LEARNING TRANSFORMATIONS

Qiang Qiu and Guillermo Sapiro

Duke University
Durham, NC 27708, USA

ABSTRACT

A low-rank transformation learning framework for subspace clustering and classification is here proposed. Many high-dimensional data, such as face images and motion sequences, approximately lie in a union of low-dimensional subspaces. The corresponding subspace clustering problem has been extensively studied in the literature, partitioning such high-dimensional data into clusters corresponding to their underlying low-dimensional subspaces. However, low-dimensional intrinsic structures are often violated for real-world observations, as they can be corrupted by errors or deviate from ideal models. We propose to address this by learning a linear transformation on subspaces using matrix rank, via its convex surrogate nuclear norm, as the optimization criteria. The learned linear transformation restores a low-rank structure for data from the same subspace, and, at the same time, forces a high-rank structure for data from different subspaces. In this way, we reduce variations within the subspaces, and increase separation between the subspaces for improved subspace clustering and classification.

1. INTRODUCTION

High-dimensional data often have a small intrinsic dimension. The ubiquitous subspace clustering problem is to partition high-dimensional data into clusters corresponding to their underlying subspaces. Standard clustering methods such as k-means in general are not applicable to subspace clustering. Various methods have been recently suggested for subspace clustering, such as Sparse Subspace Clustering (SSC) [1] (see also its extensions and analysis in [2, 3, 4, 5]), Local Subspace Affinity (LSA) [6], Local Best-fit Flats (LBF) [7], Generalized Principal Component Analysis [8], Agglomerative Lossy Compression [9], Locally Linear Manifold Clustering [10], and Spectral Curvature Clustering [11]. A recent survey on subspace clustering can be found in [12].

Low-dimensional intrinsic structures, which enable subspace clustering, are often violated for real-world data. For example, under the assumption of Lambertian reflectance, [13] show that face images of a subject obtained under a wide variety of lighting conditions can be accurately approximated with a 9-dimensional linear subspace. However, real-world face images are often captured under pose variations; in addition, faces are not perfectly Lambertian, and exhibit cast shadows and specularities [14]. Therefore, it is critical for subspace clustering to handle corrupted underlying structures of realistic data, and as such, deviations from ideal subspaces.

When data from the same low-dimensional subspace are arranged as columns of a single matrix, the matrix should be approximately low-rank. Thus, a promising way to handle corrupted data for subspace clustering is to restore such low-rank structure. In this paper, we propose to improve subspace clustering and classification by learning a linear transformation on subspaces using matrix rank, via its nuclear norm convex surrogate, as the optimization criteria. The learned linear transformation recovers a low-rank structure for data from the same subspace, and, at the same time, forces a high-rank structure for data from different subspaces. In this way, we reduce variations within the subspaces, and increase separations between the subspaces for improved subspace clustering and classification. This can be considered as a fundamental way to learn linear features for data clustering and classification. An extended version of this paper can be found in [15].

2. LEARNING LOW-RANK TRANSFORMATIONS

Let $\{\mathcal{S}_c\}_{c=1}^{C}$ be C n-dimensional subspaces of $\mathbb{R}^d$ (not all subspaces are necessarily of the same dimension, this is only here assumed to simplify notation). Given a data set $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$, with each data point $\mathbf{y}_i$ in one of the C subspaces, and in general the data arranged as columns of $\mathbf{Y}$. $\mathbf{Y}_c$ denotes the set of points in the c-th subspace $\mathcal{S}_c$, points arranged as columns of the matrix $\mathbf{Y}_c$.

As data in $\mathbf{Y}_c$ lie in a low-dimensional subspace, the matrix $\mathbf{Y}_c$ is expected to be *low-rank*, and such low-rank structure is critical for accurate subspace clustering. However, as discussed above, this low-rank structure is often violated for real data. Our proposed approach is to learn a global linear transformation on subspaces. Such transformation restores a low-rank structure for data from the same subspace, and, at the same time, encourages a high-rank structure for data from different subspaces. In this way, we reduce the variation within the subspaces and introduce separations between the

This work was partially supported by ONR, NGA, NSF, ARO, and AFOSR.

subspaces for improved subspace clustering or classification.

