledge, the analysis framework has not been exploited yet for image classification tasks. In this paper, we examine the application of the analysis model for recognition, and demonstrate that it can achieve comparable or better performance than synthesis models. Further, we show that the proposed approach can lead to a faster optimization at testing time, and the resulting sparse codes are stable under noise and occlusion.

This work was partially supported by a MURI grant from the Army Research Office under the Grant W911NF-09-1-0383.

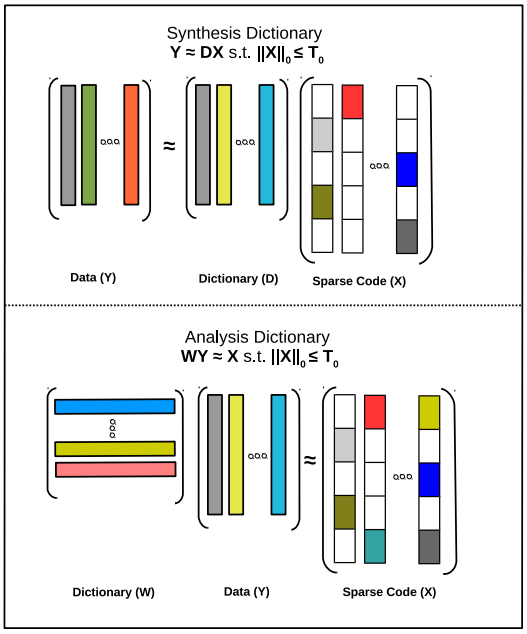


Fig. 1. An overview of synthesis versus analysis models for sparse coding.

2. ORGANIZATION

This paper is organized in six sections. We review the related works in Section 3. The proposed formulation is described in Section 4 and the optimization scheme in Section 5. The classification procedure is described in Section 6 and experimental validations and results are presented in Section 7. Finally, we conclude the paper and suggest future directions in Section 8.

3. RELATED WORKS

Analysis sparse coding models have only recently started receiving attention. A detailed analysis of analysis models was presented in [13]. An analysis K-SVD framework for learning the model was examined in [14]. Peleg *et al* [15] provided theoretical insights into the analysis model. Similarly, methods based on transform coding were proposed in [16, 17]. The idea behind transform coding is to learn transformation, instead of using off-the-shelf methods like DCT, FFT, etc, so that the resulting signal is sparse. These methods show similar performance as the previous analysis models, but have the added advantage of simpler gradient-based optimization and higher speed while testing. This paper studies analysis model

along the lines of transform coding method. However, we generalize it to different recognition scenarios.

4. FORMULATION

Given the data matrix, $\mathbf{Y} \in \mathbb{R}^{d \times N}$, whose columns represent d-dimensional training signals, in analysis dictionary framework [14], the objective is to learn $\mathbf{W} \in \mathbb{R}^{M \times d}$ which minimizes $\|\mathbf{WY}\|_0$. The optimization problem can be written as:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \|\mathbf{WY}\|_0 \text{ s.t. } \mathbf{W} \in \mathcal{A} \tag{1}$$

where, $\mathcal{A}$ is a set of constraints so that the problem is well regularized. However, the input samples can be noisy. In this case, the analysis model can be extended by expressing

$$\mathbf{Y} = \mathbf{X} + \mathbf{E}$$

where, $\mathbf{E}$ is noise and $\mathbf{WX}$ is sparse. This can be solved by the joint optimization problem:

$$\{\mathbf{W}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{W},\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_F^2 \quad \text{s.t. } \|\mathbf{WX}\|_0 \leq T_0 \, , \mathbf{W} \in \mathcal{A} \tag{2}$$

where, T_0 is the sparsity level. But, the transform coding framework [16] shows that handling the error in transformed domain as

$$\mathbf{WY} = \mathbf{X} + \mathbf{E}$$

is more general than (2). Hence, we solve the following optimization problem for analysis coding:

$$\{\mathbf{W}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{W},\mathbf{X}} \|\mathbf{WY} - \mathbf{X}\|_F^2 \quad \text{s.t. } \|\mathbf{X}\|_0 \leq T_0 \, , \mathbf{W} \in \mathcal{A} \tag{3}$$

