FACE RECOGNITION USING MULTI-MODAL LOW-RANK DICTIONARY LEARNING

Homa Foroughi[*], *Moein Shakeri*[*], *Nilanjan Ray, Hong Zhang*

Department of Computing Science, University of Alberta, Edmonton, Canada

ABSTRACT

Face recognition has been widely studied due to its importance in different applications; however, most of the proposed methods fail when face images are occluded or captured under illumination and pose variations. Recently several low-rank dictionary learning methods have been proposed and achieved promising results for noisy observations. While these methods are mostly developed for single-modality scenarios, recent studies demonstrated the advantages of feature fusion from multiple inputs. We propose a multi-modal structured low-rank dictionary learning method for robust face recognition, using raw pixels of face images and their illumination invariant representation. The proposed method learns robust and discriminative representations from contaminated face images, even if there are few training samples with large intra-class variations. Extensive experiments on different datasets validate the superior performance and robustness of our method to severe illumination variations and occlusion.

Index Terms— Multi-modal dictionary learning, Low-rank learning, Illumination invariant, Face recognition

1. INTRODUCTION

The last decade has witnessed a tremendous progress in face recognition technologies, and great recognition performance has been reported by different methods under some ideal conditions, but most of these methods are not robust to outliers, occlusions, severe illumination and pose variations. In recent years, dictionary learning (DL) algorithms have been successfully applied to different vision tasks including face recognition. DL is a feature learning technique in which, an input signal is represented with a sparse linear combination of dictionary atoms. To alleviate the effects of aforementioned variations, low-rank (LR) matrix recovery has been integrated into the DL framework, and is shown to achieve promising results when corruption existed. LR matrix recovery [1] was oroginally proposed to recover a LR matrix from corrupted observations, and have succesfuly been applied to applications like background modeling [2] and image classification [3]. Li *et al*. [4] developed a discriminative DL method by combination of the Fisher discrimination and the LR constraint on sub-dictionaries. Zhang *et al*. [5] presented a structured, sparse

and LR representation for image classification by adding a regularization term to the DL objective function. Recently, Foroughi *et al*. [3] proposed a joint projection and LR-DL method using dual graph constraints for classification of small datasets, which include considerable amount of variations.

In parallel developments, it is well established that information fusion using multiple sources can generally improve the recognition performance, since it provides a framework to combine information from different perspectives that is more tolerant to the errors of individual sources [6]. To benefit from information fusion, some methods have also successfully incorporated DL technique into the feature learning framework. Monaci *et al*. [7] proposed a multi-modal DL algorithm to extract typical templates, which represents synchronous transient structures between multi-modal features. [8] proposed an uncorrelated multi-view discrimination DL method based on the Fisher discrimination, that jointly learns multiple uncorrelated discriminative dictionaries from different views. Nevertheless, the only work that integrated LR into multi-modal DL was presented by Wu *et al*. [9] through constructing class-specific sub-dictionaries for each modality, and utilizing LR and incoherence constraints on each view.

To construct different modalities, most of the existing methods either exploit multi-view angles [9] or extract different local features [9] or weak biometrics [10] from predefined regions of face images. These methods suffer from two main disadvantages that burden extra overhead on the system. Firstly, they demand either several cameras or manual region definition and hand-crafted feature extraction and secondly, they are not applicable to millions of available face data that have already been captured under single view. By exploiting more meaningful modalities, we address these challenges and even increase the recognition rate further. Recently, Shakeri *et al*. [11] presented an illumination invariant representation of an image for outdoor place recognition. To create this representation, they use a Wiener filter derived from the power law spectrum assumption of natural images that is robust against illumination variations. Since the obtained representation may lose the chromaticity of the image, a shadow removal method based on entropy minimization is utilized. This representation showed superior performance for outdoor place recognition in various illumination and shadow variations. Inspired by this success, we design a framework for multi-modal fusion with the following contributions:

---

[*]These authors contributed equally to this work

- We design a multi-modal LR-DL method, where in each modality a discriminative and reconstructive dictionary, and a structured sparse and LR representation are learned from face images, and the collaboration between modalities is encouraged by incorporating an ideal representation term. We provide a new classification schema, which utilizes the reconstruction by LR and sparse noise components.