Let $||\mathbf{A}||_*$ denote the nuclear norm of the matrix $\mathbf{A}$, i.e., the sum of the singular values of $\mathbf{A}$. The nuclear norm $||\mathbf{A}||_*$ is the convex envelop of $rank(\mathbf{A})$ over the unit ball of matrices [16]. As the nuclear norm can be optimized efficiently, it is often adopted as the best convex approximation of the rank function in the literature on rank optimization (see, e.g., [14] and [17]).

We adopt the nuclear norm as the key learning criterion, and compute one global linear transformation on all subspaces as

$$\arg\min_{\mathbf{T}} \sum_{c=1}^{C} ||\mathbf{T}\mathbf{Y}_c||_* - ||\mathbf{T}\mathbf{Y}||_*, \ \ s.t. ||\mathbf{T}||_2 = 1. \qquad (1)$$

Intuitively, minimizing the *representation* terms $||\mathbf{T}\mathbf{Y}_c||_*$ encourages a consistent representation for the transformed data from the same subspace; and minimizing the *discrimination* term $-||\mathbf{T}\mathbf{Y}||_*$ encourages a diverse representation for transformed data from different subspaces. The normalization condition $||\mathbf{T}||_2 = 1$ on $\mathbf{T}$ prevents the trivial solution $\mathbf{T} = 0$, other normalizations could be considered as well.

Not only the variation within the subspaces is reduced, but as shown in Theorem 1, the learned transformation $\mathbf{T}$ using the objective function (1) also maximizes the separation between subspaces, leading to improved clustering and classification.

Theorem 1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be matrices of the same row dimensions, and* $[\mathbf{A}, \mathbf{B}]$ *be the concatenation of* $\mathbf{A}$ *and* $\mathbf{B}$*, we have*

$$||[\mathbf{A}, \mathbf{B}]||_* \leq ||\mathbf{A}||_* + ||\mathbf{B}||_*.$$

with equality if the column spaces of $\mathbf{A}$ *and* $\mathbf{B}$ *are orthogonal.*

The proof to Theorem 1 can be found in [15]. It is easy to see that theorem 1 can be extended for the concatenation of multiple matrices. Thus, we have,

$$\sum_{c=1}^{C} ||\mathbf{T}\mathbf{Y}_c||_* - ||\mathbf{T}\mathbf{Y}||_* \geq 0. \qquad (2)$$

Based on (2) and Theorem 1, the proposed objective function (1) reaches the minimum 0 if the column spaces of every pair of matrices are orthogonal after applying the learned transformation $\mathbf{T}$; or equivalently, (1) reaches the minimum 0 when the separation between every pair of subspaces is maximized after transformation, i.e., the smallest principal angle between subspaces equals $\frac{\pi}{2}$.

Note that, if we replace the nuclear norm with the rank function, (2) reaches the minimum when subspaces are disjoint but not necessarily maximally separated. It is also not difficult to show that, if we replace the nuclear norm with the induced 2-norm or the Frobenius norm, the objective function (2) is minimized at the trivial solution $\mathbf{T} = 0$, which is prevented by the normalization condition $||\mathbf{T}||_2 = 1$.

We have then, both intuitively and theoretically, justified the selection of the criteria (1) for learning the transform $\mathbf{T}$. We now illustrate the properties of the learned transformation $\mathbf{T}$, first using synthetic examples in Fig. 1. Here we adopt a gradient descent method described in [15] (though other modern nuclear norm optimization techniques could be considered, including recent real-time formulations [18]) to search for the transformation matrix T that minimizes (1). As shown in Fig. 1, the learned transformation $\mathbf{T}$ via (1) maximizes the distance between every pair of subspaces towards $\frac{\pi}{2}$, and reduces the deviation of the data points to the true subspace when noise is present.