To obtain a well-regularized solution, we constrain the set $\mathcal{A}$ to be matrices with row-wise norm to be unity. The unit norm condition is required to make the solution non-trivial. However, solving (3) with just these constraints may not lead to a well-conditioned solution. This is because the constraints presented above do not avoid the possibility of repeated rows or linearly dependent rows. To overcome these conditions, we add the following regularization terms to the criterion function:

$$R(\mathbf{W}) = \begin{cases} -\log(\det(\mathbf{W^T W})) & \text{if } m \geq d \\ -\log(\det(\mathbf{W W^T})) & \text{if } m < d \end{cases} \tag{4}$$

This regularization ensures that the learnt $\mathbf{W}$ has full column or row rank depending upon the matrix size. Further, the function is differentiable for cases where $\det(\mathbf{W^T W}) > 0$ or $\det(\mathbf{W W^T}) > 0$. Note that we consider both overcomplete and under-complete cases as both are common in recognition scenarios. Thus, the final optimization is given as:

$$\{\mathbf{W}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{W},\mathbf{X}} \|\mathbf{WY} - \mathbf{X}\|_F^2 + \lambda R(\mathbf{W}) \quad \text{s.t. } \|\mathbf{w}_i\|_2 = 1 \; \forall \; i = 1, \cdots, M, \; \|\mathbf{X}\|_0 \leq T_0 \tag{5}$$

where, $\mathbf{w}_i$ is the i^{th} row of dictionary matrix and $\lambda > 0$ is a hyperparameter. We now describe a strategy to solve the above optimization problem.

5. OPTIMIZATION

The overall cost function is non-convex, however, we follow the strategy of alternate minimization to optimize the cost. This can be done in two steps:

- Update sparse code, $\mathbf{X}$: Fixing $\mathbf{W}$, the solution for $\mathbf{X}$ can be obtained by a simple thresholding. The optimal solution for $\mathbf{X}$ will be given by retaining the top T_0 coefficients in each column of $\mathbf{WY}$. We can also relaxed ℓ_0 constraint to ℓ_1 to make the problem convex. In this case, we can solve the following equivalent problem:

  $$\arg\min_{\mathbf{X}} \|\mathbf{WY} - \mathbf{X}\|_F^2 + \beta\|\mathbf{X}\|_1$$

  This can be solved by applying a soft thresholding scheme as follows:

  $$\mathbf{X}_{i,j} = \begin{cases} (\mathbf{WY})_{i,j} - \frac{\beta}{2} & \text{if } (\mathbf{WY})_{i,j} \geq \frac{\beta}{2} \\ (\mathbf{WY})_{i,j} + \frac{\beta}{2} & \text{if } (\mathbf{WY})_{i,j} < -\frac{\beta}{2} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

- Update dictionary $\mathbf{W}$: Fixing $\mathbf{X}$, we now describe the update steps for $\mathbf{W}$. Even for a fixed $\mathbf{X}$, it is a non-convex problem. We solve the problem using conjugate gradient descent method [18] and then renormalizing the rows of $\mathbf{W}$ to unit norm. During the gradient descent, a small penalty of $\|\mathbf{W}\|_F^2$ can also be added to the cost term for stable solution [17]. The gradient of the function can be computed analytically and is given as:

  $$\nabla_{\mathbf{W}}(\|\mathbf{WY} - \mathbf{X}\|_F^2) = 2\mathbf{WYY^T} - 2\mathbf{YX^T} \tag{7}$$

  $$\nabla_{\mathbf{W}}(R(\mathbf{W})) = -2\mathbf{W}^{\dagger} \tag{8}$$

Thus, the optimization scheme is simple, and we found it to converge quickly during different experiments. A summary of the optimization scheme is given in Algorithm 1.

