2013 Second IAPR Asian Conference on Pattern Recognition

Joint Alignment and Clustering via Low-Rank Representation

Qi Li, Zhenan Sun, Ran He, Tieniu Tan
Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
Email: $\{$qli, znsun, rhe, tnt$\}$@nlpr.ia.ac.cn

***Abstract*—Both image alignment and image clustering are widely researched with numerous applications in recent years. These two problems are traditionally studied separately. However in many real world applications, both alignment and clustering results are needed. Recent study has shown that alignment and clustering are two highly coupled problems. Thus we try to solve the two problems in a unified framework. In this paper, we propose a novel joint alignment and clustering algorithm by integrating spatial transformation parameters and clustering parameters into a unified objective function. The proposed function seeks the lowest rank representation among all the candidates that can represent misaligned images. It is indeed a transformed Low-Rank Representation. As far as we know, this is the first time to cluster the misaligned images using the transformed Low-Rank Representation. We can solve the proposed function by linearizing the objective function, and then iteratively solving a sequence of linear problems via the Augmented Lagrange Multipliers method. Experimental results on various data sets validate the effectiveness of our method.**

***Keywords*-joint alignment and clustering; Low-Rank Representation; Augmented Lagrange Multiplier method;**

I. Introduction

In recent years, there is a dramatic increase in the amount of visual data with the development of Internet. The unprocessed visual data suffer from significant illumination variation, occlusion and misalignment [1], [2]. Among those problems, misalignment challenges many existing computer vision tasks. Images from the same object may be misclassified due to the lack of alignment. Seeking more efficient and effective solutions to solve the misalignment problem has attracted much attention in recent years. Learner-Miller's Congealing algorithm [3] employs a sum of entropy cost functions to minimize the parametric warp differences between an ensemble of images. The least squares congealing algorithm [4] is proposed to seek an alignment that minimizes the sum of squared distances between pairs of images. Sparse and low-rank decomposition [1] are used to solve the batch image alignment problem.

In many real world applications, one often encounters the situation where there are multiple object classes in an image ensemble. Traditional image alignment methods treat all of the images as a single class of objects which concerns about the increase of image similarity whereas neglects the discriminate information among them. Thus most of alignment algorithms have poor performance in dealing with complex data set. The problem addressed here is to align and cluster the complex data set, that is to remove geometrical variability whereas to preserve the useful discriminate information to cluster the images. Frey and Jojic's work [5] and Liu *et al.*'s work [6] have the greatest relevance to our work. In [5], a transformed mixture of Gaussian models is used to normalize the input data for global transformations and cluster the normalized data. In [6], a unified objective function which consists of the within-cluster difference and the between-cluster difference is proposed to simultaneously align and cluster misaligned images.

The main contributions of this work are summarized as follows. We integrate image alignment and clustering assignments into a unified objective function which can be seen as finding the lowest rank representation among all the candidates that can represent the misaligned images. Our algorithm inherits the benefit of clustering misaligned images in an unsupervised manner by using only pixel information. Experimental results on various data sets have validated the effectiveness of our algorithm in terms of both alignment performances and clustering results.

The remainder of this paper is organized as follows. Section II describes the technical details of our algorithm. Experimental results and analysis are presented in Section III. Finally we draw some conclusions in Section IV.

II. Joint Image Alignment and Clustering Algorithm

A. Subspace Recovery by Low-Rank Representation

Given an ensemble of well aligned images, a reasonable assumption is that the images are drawn from a mixture of several low-rank subspaces. Recent development of Low-Rank Representation (LRR) [7] algorithm provides promising applications in subspace clustering area. LRR algorithm seeks the lowest rank representation among all the candidates that can represent data samples as linear combinations of the bases in a given dictionary, and it has been proven that LRR algorithm can exactly recover true subspace structures of all the images under certain conditions.

For n well aligned images, we denote the operator $\mathbb{R}^{w\times h} \rightarrow \mathbb{R}^{m}$ as selecting an m-pixel region of interest from the i-th $(i = 1, \cdots, n)$ input image and stack it as a vector representing as I_i. Then we store all of the n well

978-1-4799-2190-4/13 $31.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.66

aligned images as $X = [I_1, ..., I_n] \in \mathbb{R}^{m\times n}$. The low-rank representation algorithm is formulated as:

$$\begin{aligned} \min_{Z,E} \quad & rank(Z) + \lambda \|E\|_{2,1}, \\ s.t. \quad & X = AZ + E, \end{aligned} \tag{1}$$

where $A = [A_1, \cdots, A_n]$ is a given dictionary, $Z = [Z_1, \cdots, Z_n]$ is the coefficient matrix with Z_i being the representation of X_i, $\|E\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} (E_{ij})^2}$ is used to model the sample specific corruptions and outliers, the parameter λ is used to balance the above two terms. LRR algorithm can recover the underlying subspaces of the well aligned images.