It is noted that, given data $\mathbf{Y} \subseteq \mathbb{R}^d$, so far, we considered a square linear transformation $\mathbf{T}$ of size $d \times d$. If we learn a "fat" linear transformation $\mathbf{T}$ of size $r \times d$, where $(r < d)$, we enable dimension reduction along with transformation, thereby connecting this work with compressed sensing for classification [19].

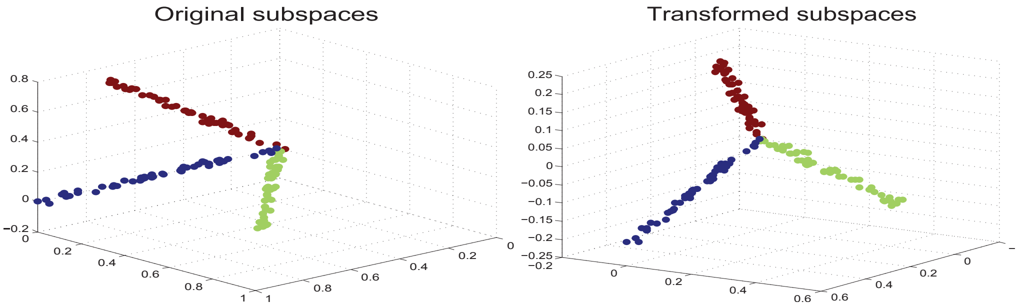
3. SUBSPACE CLUSTERING AND CLASSIFICATION USING LOW-RANK TRANSFORMATIONS

Subspace Clustering: In clustering tasks, meaning to partition the data set $\mathbf{Y}$ into C clusters corresponding to their underlying subspaces, the data labeling is not known beforehand in practice. The proposed algorithm, Algorithm 1, iterates between two stages: In the first assignment stage, we obtain clusters using any subspace clustering methods, e.g., SSC [1], LSA [6], or LBF [7]. In particular, in this paper we use the technique R-SSC introduced in [15], which is a fast subspace clustering technique that fully exploits the low-rank structure of (learned) transformed subspaces. In the second update stage, based on the current clustering result, we compute the optimal subspace transformation that minimizes (1). The algorithm is repeated until the clustering assignments stop changing. Algorithm 1 is a general procedure to enhance the performance of any subspace clustering methods. While formally studying its convergence is the subject of future research, the experimental validation later presented demonstrates excellent performance.

Classification: When a global transformation matrix $\mathbf{T}$ is learned, we can perform classification in the transformed space by simply considering the transformed data $\mathbf{T}\mathbf{Y}$ as the new features. For example, when a Nearest Neighbor (NN) classifier is used, a testing sample $\mathbf{y}$ uses $\mathbf{T}\mathbf{y}$ as the feature and searches for nearest neighbors among $\mathbf{T}\mathbf{Y}$.

4. EXPERIMENTAL EVALUATION

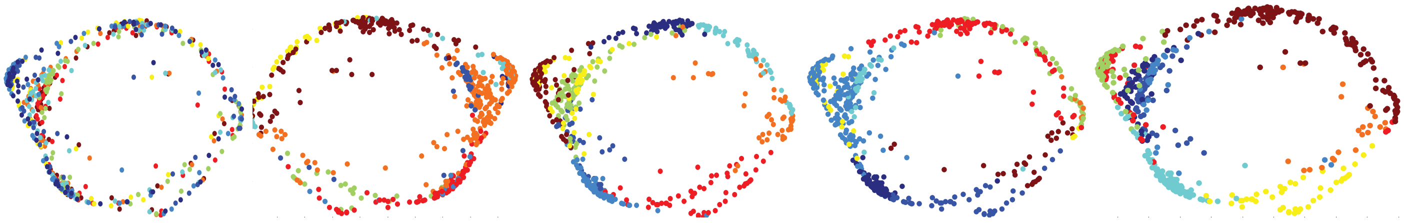
This section presents experimental evaluations on subspace clustering and classification using the public Extended YaleB



(a) $\begin{bmatrix} \theta_{AB} = 0.79, & \theta_{AC} = 0.79, & \theta_{BC} = 1.05 \\ \epsilon_A = 0.0141, & \epsilon_B = 0.0131, & \epsilon_C = 0.0148 \end{bmatrix}$.