Input: Data $\mathbf{Y}$, sparsity level T_0, dictionary size M, λ, initial $\mathbf{W}$
Procedure:
Iterate till convergence,
1. *Update* $\mathbf{X}$*:* Update $\mathbf{X}$ using thresholding method.
2. *Update* $\mathbf{W}$*:* Perform conjugate gradient descent followed by renormalizing $\mathbf{W}$ row-wise.
Output: Analysis dictionary $\mathbf{W}$, sparse codes $\mathbf{X}$

Algorithm 1: Analysis Dictionary Learning (ADL)

5.1. Test Sparse Coding

At testing stage, given the test data $\mathbf{y_{te}}$ and trained dictionary $\mathbf{W_{tr}}$, the sparse code can be obtained by solving the optimization problem:

$$\mathbf{x}^*_{\mathbf{te}} = \arg\min_{\mathbf{x}} \|\mathbf{W_{tr} y_{te}} - \mathbf{x}\|_F^2 \text{ s.t. } \|\mathbf{x}\|_0 \leq T_0$$

This can be solved using the thresholding method described above. Hence, the encoding is efficient.

6. CLASSIFICATION

Given training samples from C classes $\{\mathbf{Y}_i\}_{i=1}^{C}$, we concatenate all the training samples to obtain a training matrix as follows

$$\mathbf{Y_{tr}} = [\mathbf{Y}_1, \cdots, \mathbf{Y}_C] \in \mathbb{R}^{d \times N}.$$

We then apply Algorithm 1 to learn the analysis dictionary $\mathbf{W_{tr}}$. Note that we do not employ any discriminative cost while learning. Once $\mathbf{W_{tr}}$ is found, we apply (6) on the training data $\mathbf{Y_{tr}}$ and test data $\mathbf{Y_{te}}$ to obtain feature vectors $\mathbf{X_{tr}}$ and $\mathbf{X_{te}}$, respectively. Once the sparse codes are found, we train a Support Vector Machine (SVM) classifier on $\mathbf{X_{tr}}$ and test it on $\mathbf{X_{te}}$. The entire procedure for classification is summarized in Algorithm 2.

Input: Train Data $\mathbf{Y_{tr}}$, train label, $\ell_{\mathbf{tr}}$, test data $\mathbf{Y_{te}}$, T_0, λ, M.
Procedure:
1. Learn dictionary $\mathbf{W}$ from training data $\mathbf{Y_{tr}}$ and input parameters using Algorithm 1.
2. Obtain sparse codes $\mathbf{X_{tr}}$ and $\mathbf{X_{te}}$ using Eq. (6) and $\mathbf{W_{tr}}$.
3. Train SVM using $\mathbf{X_{tr}}$ and $\ell_{\mathbf{tr}}$ and test on $\mathbf{X_{te}}$.
Output: Test labels, $\ell_{\mathbf{te}}$

Algorithm 2: Classification using ADL.

7. EXPERIMENTS

We conducted experiments on digit and face datasets to demonstrate the efficacy of the proposed method. We compare the proposed method with different synthesis based algorithms like SRC [19], K-SVD [2], discriminative K-SVD (DKSVD) [6], Fisher discrimnat dictionary learning (FDDL) [3], supervised dictionary learning (SDL-G) [4] and incoherent dictionary learning [5]. Note that many of these algorithms use class-wise reconstruction error for classification. For a fair comparison, we report SVM-based classification for K-SVD [2] and FDDL [3] algorithms. The results for other methods are, however, reproduced as reported in literature.

7.1. USPS Digit Dataset

The USPS digit dataset [20] contains images of handwritten digits. The dataset is split into 7291 training and 2007 testing samples. We present results on recognition experiment as well as synthetic experiments to test robustness of the method to noise and missing pixels.

7.1.1. Convergence and Learnt Dictionary

Figure 2 shows the convergence of the optimization and learnt atoms of the dictionary. It can be seen that the cost converges smoothly. The output sparse codes also demonstrate that the learnt dictionary is meaningful, as there are few significant non-zero elements for each digit sample.

7.1.2. Overall Recognition

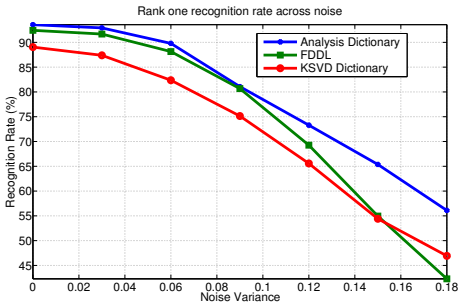
We then compared the recognition rate of proposed method with different synthesis dictionary-based algorithms. We trained an RBF-kernel based SVM classifier, tuning the parameters through cross-validation. The final result is reported for 900 atoms dictionary with $T_0 = 600, \lambda = 0.1$. It can be seen in Table 1 that the accuracy of the proposed method is comparable to other methods. In particular, the proposed method performs better than [4] and is comparable to [3] even though no discriminative cost has been used in training the method. Note that [5] uses reconstruction error for classification, hence, it is not directly comparable to the proposed method.