- By adopting illumination invariant representation of images as one of the modalities, the model learns robust and discriminative representations from noisy images, even when the kind of variation is different in the training and test sets. The proposed method achieves superior performance for small datasets that have large intra-class variation.

2. THE PROPOSED MM-SLDL METHOD

We propose a *M*ulti-*M*odal *S*trcutured *L*ow-rank *D*ictionary *L*earning method (MM-SLDL) for face recognition, in which we use two modalities. While the first modality is constructed by the raw pixels of face images, the second is formed by illumination invariant images [11]. Denote X_K ($K = 1, 2$) the training data from the Kth modality including C classe, as $X_K = \{X_K^1, X_K^2, \ldots, X_K^C\}$, where X_K^i corresponds to class i in the Kth modality. In each modality, we use a supervised learning method to learn a discriminative and reconstructive dictionary D_K, and a structural sparse and LR image representation Z_K. LR matrix recovery helps to decompose the corrupted matrix X_K into a LR component DZ and a sparse noise component E, *i.e.*, $X_K = D_K Z_K + E_K$. With respect to dictionary D_K, the optimal representation matrix Z_K for X_K should be block-diagonal [12], *i.e.*, $Z_K^{*\,ii} = Z_K^i$. In each modality, the dictionary D_K contain C sub-dictionaries as $D_K = \{D_K^1, D_K^2, \ldots, D_K^C\}$, where D_K^i corresponds to the ith class. Let $Z_K^i = \{Z_K^{i,1}, Z_K^{i,2}, \ldots, Z_K^{i,C}\}$ be the representation of X_K^i with respect to D_K, then $Z_K^{i,j}$ denotes coefficients for D_K^j. To learn robust representations from images, D_K should have discriminative and reconstructive power. Firstly, D_K^i should well represent the samples in class i, and ideally be exclusive to each subject i. Secondly, every class i needs to be well represented by its sub-dictionary, such that $X_K^i = D_K^i Z_K^{i,i} + E_K^i$, and finally $Z_K^{i,j}$, the coefficients for D_K^j ($i \neq j$), are nearly all zero. So, the objective function of MM-SLDL is defined as:

$$\min_{D_K, Z_K, E_K} \sum_{K=1}^{2} (\|Z_K\|_* + \beta\|Z_K\|_1 + \lambda\|E_K\|_1)$$
$$+\alpha\|Z_1 Z_2^T - Q\|_F^2 \quad s.t. \quad X_K = D_K Z_K + E_K \quad K = 1, 2 \tag{1}$$

The main objective function simultaneously trains two dictionaries and representations under the joint ideal regularization prior. Q is an ideal representation built from training data in block-diagonal form, and defined as $Q = \{q_1, q_2, \ldots, q_C\} \in R^{C \times C}$. Here C is the size of dictionary, and q_i is the code for

sample x_K^i in the form of $[0 \ldots p_i, p_i, p_i, \ldots]^t \in R^C$, where p_i is the number of training samples in class i. This means that if x_K^i belongs to class L, then the coefficients in q_i for D_K^L are all p_is, while the others are all 0s. We add the regularization term $\|Z_1 Z_2^T - Q\|_F^2$ for two reasons: first, to include the structure information into the dictionary learning process and second, to enforce collaboration between two modalities. It encourages the training images of the same class to have the same representation Z_K in different modalities, despite of intra-class variations.

Classification Scheme: After dictionaries D_K are learned, the LR sparse representations Z_K of training data X_K and Z_K^{ts} of test data X^{ts} are calculated by solving (2) separately using Algorithm 1 with $\alpha = 0$. The representation $Z_{K,i}^{ts}$ of the ith test sample is the ith column of Z_K^{ts}. Using the multivariate ridge regression model [13], we obtain a linear classifier $\hat{W}_K$ as:

$$\hat{W}_K = \underset{W_K}{\operatorname{argmin}} \|H - W_K Z_K\|_2^2 + \lambda\|W_K\|_2^2 \tag{2}$$

where H is the class label matrix of X_K. This yields $\hat{W} = HZ_K^T(Z_K Z_K^T + \lambda I)^{-1}$. The estimated label of the Kth modality is obtained as:

$$c_K = \underset{c_K}{\operatorname{argmax}} \left(s = (\hat{W}_K + Q)Z_{K,i}^{ts}\right) \tag{3}$$

where s is the class label vector. We then use LR matrix recovery to obtain LR and sparse noise components of potential classes c_1, c_2, and then compute the reconstruction error of the given query sample X_i^{ts} in both modalities:

$$\|L(c_K) - \left(X_{K,i}^{ts} - \bar{S}(c_K)\right)\|_F^2 \tag{4}$$

where $L(c_K)$ is the LR component of class c_K in the Kth modality, and $\bar{S}(c_K)$ is the average sparse noise of that class. Since the data range is different in two modalities, we use a normalization step, and the winner class is the one that minimizes the ratio of (4) between two modalities.