(b) $\mathbf{T} = \begin{bmatrix} 0.39 & -0.16 & -0.16 \\ -0.13 & 0.90 & 0.11 \\ -0.23 & 0.11 & 0.57 \end{bmatrix}$; $\begin{bmatrix} \theta_{AB} = 1.51, & \theta_{AC} = 1.49, & \theta_{BC} = 1.57 \\ \epsilon_A = 0.0091, & \epsilon_B = 0.0085, & \epsilon_C = 0.0114 \end{bmatrix}$.

Fig. 1: The learned transformation $\mathbf{T}$ using (1) with the nuclear norm as the key criterion. Three subspaces in $\mathbb{R}^3$, and data points in each subspace are $\mathbf{A}$(red), $\mathbf{B}$(blue), $\mathbf{C}$(green). We denote the angle between subspaces $\mathbf{A}$ and $\mathbf{B}$ as θ_{AB} (and analogous for the other pairs of subspaces). Using (1), we transform $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ in (a) to (b). Data points in (a) are associated with random noises $\sim \mathcal{N}(0, 0.01)$. We denote the root mean square deviation of points in $\mathbf{A}$ from the true subspace as ϵ_A (and analogous for the other subspaces). We observe that the learned transformation $\mathbf{T}$ maximizes the distance between every pair of subspaces towards $\frac{\pi}{2}$, and reduces the deviation of points from the true subspace when noise is present.



(a) Ground truth.

(b) SSC [1], $e = 71.25\%$, $t = 714.99$ *sec.*

(c) LBF [7], $e = 76.37\%$, $t = 460.76$ *sec.*

(d) LSA [6], $e = 71.96\%$, $t = 22.57$ *sec.*

(e) R-SSC [15], $e = \mathbf{67.37}\%$, $t = \mathbf{1.83}$ *sec.*

Fig. 2: Misclassification rate (*e*) and running time (*t*) on clustering 9 subjects using different subspace clustering methods. The R-SSC [15] outperforms state-of-the-art methods both in accuracy and running time. The proposed LRSC in Algorithm 1 (that is, learning the transform) reduces the error to 4.94%, see Fig. 3.

face dataset [20]. The Extended YaleB face dataset contains 38 subjects with near frontal pose under 64 lighting conditions. All the images are resized to 16×16 for clustering and 20×20 for classification. We adopt a NN classifier unless otherwise specified.