Method	Recognition rate (%)
ADL-SVM	94.5
KSVD-SVM [2]	92.1
FDDL-SVM [3]	94.7
SDL-G [4]	93.3
Ramirez *et al* [5]	96.0

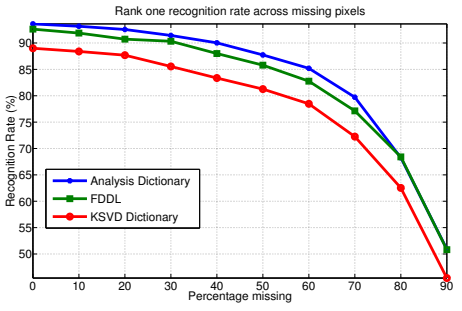
Table 1. Recognition rates for USPS dataset.

7.1.3. Stability under noise and occlusion

We compare the stability of sparse codes generated by the proposed method to those generated by different synthesis coding methods, *viz.*, K-SVD [2] and FDDL [3] under different distortions. In the first experiment, we added random Gaussian noise of increasing variance, and in the second experiment, we randomly set increasing percentage of pixels to zero. We compared the rank-one recognition rates of these methods using the NN-classifier. It can be seen from Figure 3 that the proposed method is more stable, esp. under addition of noise. Thus, analysis method are useful as often sparse codes are used as building blocks for recognition systems [21].

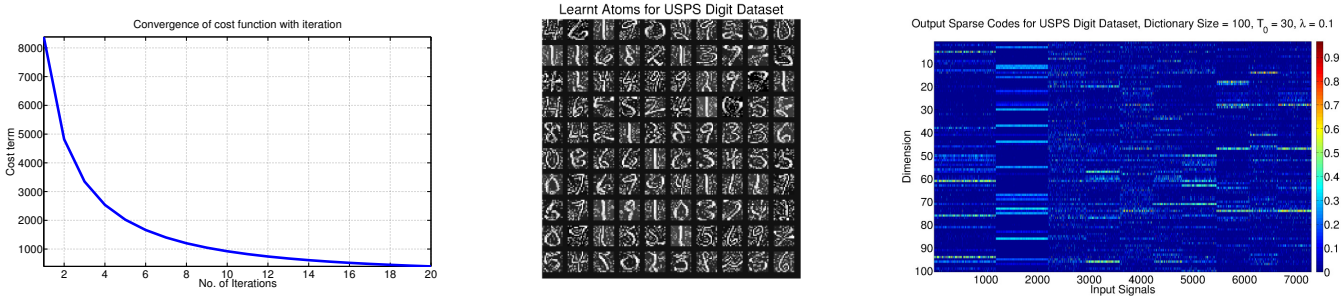


(a)



(b)

Fig. 3. Stabiliy of different sparse coding algorithms under (a) noise, (b) missing pixels.



(a) (b) (c)

Fig. 2. (a) Convergence of the proposed analysis dictionary algorithm, (b) examples of the atoms learnt and (c) absolute value of output sparse codes produced by the algorithm.

7.1.4. Encoding Speed

A significant advantage of the proposed framework over synthesis methods is the simple encoding scheme at test time. We compare the encoding time for the test images of the dataset with algorithms used in sparse coding in synthesis dictionaries, like OMP [22] and SPAMS [12]. Table 2 shows that the proposed ADL alogrithm is much faster than previous methods. All the tests were done on a 2.13 GHz Intel Xeon processor machine using Matlab programming interface.

Method	Time (s)
ADL	0.09
SPAMS [12]	0.15
OMP [22]	2.28

Table 2. Encoding speed for different methods for dictionary size $300, T_0 = 10$, number of samples = 2007.