3. OPTIMIZATION OF MM-SLDL

In each iteration we update the variables of the Kth modality, while fixing the variables of other modality, and for the Kth modality, the variables are updated alternatively. To solve optimization problem (2), we first introduce an auxiliary variable W_K to make it separable:

$$\min_{D_K, Z_K, E_K} \sum_{K=1}^{2} (\|Z_K\|_* + \beta\|W_K\|_1 + \lambda\|E_K\|_1) \tag{5}$$
$$+\alpha\|Z_1 Z_2^T - Q\|_F^2 \quad s.t. \quad X_K = D_K Z_K + E_K, W_K = Z_K$$

The augmented Lagrangian function L of (5) is defined as:

$$L = \|Z_K\|_* + \beta\|W_K\|_1 + \lambda\|E_K\|_1 + \alpha\|Z_1 Z_2^T - Q\|_F^2$$
$$+ <Y_K, X_K - D_K Z_K - E_K> + <M_K, Z_K - W_K>$$
$$+ \frac{\mu}{2}(\|X_K - D_K Z_K - E_K\|_F^2 + \|Z_K - W_K\|_F^2) \quad K = 1, 2 \tag{6}$$

where $< A, B >= tr(A^T B)$, Y_K and M_K are Lagrange multipliers and μ is a balance parameter. The optimization problem (3) can be divided into two sub-problems as follows:

- **Updating Coding Coefficient Z_K:** With D_K fixed, we use the linearized alternating direction method with adaptive penalty (LADMAP) [14] to solve for Z_K and E_K. The augmented Lagrangian function (3) would reduce to:

$$\|Z_K\|_* + \beta\|W_K\|_1 + \lambda\|E_K\|_1 + \alpha\|Z_1 Z_2^T - Q\|_F^2$$
$$+\frac{\mu}{2}(\|X_K - D_K Z_K - E_K + \frac{Y_K}{\mu}\|_F^2$$
$$+\|Z_K - W_K + \frac{M_K}{\mu}\|_F^2) - \frac{1}{2\mu}(\|Y_K\|_F^2 + \|M_K\|_F^2) \quad (7)$$

The function (3) should be minimized by alternative updating variables Z_K, W_K, E_K as follows:

$$Z_K^{j+1} = \underset{Z_K}{\operatorname{argmin}} \; \frac{1}{\eta\mu}\|Z_K\|_* + \frac{1}{2}\|Z_K - Z_K^j + \mu[-D_K^T(X_K$$
$$-D_K Z_K^j - E_K^j + \frac{Y_K^j}{\mu}) + \frac{2\alpha(Z_K Z_l^T - Q)Z_l}{\eta\mu}$$
$$+\mu(Z_K - W_K^j + \frac{M_K^j}{\mu})]/\eta\mu\|_F^2 \quad \text{where} \quad l \neq K \quad (8)$$

where $\eta = \|D_K\|_2^2$, and we notice that $Q = Q^T$.

$$W_K^{j+1} = \underset{W_K}{\operatorname{argmin}} \; \frac{\beta}{\mu}\|W_K\|_1 + \frac{1}{2}\|W_K - Z_K^{j+1} - \frac{M_K^j}{\mu}\|_F^2 \quad (9)$$

$$E_K^{j+1} = \underset{E_K}{\operatorname{argmin}} \; \frac{\lambda}{\mu}\|E_K\|_1 + \frac{1}{2}\|E_K - (\frac{Y_K^j}{\mu} + X_K$$
$$-D_K Z_K^{j+1})\|_F^2 \quad (10)$$