B. Joint Image Alignment and Clustering

Due to changes of poses in practical applications, images are usually misaligned with each other. Thus the above model has a poor performance in clustering misaligned images. Considering the assignment of clustering misaligned images, we propose a novel algorithm that can jointly align and cluster these misaligned images. Let us denote the warping parameters for the entire n misaligned images as $\tau = [\tau_1, \cdot\cdot\cdot, \tau_n]$. τ_i is a p-dimensional vector which allows the alignment from each image to a predefined common coordinate space depending on the specific transformation type. The warping function $W(x; \tau_i)$ takes the pixel x in the coordinate of the original input image and maps it to the coordinate of the common space. Thus the warped image vector of m-pixel region of interest from each input image $I_i(x)$ can be represented as $I_i(W(x; \tau_i)) \in \mathbb{R}^m$. For the rest of this paper, we use $I_i \circ \tau_i$ to represent $I_i(W(x; \tau_i))$ for convenience. The overall warped images can be represented as $X \circ \tau = [I_i \circ \tau_i, \cdots, I_i \circ \tau_i] \in \mathbb{R}^{m\times n}$.

Then we can model the joint image alignment and clustering problem in the following form:

$$\begin{aligned} \min_{Z,E,\tau} \quad & rank(Z) + \lambda \|E\|_{2,1}, \\ s.t. \quad & X \circ \tau = AZ + E. \end{aligned} \tag{2}$$

RASL algorithm [1] also formulates the image alignment algorithm as a low-rank optimization problem. The differences between RASL and our algorithm are that by setting $A = I$ and using the ℓ_0 norm of the matrix E, Equation (2) will become the same objective function used in RASL algorithm. Note that compared with RASL algorithm, our algorithm can be seen as a more general one and $l_{2,1}$ norm of the matrix E is used to model the sample specific errors so as to recover the underlying subspaces of misaligned images.

C. Solving The Optimization Problem

The above formulation is difficult to solve due to the discrete property of the rank function and the nonlinear nature of the transformation parameter. In order to solve the first problem, nuclear norm $\|Z\|_*$ is used to replace the rank minimization function. For the second problem, similar to Lucas-Kanade [8] algorithm, we can approximate the parameter τ: suppose that the current estimate of τ is known and then solve for increments of τ iteratively. For the current estimate of τ, we linearize the parameter by:

$$X \circ (\tau + \Delta\tau) = X \circ \tau + \sum\nolimits_{i=1}^{n} J_i \Delta\tau_i \varepsilon_i \varepsilon_i^T, \tag{3}$$

where $J_i = \frac{\partial}{\partial \xi}(I_i \circ \xi)|_{\xi=\tau_i} \in \mathbb{R}^{m\times p}$ is the Jacobian of the i-th image with respect to τ_i, $\Delta\tau_i \in \mathbb{R}^p$ is the increments of τ_i. ε_i is the standard basis for $\in \mathbb{R}^n$. Substitute Equation (3) into Equation (2) we get the following formulation:

$$\begin{aligned} \min_{Z,E,\Delta\tau} \quad & \|Z\|_* + \lambda \|E\|_{2,1}, \\ s.t. \quad & X \circ \tau + \textstyle\sum_{i=1}^{n} J_i \Delta\tau_i \varepsilon_i \varepsilon_i^T = AZ + E. \end{aligned} \tag{4}$$

Inspired by [7], we use the current estimate of the warped images $X \circ \tau$ as the dictionary A. Then Equation (4) becomes:

$$\begin{aligned} \min_{Z,E,\Delta\tau} \quad & \|Z\|_* + \lambda \|E\|_{2,1}, \\ s.t. \quad & X \circ \tau + \textstyle\sum_{i=1}^{n} J_i \Delta\tau_i \varepsilon_i \varepsilon_i^T = (X \circ \tau)Z + E. \end{aligned} \tag{5}$$

