SUBSPACE CLUSTERING APPLIED TO FACE IMAGES

Constantine Kotropoulos[*], *Konstantinos Pitas*[†]

Aristotle University of Thessaloniki
Department of Informatics
Thessaloniki 54124, GREECE

costas@aiia.csd.auth.gr, kpitas@auth.gr

ABSTRACT

In this paper, two state-of-the-art subspace clustering techniques, namely the Sparse Subspace Clustering and the Elastic Net Subspace Clustering, are tested for clustering. Both algorithms are frequently implemented using the linearized alternating directions method. An efficient implementation of the Elastic Net Subspace Clustering is derived, employing the fast iterative shrinkage algorithm. Random projections are also used to reduce significantly the computation time. Figures of merit are reported for two publicly available face image datasets, i.e., the Extended Yale B dataset and the Hollywood dataset.

Index Terms— Subspace clustering, face clustering, clustering assessment

1. INTRODUCTION

Given face images of multiple subjects, face clustering aims at grouping the images of the same subject together. The subjects depicted in these images could have a fixed or varying pose. In addition, the illumination could vary during image acquisition. Face clustering is a challenging research topic in computer vision. It is applied to extract semantic information from videos, assisting video indexing and content analysis (e.g., facial expression recognition or human action recognition) as well as to preprocess images for face recognition and surveillance.

An agglomerative or bottom-up hierarchical clustering was applied to the similarity matrix based on the matching between scale-invariant features key-points [1]. The clustering quality was assessed with respect to the F_1 measure (that is, the harmonic mean of recall and precision rates), the overall entropy, and the Γ statistic [2], a well known cluster validity index. A recursive normalized cut algorithm with properly adjusted thresholds was applied to a similarity graph based on the mutual information between image pairs [3–5]. Next, tentative mergers between any two clusters created by the just-mentioned spectral graph clustering technique were examined. A new dissimilarity measure between two face images using their neighboring information in the dataset was proposed in [6]. The so-called rank-order distance is motivated by the observation that two faces of the same person tend to have many shared top neighbors. For each face, an ordered list is generated by sorting all other faces in the dataset. Then, the rank-order distance between two faces is calculated, using their ranking orders.

It is well known that various tasks in computer vision, such as motion segmentation, face clustering under varying illumination, handwritten character recognition, image segmentation and compression, and feature selection, may be solved as low-dimensional linear subspace clustering problems [7]. Subspace clustering or hybrid linear modeling [8] major premise is that the total variance of the data in the aforementioned tasks is contained in a small number of principal axes. Even if the measured data are high-dimensional, their intrinsic dimensionality is usually much lower. Accordingly, the data from different classes are assumed to lie in a union of linear or affine subspaces rather than in a single subspace. Several solutions to the subspace clustering problem have been proposed, such as the spectral curvature clustering [9], the sparse subspace clustering (SSC) [10, 11], and the low-rank representation clustering [12]. The former method describes every point by a set of sparse linear combinations of points from the same subspace. The sparsity information is then used as a point clustering affinity, while the latter tries to recover a low-rank representation of the data points, able to handle the effects of unobserved (i.e., "hidden") data, by solving a convex minimization problem. Another method is the so called discriminative subspace clustering, which solves the problem by using a quadratic classifier trained by unlabeled data (i.e, clustering by classification) [13]. Labels are generated by exploiting the locality of points from the same subspace and a basic affinity criterion. Several classifiers are then diversely trained from different partitions of the data and their results are combined together in an ensemble in order to obtain the final clustering result.