- **Updating Dictionary D_K:** When Z_K, W_K, E_K are fixed, we would be able to update D_K. The Lagrangian function (3) is further reduced to:

$$\frac{\mu}{2}(\|X_K - D_K Z_K - E_K + \frac{Y_K}{\mu}\|_F^2 + \|Z_K - W_K\|_F^2)$$
$$+C(Z_K, W_K, E_K, Q) \quad (11)$$

where $C(Z_K, W_K, E_K, Q)$ is fixed. Equation (3) is in the quadratic form and D_K can be solved directly as follows:

$$D_K^{j+1} = \gamma D_K^j + (1-\gamma)D_K^{update} \quad (12)$$

where $D_K^{update} = \frac{1}{\mu}(Y_K + \mu(X_K - E_K))Z_K^T(Z_K Z_K^T)^{-1}$. We initialize the dictionary using KSVD method on training samples of each class and combining all the classes.

4. RESULTS AND DISCUSSION

The performance of MM-SLDL method is evaluated on three face datasets. We compare our method with three types of methods: (1)Multi-modal LR-DL method MLDL [9] (2)Multi-modal DL methods including UMD^{2}L [8] and MSDL [15] (3)Single-modality LR-DL methods such as JP-LRDL [3], D^{2}L^{2}R^2 [4] and SLRDL [5]. For constructing the

---

Algorithm 1 MM-SLDL Method in the Kth Modality

Input: Data X_K, Parameters $\lambda, \beta, \alpha, \gamma$
Output: D_K, Z_K
1: **Initialize:** D_K^0; $Z_K^0 = W_K^0 = E_K^0 = Y_K^0 = M_K^0 = 0$; $\mu = 10^{-6}$; $max_\mu = 10^{30}$; $\epsilon_s = 10^{-8}$; $\rho = 1.1$; $\epsilon_d = 10^{-5}$
2: **while** not converged **do**
3:     Fix other variables and update Z_K by Equation (3)
4:     Fix other variables and update W_K by Equation (9)
5:     Fix other variables and update E_K by Equation (3)
6:     Update Y_K, M_K as:
        $Y_K = Y_K + \mu(X_K - D_K Z_K - E_K)$
        $M_K = M_K + \mu(Z_K - W_K)$
7:     Update μ as: $\mu = min(\rho\mu, max_\mu)$
8:     Check stopping conditions as:
  $\|X_K - D_K Z_K - E_K\|_\infty < \epsilon_s$ and $\|Z_K - W_K\|_\infty < \epsilon_s$
9: **end while**
10: **while** not converged **do**
11:     Fix other variables and update D_K by Equation (12)
        $\|D_K^{j+1} - D_K^j\|_\infty < \epsilon_d$
12: **end while**

---

training set, we select images randomly and the selection is repeated 10 times and we report the average recognition rates for all methods. We set the number of dictionary atoms of each class as training size, and choose the tuning parameters of all methods by 5-fold cross validation. Convolutional Neural Networks (CNNs) have significantly improved the face recognition rates, and the most important ingredient for the success of such methods is the availability of large quantities of training data; however, transfer learning is a powerful tool to train small target datasets. [18] revealed when the target dataset is small and similar to original dataset, it is better to treat CNN as fixed feature extractor and train a linear classifier on the CNN features. We compare MM-SLDL with two deep methods (1)Deep features generated by VGG-Face descriptor [17], that is based on a 16-layer CNN trained on 2.6M images, followded by a nearest neighbor classifier. (2)We use 8-layer AlexNet [16] trained on 1.2M images of the ImageNet dataset, and fine-tune it on the target data.

AR Dataset [19] includes over $4,000$ face images from 126 individuals, 26 images for each person in two sessions. Among the images of each session, 3 are obscured by scarves, 3 by sunglasses, and the remaining faces are of different facial expressions or illumination variations, which we refer to as unobscured images. Following [5], experiments are con-