For computation convenience, we introduce an auxiliary variable M and convert Equation (5) to the following equivalent problem:

$$\begin{aligned} \min_{M,Z,E,\Delta\tau} \quad & \|M\|_* + \lambda \|E\|_{2,1}, \\ s.t. \quad & X \circ \tau + \textstyle\sum_{i=1}^{n} J_i \Delta\tau_i \varepsilon_i \varepsilon_i^T = (X \circ \tau)Z + E, \\ & Z = M, \end{aligned} \tag{6}$$

Then we can repeatedly linearize our estimate of τ and solve problem (6) to get the transformation parameter τ and the clustering parameter Z. Problem (6) can be solved by various methods. In this paper, we adopt the Augmented Lagrangian Multiplier (ALM) [9] method to solve the problem because of its fast speed and high accuracy. The augmented Lagrangian functions take the form of:

$$\begin{aligned} f(Z, E, \Delta\tau) &= X \circ \tau + \textstyle\sum_{i=1}^{n} J_i \Delta\tau_i \varepsilon_i \varepsilon_i^T - (X \circ \tau)Z - E, \\ g(M, Z) &= Z - M. \end{aligned} \tag{7}$$

By applying the ALM method, we can rewrite Equation (6) as:

$$\begin{aligned} & \mathcal{L}_\mu(M, Z, E, \Delta\tau, Y_1, Y_2) = \|M\|_* + \lambda \|E\|_{2,1}, \\ & + tr\left(Y_1^T f(Z, E, \Delta\tau)\right) + tr\left(Y_2^T g(M, Z)\right), \\ & + \tfrac{\mu}{2}\left(\|f(Z, E, \Delta\tau)\|_F^2 + \|g(M, Z)\|_F^2\right), \end{aligned} \tag{8}$$

where Y_1 and Y_2 are Lagrangian multipliers, tr represents the trace of a matrix, $\mu > 0$ is a penalty parameter. The above unconstrained optimization problem can be minimized through an alternative strategy with respect to M, Z, E and $\Delta\tau$ by fixing the other variables and then update Y_1, Y_2 and

μ as the following form:

$$M^{k+1} = \arg\min_{M} \;\; L_{\mu^k}(M, Z^k, E^k, \Delta\tau^k, Y_1^k, Y_2^k), \tag{9}$$

$$Z^{k+1} = \arg\min_{Z} \;\; L_{\mu^k}(M^{k+1}, Z, E^k, \Delta\tau^k, Y_1^k, Y_2^k), \tag{10}$$

$$E^{k+1} = \arg\min_{E} \;\; L_{\mu^k}(M^{k+1}, Z^{k+1}, E, \Delta\tau^k, Y_1^k, Y_2^k), \tag{11}$$

$$\Delta\tau^{k+1} = \arg\min_{\Delta\tau} \;\; L_{\mu^k}(J^{k+1}, Z^{k+1}, E^{k+1}, \Delta\tau, Y_1^k, Y_2^k), \tag{12}$$

$$Y_1^{k+1} = Y_1^k + \mu^k f\left(Z^{k+1}, E^{k+1}, \Delta\tau^{k+1}\right), \tag{13}$$

$$Y_2^{k+1} = Y_2^k + \mu^k g\left(M^{k+1}, Z^{k+1}\right), \tag{14}$$

$$\mu^{k+1} = \rho\mu^k, \tag{15}$$

where ρ is an incremental factor for the parameter μ. At each step we solve Equation (9)-(15) to get the corresponding parameters and then update the parameters until the whole process converged.

The parameter M can be obtained by solving Equation (9) using the soft threshold methods [10]:

$$\begin{aligned} M^{k+1} &= U\mathcal{S}_{\frac{1}{\mu^k}}\left[\Sigma\right]V^T, \\ (U, \Sigma, V) &= \text{svd}\left(Z^k + Y_2^k/\mu^k\right), \end{aligned} \tag{16}$$

where $\mathcal{S}$ denotes the soft threshold operator which acts elementwise. Equation (10) can be solved efficiently as a least squares problem:

$$\begin{aligned} &Z^{k+1} = \Big((X\circ\tau)^T X\circ\tau + I\Big)^{-1} \\ &\Big((X\circ\tau)^T(X\circ\tau + \textstyle\sum_{i=1}^n J_i\Delta\tau_i^k\varepsilon_i\varepsilon_i^T - E^k) \\ &+M^{k+1} + (X\circ\tau)^T Y_1^k/\mu^k - Y_2^k/\mu^k\Big) \end{aligned} \tag{17}$$

Equation (11) can be solved through the following lemma.