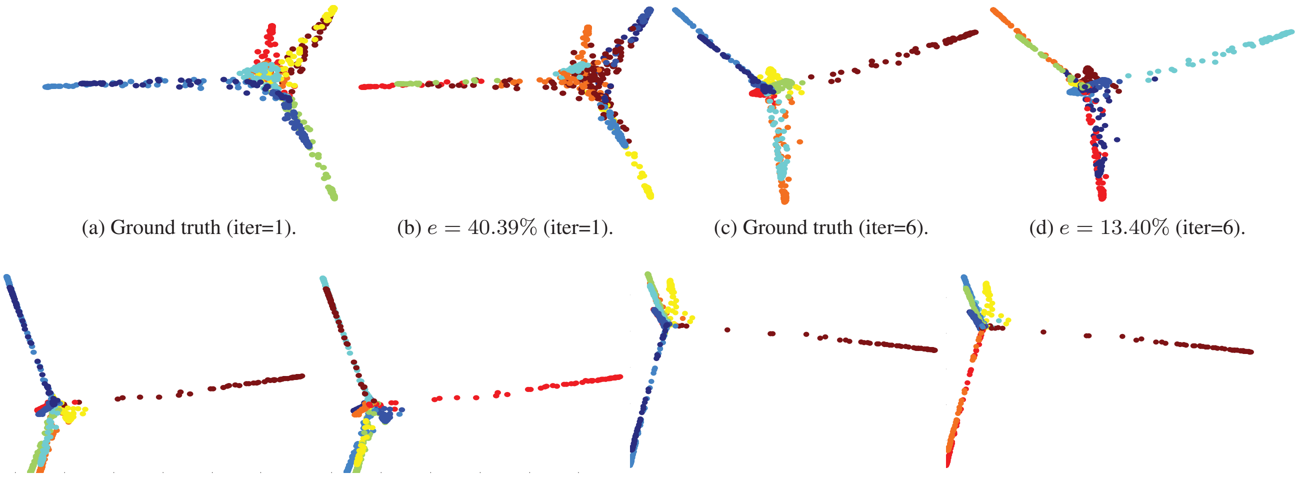
4.1. Application to Face Clustering

In the Extended YaleB dataset, each of the 38 subjects is imaged under 64 lighting conditions. We conduct the face clustering experiments on the first 9 subjects, since it requires $O(C!)$ data cluster comparisons to access the clustering errors given C subspaces. Thus, results are usually reported for no more than 10 subspaces in literature (see, e.g., [7]).

Fig. 2 shows error rate (*e*) and running time (*t*) on clustering subspaces of 9 subjects using different subspace clustering methods from the literature. The R-SSC techniques outperforms state-of-the-art methods both in accuracy and running time. As shown in Fig. 3, using the proposed LRSC algorithm (that is, learning the transform), the misclassification errors of R-SSC are further reduced significantly, for example, from 67.37% to 4.94% for the 9 subjects, simply by preprocessing with the learned matrix. As expected from the theory presented before, such dramatic performance improvement is because the learned subspace transformation increases the distance (the smallest principal angle) between subspaces and, at the same time, reduces the nuclear norms of subspaces.

4.2. Application to Face Recognition

For the Extended YaleB dataset, we adopt a similar setup as described in [21, 22]. We split the dataset into two halves by randomly selecting 32 lighting conditions for training, and



(a) Ground truth (iter=1). (b) $e = 40.39\%$ (iter=1). (c) Ground truth (iter=6). (d) $e = 13.40\%$ (iter=6).

(e) Ground truth (iter=8). (f) $e = 6.17\%$ (iter=8). (g) Ground truth (iter=12). (h) $e = \mathbf{4.94}\%$ (iter=12).

Fig. 3: Misclassification rate (e) on clustering 9 subjects using the proposed LRSC framework in Algorithm 1. We adopt the R-SSC technique in [15] for the clustering step. With the proposed LRSC framework, the clustering error of R-SSC is further reduced significantly, e.g., from 67.37% to 4.94% for the 9-subject case. Note how the classes are clustered in clean subspaces in the transformed domain.

```
Input: A set of data points $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \subseteq \mathbb{R}^d$ in a union of $C$
       subspaces.
Output: A partition of $\mathbf{Y}$ into $C$ disjoint clusters $\{\mathbf{Y}_c\}_{c=1}^C$ based
        on underlying subspaces.
begin
    1. Initial a transformation matrix $\mathbf{T}$ as the identity matrix ;
    repeat
        Assignment stage:
        2. Assign points in $\mathbf{TY}$ to clusters with any subspace
        clustering methods, e.g., the proposed R-SSC;
        Update stage:
        3. Obtain transformation $\mathbf{T}$ by minimizing (1) based on the
        current clustering result ;
    until assignment convergence;
    4. Return the current clustering result $\{\mathbf{Y}_c\}_{c=1}^C$ ;
end
```

Algorithm 1: Learning a robust subspace clustering (LRSC) framework.

the other half for testing. We learn a global low-rank transformation matrix from the training data. We report recognition accuracies in Table 1. The recognition accuracy is increased from 91.77% to 99.10% by simply applying the learned transformation matrix to the original face images. Our method also outperforms state-of-the-art sparse representation based face recognition methods.

Table 1: Recognition accuracies (%) under illumination variations for the Extended YaleB dataset. The recognition accuracy is increased from 91.77% to 99.10% by simply applying the learned low-rank transformation (LRT) matrix to the original face images.