Table 1: Recognition rates (%) on AR dataset

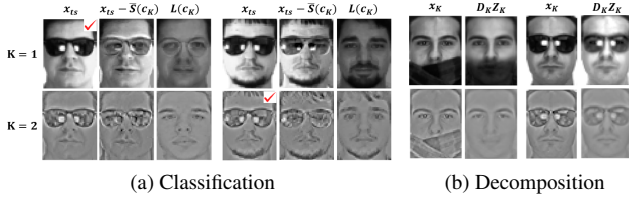| Method	Sunglasses	Scarf	Mixed	Misc.
MLDL [9]	90.51	91.51	91.32	76.33
UMD^{2}L [8]	88.26	87.40	88.30	71.30
MSDL [15]	83.20	80.65	79.50	68.44
D^{2}L^{2}R^2 [4]	92.20	90.40	91.30	75.30
SLRDL [5]	87.35	83.40	82.47	72.30
JP-LRDL [3]	93.20	93.00	93.30	78.23
AlexNet [16]	30.33	30.12	30.17	25.55
VGG-Face [17]	85.90	85.01	87.30	79.83
MM-SLDL	**96.70**	**96.41**	**96.30**	**85.30**

(a) Classification   (b) Decomposition

Fig. 1: Image decomposition and classification on AR dataset

ducted under three scenarios. **Sunglasses:** We select 7 unobscured images and 1 with sunglasses from the first session as training samples for each person, and the rest of unobscured and sunglasses images are used for testing. **Scarf:** We choose 8 training images (7 unobscured and 1 with scarf) from the first session for training, and 12 test images including the rest of unobscured and scarf images. **Mixed:** We select 7 unobscured, plus 2 occluded images (1 with sunglasses, 1 by scarf) from the first session for training, and the remaining 17 images in two sessions for testing. We design a challenging scenario **Misc.**, in which we select 7 unobscured, and 1 scarf images from the first session for training, and utilize the remaining 7 unobscured and 6 sunglasses images for testing. Here, the type of noise is different in training and test sets. According to Table 1, MM-SLDL achieves the best performance in all scenarios, and the improvement is significant in "Misc." scenario, where all the other methods fail. Fig. 1b illustrates
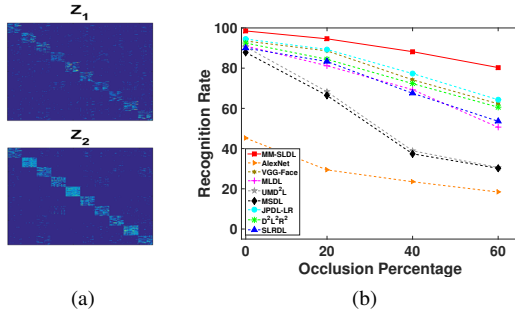

(a)   (b)

Fig. 2: (a) Representations for testing samples and (b) Recognition rates (%) on Extended YaleB dataset

examples of image decomposition on AR dataset. The first and second rows show training images and learned LR component $D_K Z_K$ in two modalities. While the first modality keeps more details, the illumination invariant modality better separates occlusions from the original images; hence, a robust representation is learned by their fusion. Fig. 1a demonstrates a testing sample x_{ts}, and $x_{ts} - \bar{S}(c_K), L(c_K)$ components, which their difference determines the winner, that is illustrated by a red tick mark.

Extended YaleB Dataset [20] contains $2,414$ face images of 38 human subjects captured under different illumination conditions. There are $59 \sim 64$ images for each subject, and we randomly select 20 of them for training. We simulate various levels of contiguous occlusion from 20% to 60%, by replacing a randomly located square block of each train image with an unrelated image, as seen in Fig. 3a. To have a real

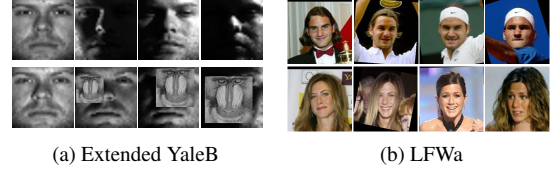
(a) Extended YaleB   (b) LFWa

Fig. 3: Sample images of Extended YaleB and LFWa datasets

challenge, test images are not occluded. We visualize the representation Z for two modalities for testing images of the first 10 classes under 40% occlusion training scenario in Fig. 2a. Testing images automatically generate a block diagonal structure, and the second modality learns a better representation here. Fig. 2b illustrates the recognition rates of all methods across different occlusion levels, and MM-SLDL outperforms other counterparts, especially for severely occluded images.