Lemma 1: ([7], [11]) Let Q be a given matrix. If the optimal solution to

$$\min_W \;\; \alpha\|W\|_{2,1} + \frac{1}{2}\|W - Q\|_F^2$$

is W^*, then the i-th column of W^* is

$$W^*_{:,i} = \begin{cases} \frac{\|Q_{:,i}\|_2 - \alpha}{\|Q_{:,i}\|_2} Q_{:,i}, & \text{if } \|Q_{:,i}\|_2 > \alpha \\ 0, & \text{otherwise} \end{cases}$$

Then we can write the solution to Equation (11) as:

$$E_{:,i} = \begin{cases} \frac{\|K_{:,i}\|_2 - \alpha}{\|K_{:,i}\|_2} K_{:,i}, & \text{if } \|K_{:,i}\|_2 > \frac{\lambda}{\mu} \\ 0, & \text{otherwise} \end{cases}$$

where $K = X\circ\tau + \sum_{i=1}^n J_i\Delta\tau_i^k\varepsilon_i\varepsilon_i^T - X\circ\tau * Z^{k+1} + Y_1^k/\mu^k$. Equation (12) can be solved as a least squares problem:

$$\begin{aligned} \Delta\tau^{k+1} = &\textstyle\sum_{i=1}^n J_i^+\left(X\circ\tau * Z^{k+1} + E^{k+1}\right. \\ &\left.-X\circ\tau - Y_1^{k+1}/\mu^k\right)\varepsilon_i\varepsilon_i^T. \end{aligned} \tag{18}$$

Algorithm 1 The framework of our algorithm.

```
Input:
    The set of the misaligned images $I_1, \cdots, I_n$, initial transformations $\tau_1, \cdots, \tau_n$.
while not converged do
    Compute the Jacobian matrix with respect to the specific transformation:
    $J_i = \frac{\partial}{\partial\xi}\left(\frac{I_i\circ\xi}{\|I_i\circ\xi\|_2}\right)|_{\xi=\tau_i}, i = 1, \cdots, n.$
    Warp and normalize the images:
    $D\circ\tau = [\frac{I_1\circ\tau_1}{\|I_1\circ\tau_1\|_2}|\cdots|\frac{I_n\circ\tau_n}{\|I_n\circ\tau_n\|_2}].$
    Solve for the parameters $J$, $Z$, $E$, $\Delta\tau$:
        $\min_{M,Z,E,\Delta\tau} \quad \|M\|_* + \lambda\|E\|_{2,1},$
        $s.t. \quad X\circ\tau + \sum_{i=1}^n J_i\Delta\tau_i\varepsilon_i\varepsilon_i^T = (X\circ\tau)Z + E,$
        $\quad Z = M.$
    Update the parameter: $\tau = \tau + \Delta\tau$.
end while
Output:
    The final parameter $\tau = [\tau_1, \cdots, \tau_n]$ and $Z = [Z_1, \cdots, Z_n]$.
```

where J_i^+ denotes the Moore-Penrose pseudo inverse of J_i. The framework of our algorithm is summarized in Algorithm 1. After obtaining the clustering parameter Z, we use the method proposed by [7] to get the final clustering result.

III. Experiments

In order to evaluate the performance of our algorithm, we have conducted experiments on the MNIST data set [12] and the Labeled Faces in the Wild (LFW) data set [13]. The parameters ρ and μ_0 are set to be 1.1 and 10^{-6} respectively. The stopping criterion of the inner loop of our algorithm is that the difference value of the cost function between two consecutive iterations is less than 10^{-7}. The parameter λ and the stopping criterion of the outer loop are tuned empirically.

A. Results on The MNIST Data Set

In this experiment, we validate performances of our algorithm by aligning and clustering the images from the MNIST data set. We choose 200 images from 10 digit classes which are also used by Liu *et al.* [6]. Some of the digits are showed in Figure 1(a). In order to align and cluster the digits, we set the initial transformation as identity matrix, and Euclidean transformation is used in this experiment. The results are evaluated both visually and quantitatively based on the metrics suggested by [6]. Alignment score measures the distance between pairs of the warped images which are assigned to the same cluster. The mean and standard deviation of all the distances are reported. Rand index is used to evaluate the clustering accuracy with respect to the correct labels. It is computed from the estimated membership

Algorithm	Alignment	Clustering
TIC	6.0 ± 1.1	35.5%
USAC [6]	3.8 ± 0.9	56.5%
SSAC [6]	not reported	73.7%
RASL [1]+K-means	3.3 ± 1.4	68.0%
Our algorithm	3.1 ± 1.6	74.0%

Table I: Alignment and clustering results on the MNIST data set. All of the alignment scores are divided by 10^6. The results of TIC, USAC and SSAC are reported in [6].

vectors. Confusion matrix is also used to evaluate the clustering performance of different algorithms.