7.2. AR Face Dataset

The AR face data set [23] consists of faces with varying illumination, expression, and occlusion conditions, captured in two sessions. We evaluated our algorithms on a set of 100 users. Images from the first session, seven for each subject, were used as training and the images from the second session, again seven per subject, were used for testing.

7.2.1. Recognition Comparison

Table 3 shows a comparison with different methods. The proposed method compares favorably with previously proposed synthesis sparse coding methods. Again it should be noted that SRC [19] uses reconstruction error for classification, and hence is not directly comparable. The proposed method however outperforms [6], which is a discriminative dictionary method.

7.2.2. Output Sparse Code

Figure 4 shows the output sparse codes for first 50 test samples. It can be seen that by exploiting the low-dimensional structure of face images, the proposed method is able to learn meaningful sparse codes.

Method	Recognition rate (%)
ADL-SVM	87.7
KSVD-SVM [2]	88.0
FDDL-SVM [3]	88.2
DKSVD [6]	85.4
SRC [19]	88.8

Table 3. Recognition rates for AR Face dataset.

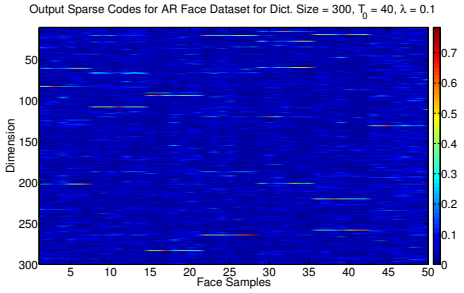


Fig. 4. Output sparse codes produced by the proposed method on AR Face data.

8. CONCLUSION AND FUTURE DIRECTIONS

We have demonstrated some applications of analysis sparse coding to image classification. The proposed approach compares favorably with previous synthesis sparse coding methods and is robust to noise and missing pixels. The method, further, has the advantage of simple encoding scheme at testing, thus, making it efficient.

In this paper, we explored a basic formulation for analysis sparse coding. Future directions include exploring discriminative methods as well as methods to handle to non-linearity in data through kernel approaches. The method can also be extended for other vision tasks, like object detection, tracking, etc for which traditional sparse coding methods have been explored. The proposed method being efficient, looks promising for these applications that require both speed and accuracy.

9. REFERENCES

[1] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision re-*

search, vol. 37, no. 23, pp. 3311–3325, 1997.

[2] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[3] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher Discrimination Dictionary learning for sparse representation," in *IEEE International Conference on Computer Vision*, Nov. 2011, pp. 543–550.

[4] J. Mairal, F. Bach, J. Ponce, A. Zisserman, and G. Sapiro, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems*, 2008.

[5] I. Ramírez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3501–3508.

[6] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *CVPR*, 2010.

[7] F. Bach, J. Mairal, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[8] S. Shekhar, V. M. Patel, and R. Chellappa, "Synthesis-based recognition of low resolution faces," in *International Joint Conference on Biometrics*, 2011, pp. 1–6.

[9] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Design of non-linear kernel dictionaries for object recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5123–5135, 2013.

[10] V. M. Patel and R. Chellappa, "Sparse representations, compressive sensing and dictionaries for pattern recognition," in *Asian Conference on Pattern Recognition(ACPR)*, 2010.

[11] V. M. Patel and R. Chellappa, *Sparse representations and compressive sensing for imaging and vision*, SpringerBriefs, 2013.

[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[13] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[14] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 661–677, 2013.

[15] T. Peleg and M. Elad, "Performance guarantees of the thresholding algorithm for the co-sparse analysis model," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1832–1845, 2013.

[16] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2013.

[17] S. Ravishankar and Y. Bresler, "Learning overcomplete sparsifying transforms for signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3088–3092.

[18] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Tech. Rep., Pittsburgh, PA, USA, 1994.

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

[20] J.J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[21] L. Bo, X. Ren, and D. Fox, "Hierarchical matching pursuit for image classification: Architecture and fast algorithms," in *Advances in Neural Information Processing Systems*, 2011, pp. 2115–2123.

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

[23] A. M. Martinez and R. Benavente, "The AR face database," Tech. Rep., 1998.