Method	Accuracy (%)
D-KSVD [22]	94.10
LC-KSVD [21]	96.70
SRC [23]	97.20
Original+NN	91.77
LRT+NN	**99.10**

5. CONCLUSION

We introduced a subspace low-rank transformation approach for subspace clustering and classification. Using matrix rank as the optimization criteria, via its nuclear norm convex surrogate, we learn a subspace transformation that reduces variations within the same-class subspaces, and increases separations between the different-class subspaces. We demonstrated that the proposed approach significantly outperforms state-of-the-art methods for subspace clustering and classification, and provided some theoretical support to these experimental results. The same approach has been studied in the context of random forests [24].

Numerous venues of research are opened by the framework here introduced. At the theoretical level, extending the analysis to the noisy case is needed. Furthermore, the study of the framework in its compressed dimensionality form is of critical significance. Beyond this, considering the proposed approach as a feature extraction technique, and its combination with other successful clustering and classification techniques, is the subject of current research.

6. REFERENCES

[1] E. Elhamifar and R. Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 2013, To appear.

[2] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in *International Conference on Machine Learning*, Haifa, Israel, 2010.

[3] M. Soltanolkotabi and E. J. Candes, “A geometric analysis of subspace clustering with outliers,” *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

[4] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, “Robust subspace clustering,” *CoRR*, vol. abs/1301.2603, 2013.

[5] Y. Wang and H. Xu, “Noisy sparse subspace clustering,” in *International Conference on Machine Learning*, Atlanta, USA, 2013.

[6] J. Yan and M. Pollefeys, “A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and non-degenerate,” in *Proc. European Conference on Computer Vision*, Graz, Austria, 2006.

[7] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, “Hybrid linear modeling via local best-fit flats,” *International Journal of Computer Vision*, vol. 100, no. 3, pp. 217–240, 2012.

[8] R. Vidal, Yi Ma, and S. Sastry, “Generalized principal component analysis (GPCA),” in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn.*, Madison, Wisconsin, 2003.

[9] Y. Ma, H. Derksen, W. Hong, and J. Wright, “Segmentation of multivariate mixed data via lossy data coding and compression,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 29, no. 9, pp. 1546–1562, 2007.

[10] A. Goh and R. Vidal, “Segmenting motions of different types by unsupervised manifold clustering,” in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn.*, Minneapolis, Minnesota, 2007.

[11] G. Chen and G. Lerman, “Spectral curvature clustering (SCC),” *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[12] R. Vidal, “Subspace clustering,” *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 52–68, 2011.

[13] R. Basri and D. W. Jacobs, “Lambertian reflectance and linear subspaces,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 25, no. 2, pp. 218–233, February 2003.

[14] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[15] Q. Qiu and G. Sapiro, “Learning transformations for clustering and classification,” *CoRR*, vol. abs/1309.2074, 2013.

[16] M. Fazel, “Matrix Rank Minimization with Applications,” *PhD thesis, Stanford University*, 2002.

[17] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[18] P. Sprechmann, A. M. Bronstein, and G. Sapiro, “Learning efficient sparse and low rank models,” *CoRR*, vol. abs/1212.3631, 2012.

[19] H. Reboredo, F. Renna, R. Calderbank, and M. R. D. Rodrigues, “Compressive classification of a mixture of Gaussians: Analysis, designs and geometrical interpretation,” *pre-print*, 2014.

[20] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 23, no. 6, pp. 643–660, June 2001.

[21] Z. Jiang, Z. Lin, and L. S. Davis, “Learning a discriminative dictionary for sparse coding via label consistent K-SVD,” in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn.*, Colorado springs, CO, June 2011.

[22] Q. Zhang and B. Li, “Discriminative k-SVD for dictionary learning in face recognition,” in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn.*, San Francisco, CA, June 2010.

[23] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[24] Q. Qiu and G. Sapiro, “Learning transformations for classification forests,” *CoRR*, vol. abs/1312.5604, 2013.