However, the classifiers that are based on either sparse representations (SR) or low-rank representations (LRR) are not the best choices for face clustering, when groups of a few contiguous dictionary atoms (i.e., column image vectors) are expected to be highly collinear. As a result, the sparsity constraint is not appropriate for selecting the relevant dictionary atoms efficiently, since the least absolute shrinkage and selection operator (LASSO) does not discriminate between collinear entries adequately. Neither the LRR is consistent with the underlying group collinear structure, because the LRR tends to produce a holistic dense representation. It has been shown that the elastic net (EN) criterion is able to handle collinear dictionary atoms [14]. Accordingly, a novel subspace clustering method that employs the joint elastic net representation of the features is exploited here for face clustering. Such a representation shares the same advantages with the SR and the LRR. That is, when the data are noise-free, the elastic net representation exhibits nonzero within-subspace affinities and zero between-subspace affinities. Here, the Elastic Net Subspace Clustering (ENSC) algorithm proposed in [14] is tested for face clustering. The ENSC algorithm extends the elastic net (i.e., the sum of ℓ_1 and squared ℓ_2 regularized regression

[*] Supported by the COST Action IC 1106 "Integrating Biometrics and Forensics for the Digital Age".

[†] Supported by a Short Term Scientific Mission Scholarship within the COST Action IC 1106.

in compressive sensing to the more general setting of matrix subspace recovery, employing the sum of the ℓ_1 norm and the squared Frobenius norm of a matrix. The joint EN representation is obtained as the solution of an appropriate convex problem by employing the convergent Linearized Alternating Directions Method (LADM) [15]. Having found the joint EN representation, an affinity matrix is constructed. Face clustering is revealed by applying the normalized cuts to the EN-based affinity matrix. Both the SSC and the ENSC are frequently implemented using the LADM. Here, a less time-consuming iterative algorithm for the ENSC than the LADM-based implementation is theoretically derived by employing the fast iterative shrinkage algorithm (FISTA) [16]. In addition, random projections [17] are used to reduce significantly the computation time.

Figures of merit are reported for face clustering by applying the SSC and the ENSC to two publicly available datasets, namely the Extended Yale B dataset [18] and the Hollywood Human Actions dataset [19]. The former database contains faces taken from the same viewpoint under varying illumination conditions, while the latter contains face images retrieved by using a face detector and face tracker that are totally uncalibrated. The experiments have been conducted for a *fixed* number of clusters, that defined in the associated ground truth for each dataset. For the Extended Yale B dataset, the experiments have been conducted by adhering to the experimental protocol set in [11]. The SSC implementation is that of the authors'[1]. It is demonstrated that ENSC performs comparably to the SSC. A slight superiority with respect to the clustering error is noticed for the subsets containing images of 8 and 10 individuals in the Extended Yale B dataset. By reducing the size of image vectors to one third using random projections significant time savings are obtained. Unlike the previous works addressed face image clustering in the Hollywood Human Actions dataset employing a number of clusters different than that of the ground truth [3–5], here the figures of merit reported for the SSC and the ENSC were obtained by fixing the number of clusters to that of the ground truth. Although different regularization parameters were employed in the LADM to solve the SSC for the subsets of face images extracted from the different movies, the ENSC is demonstrated to achieve comparable performance with the *same* regularization parameters in the LADM across the dataset.

The outline of the paper is as follows. Section 2 is devoted to Subspace Clustering. To make the paper self-contained, SSC is briefly reviewed and emphasis is given to the ENSC. Experimental results are reported in Section 3. Conclusions are drawn and future research directions are highlighted in Section 4.

2. SUBSPACE CLUSTERING

Let $\mathbf{Y} \in \mathbb{R}^{M \times N}$ be a matrix containing the data points (i.e., image representations as columns), where M is the image vector size and N is the number of images. $\mathbf{Y}$ is called dictionary hereafter. Let $\{\mathcal{G}_k\}_{k=1}^{K}$ be an arrangement of K linear subspaces of dimensions $\{d_k\}_{k=1}^{K}$. Denote the dictionary as

$$\mathbf{Y} \overset{\triangle}{=} [\mathbf{y}_1|\mathbf{y}_2|\ldots|\mathbf{y}_N] = [\mathbf{Y}_1|\mathbf{Y}_2|\ldots|\mathbf{Y}_K]\,\mathbf{\Gamma} \qquad (1)$$

where $\mathbf{Y}_k \in \mathbb{R}^{M \times N_k}$ in a rank-d_k matrix of the $N_k > d_k$ points that lie in $\mathcal{G}_k$, $\mathbf{\Gamma} \in \mathbb{R}^{N \times N}$ is an unknown permutation matrix, and $N = \sum_{k=1}^{K} N_k$. We assume that the bases of the subspaces is not known a priori nor is it known which data points belong to any subspace. In principle, the subspace clustering problem aims at finding the number of the subspaces, their dimensions, a basis for each subspace, and the segmentation of the data in $\mathbf{Y}$ [7, 11].