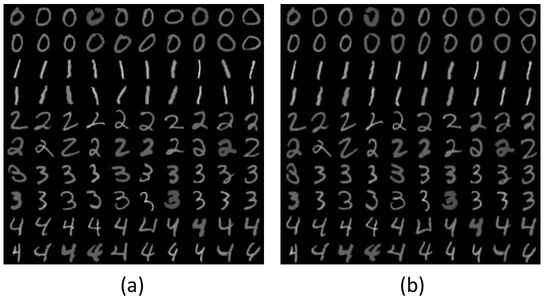

Figure 1: (a) Some of the digits before alignment. (b) The corresponding digits after alignment using our algorithm.

The compared algorithms are briefly introduced as follows. The Transformation-invariant Clustering (TIC) [5] algorithm is used as the baseline in [6]. Unsupervised simultaneous alignment and clustering algorithm (USAC) and semi-supervised simultaneous alignment and clustering algorithm (SSAC) are proposed by [6]. The difference between USAC and SSAC is that parts of the ensemble of digits are manually labeled in terms of clustering labels for the SSAC algorithm. We also compare our algorithm with RASL [1] + K-means algorithm: digits are aligned using the Euclidean transformation according to their released code (the parameters are set as default), then clustering performances are reported using the best of 100 K-means runs.

Table I summarizes the alignment score and clustering accuracy. Figure 1(b) shows some of the digits after alignment using our algorithm. Figure 2 plots the average digits after alignment using the estimated labels. From Table I, we can see that the clustering performance of our algorithm is the best and the alignment performance is similar to RASL algorithm. The reason is that both RASL and our algorithm seek the low-rank representation among the images whereas our algorithm is a more general form which can be used in clustering misaligned images. We can also find that our algorithm is even better than SSAC which utilizes the manually labeled information in terms of the clustering accuracy. Besides, we used the pixel information directly rather than the HOG features used in [6]. Figure 3 further plots the confusion matrices of different algorithms. As shown in Figure 3, when our algorithm is used to cluster the digits, the between-class similarity of digit ”2” and digit ”7” is higher than the within-class similarity of digit ”6”. Thus digit ”2” and digit ”7” are confused, and digit ”6” is clustered into two different clusters. Most of the other digits are clustered correctly by our algorithm.

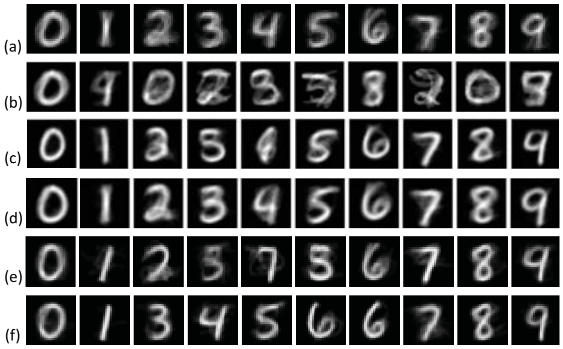

Figure 2: Average digits before and after alignment. (a) The average digits before alignment using the ground truth cluster labels. The average digits after alignment using the estimated cluster labels by (b) TIC, (c) USAC, (d) SSAC, (e) RASL+K-means and (f) our algorithm.

B. Results on The LFW Data Set

We also pursue an evaluation of our algorithm on the LFW data set. The LFW data set is taken under the unconstrained environments with variability in pose, lighting and occlusion. We choose 5 subjects from this data set, and each of the selected subjects has 35 images. We obtain the initial estimate of the transformation by using the Viola-Jones face detector [14]. Then affine transformation is used to align the images. We compare our algorithm with RASL+K-means algorithm in this experiment.

The average images and clustering accuracy are used to evaluate different algorithms. Figure 4 plots the average images using the estimated cluster labels. From Figure 4 we can see that the average images of our algorithm is better than RASL+K-means algorithm though RASL is useful for alignment and K-means is powerful for clustering. The clustering accuracy of our algorithm is 69% whereas the RASL+K-means is 46%. From this experiment we can conclude that our algorithm is effective for aligning and clustering the complex data set, such as the digit classes and faces taken under unconstrained environment.