LFW Dataset [21] contains $13,233$ unconstrained face images of $5,749$ different individuals, collected from the web with large variations in pose, expression, illumination, clothing, hairstyles, occlusion, etc. We use an aligned version of LFW called LWFa [22], and exploit 143 subject with no less than 11 samples per subject to perform the experiment. Some of these images are shown in Fig. 3b. A central 170×140 region is cropped from each of images and the first 10 samples per class are selected for training, while the rest is used for testing. Table 2 shows the recognition rates of all compared methods. Although VGG-Face has already been trained on extra 2.6M images, MM-SLDL achieves competitive results using too much smaller training data, which have large intra-class variation. Also, as expected fine-tuned AlexNet is prone to overfitting because target data is small and very different in content compared to the ImageNet. Finally, to verify the role of illumination invariant modality, we just use $K = 2$ in the objective function (2) and change the ideal representation term to $\|Z_K - Q\|_F^2$. The results are reported under *SLDL-Mod2*, and as observed is not competitive.

Table 2: Recognition rates (%) on LFWa dataset

Method	Rec. Rate	Method	Rec. Rate
MLDL [9]	74.10	UMD^{2}L [8]	70.43
MSDL [15]	64.25	D^{2}L^{2}R^2 [4]	75.20
SLRDL [5]	74.20	JP-LRDL [3]	79.87
AlexNet [16]	40.31	VGG-Face [17]	90.01
SLDL-Mod2	76.77	**MM-SLDL**	**88.04**

5. CONCLUSIONS

We proposed a face recognition method that learns discriminative dictionaries and structured sparse LR representations from contaminated face image in two modalities. Adopting the illumination invariant representation of images as a modality, additionally empowers the model. Experimental results indicate that MM-SLDL is robust, achieving state-of-art performance in the presence of occlusion, illumination and pose changes, using a few training samples.

6. REFERENCES

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[2] Moein Shakeri and Hong Zhang, "Corola: a sequential solution to moving object detection using low-rank approximation," *Computer Vision and Image Understanding*, vol. 146, pp. 27–39, 2016.

[3] Homa Foroughi, Nilanjan Ray, and Hong Zhang, "Object classification with joint projection and low-rank dictionary learning," *arXiv preprint arXiv:1612.01594*, 2016.

[4] Liangyue Li, Sheng Li, and Yun Fu, "Discriminative dictionary learning with low-rank regularization for face recognition," in *Automatic Face and Gesture Recognition, 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[5] Yangmuzi Zhang, Zhuolin Jiang, and Larry Davis, "Learning structured low-rank representations for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 676–683.

[6] David L Hall and James Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[7] Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhe, Sylvain Lesage, and Rémi Gribonval, "Learning multi-modal dictionaries: Application to audiovisual data," in *International Workshop on Multimedia Content Representation, Classification and Security*. Springer, 2006, pp. 538–545.

[8] Xiao-Yuan Jing, Ruimin Hu, Fei Wu, Xi-Lin Chen, Qian Liu, and Yong-Fang Yao, "Uncorrelated multi-view discrimination dictionary learning for recognition.," in *AAAI*, 2014, pp. 2787–2795.

[9] Fei Wu, Xiao-Yuan Jing, Xinge You, Dong Yue, Ruimin Hu, and Jing-Yu Yang, "Multi-view low-rank dictionary learning for image classification," *Pattern Recognition*, vol. 50, pp. 143–154, 2016.

[10] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and William Kenneth Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.

[11] Moein Shakeri and Hong Zhang, "Illumination invariant representation of natural images for visual place recognition," in *Intelligent Robots and Systems, 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 466–472.

[12] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma, "Robust recovery of subspace structures by low-rank representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 171–184, 2013.

[13] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary, "Tikhonov regularization and total least squares," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 1, pp. 185–194, 1999.

[14] Zhouchen Lin, Risheng Liu, and Zhixun Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Advances in neural information processing systems*, 2011, pp. 612–620.

[15] Mehrdad J Gangeh, Pouria Fewzee, Ali Ghodsi, Mohamed S Kamel, and Fakhri Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1056–1068, 2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition.," in *BMVC*, 2015, vol. 1, p. 6.

[18] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

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

[20] Athinodoros S. Georghiades, Peter N. Belhumeur, and David Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[21] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report, University of Massachusetts, Amherst, 2007.

[22] Lior Wolf, Tal Hassner, and Yaniv Taigman, "Similarity scores based on background samples," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 88–97.