IV. Conclusion

In this paper, we have proposed an efficient joint alignment and clustering algorithm via transformed Low-Rank Representation. We model the misalignments as domain

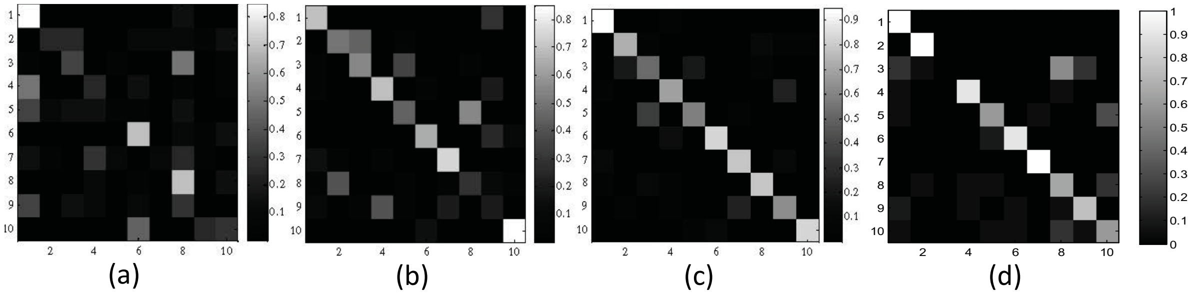

Figure 3: The confusion matrices of digits estimated by (a) TIC, (b) USAC, (c) SSAC and (d) our algorithm. (a), (b) and (c) are published in [6]. The index "1-10" correspond to 10 digit classes.

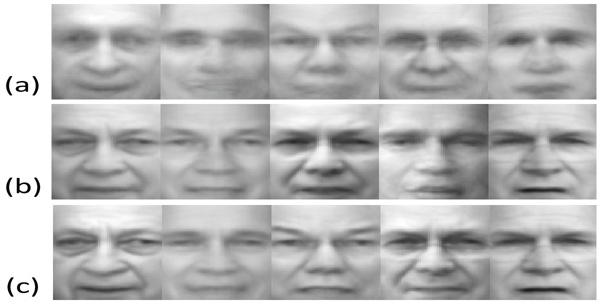

Figure 4: Average images on the LFW data set. (a) The average images before alignment using the ground truth labels. (b) The average images after alignment using the estimated cluster labels by RASL+K-means algorithm. (c) The average images after alignment using the estimated cluster labels by our algorithm.

transformations, and integrate the domain transformations into Low-Rank Representation. Then a unified objective function is proposed to cluster the misaligned images. The Augmented Lagrange Multiplier method is adopted to optimize the objective function. Experimental results on the MNIST and LFW data sets have validated the effectiveness of our algorithm.

Acknowledgment

This work is funded by the National Natural Science Foundation of China (Grant No. 61273272,61103155), the International S&T Cooperation Program of China (Grant No. 2010DFB14110) and the Instrument Developing Project of the Chinese Academy of Sciences (Grant No. YZ201266).

References

[1] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE TPAMI*, vol. 34, no. 11, pp. 2233–2246, 2012.

[2] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE TPAMI*, vol. 33, no. 8, pp. 1561–1576, 2011.

[3] E. G. Learned-Miller, "Data driven image models through continuous joint alignment," *IEEE TPAMI*, vol. 28, no. 2, pp. 236–250, 2006.

[4] M. Cox, S. Sridharan, S. Lucey, and J. Cohn, "Least squares congealing for unsupervised alignment of images," in *CVPR*, 2008, pp. 1–8.

[5] B. J. Frey and N. Jojic, "Transformation-invariant clustering using the em algorithm," *IEEE TPAMI*, vol. 25, no. 1, pp. 1–17, 2003.

[6] X. Liu, Y. Tong, and F. W. Wheeler, "Simultaneous alignment and clustering for an image ensemble," in *ICCV*, 2009, pp. 1327–1334.

[7] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE TPAMI*, vol. 35, no. 1, pp. 171–184, 2013.

[8] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, 1981.

[9] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[10] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[11] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 569–592, 2009.

[12] 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, 1998.

[13] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller *et al.*, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

[14] P. Viola and M. J. Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.