2.1. Sparse Subspace Clustering

Seeking a sparse representation of each data point in the dictionary in terms of the other points leads to selecting a few points from the same subspace where the data point lies to. Accordingly, SSC attempts to find a non-trivial representation of $\mathbf{y}_i$ by minimizing the tightest convex relaxation of the ℓ_0 norm [10, 11]:

$$\min ||\mathbf{c}_i||_1 \quad \text{s.t. } \mathbf{y}_i = \mathbf{Y}\,\mathbf{c}_i \text{ and } \mathbf{c}_{ii} = 0. \qquad (2)$$

For the entire dictionary, the sparse optimization problem (2) can be rewritten as

$$\min ||\mathbf{C}||_1 \quad \text{s.t. } \mathbf{Y} = \mathbf{Y}\,\mathbf{C} \text{ and } \operatorname{diag}(\mathbf{C}) = \mathbf{0} \qquad (3)$$

where $\mathbf{C} = [\mathbf{c}_1|\mathbf{c}_2|\ldots|\mathbf{c}_N] \in \mathbb{R}^{N \times N}$ is the matrix whose i-th column corresponds to the sparse representation of $\mathbf{y}_i$ and $\operatorname{diag}(\mathbf{C}) \in \mathbb{R}^{N \times 1}$ is the vector of the diagonal elements of $\mathbf{C}$. The constraint in (3) ensures that the result is not trivial. An efficient solution for the sparse optimization problem (3) using LADM was derived in [11]. Having found $\mathbf{C}$, one obtains a symmetric similarity (i.e., affinity) matrix as $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^T$. The ideal similarity matrix has a block diagonal structure

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{W}_K \end{bmatrix} \mathbf{\Gamma} \qquad (4)$$

where $\mathbf{W}_k$, $k = 1, 2, \ldots, K$ is the similarity matrix of the data points in $\mathcal{G}_k$. Accordingly, the associated graph has ideally K connected components corresponding to the subspaces. Clustering of the data into subspaces can be done by applying spectral clustering to $\mathbf{W}$ [20, 21]. Let $\mathbf{1} \in \mathbb{R}^N$ denote a vector of ones and $\mathbf{D}$ be the degree matrix, i.e., $\operatorname{diag}(\mathbf{D}) = \mathbf{W}\mathbf{1}$. In [20], the K-means algorithm [22] is applied to the row vectors of the matrix formed by the bottom K eigenvectors of $\mathbf{D}^{-1}\mathbf{L}$, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the unnormalized graph Laplacian. In [21], the K-means algorithm is applied to the row vectors of the matrix formed by the bottom K eigenvectors of the normalized graph Laplacian $\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$. The SSC is well documented and has achieved impressive performance in many clustering applications [11].

2.2. Elastic Net Subspace Clustering

In the ideal case, if the data points belong to an arrangement (i.e., a cluster), they will lie into the same union of subspaces. Accordingly, it is assumed that $\mathbf{y}_i$ are drawn from a union of K unions of independent linear subspaces of unknown dimensions. Moreover, groups of a few contiguous $\mathbf{y}_i$ are expected to be quite similar and thus highly correlated. Based on the just mentioned assumptions, one would like to learn the representation matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$, such that $\mathbf{Y} = \mathbf{YC}$, with $c_{ij} = 0$ if $\mathbf{y}_i$ and $\mathbf{y}_j$ lie on different unions of subspaces and nonzero c_{ij} otherwise. Such a representation matrix $\mathbf{C}$ measures the similarity between all the features, unveiling the hidden subspace structure. It is obtained by solving [14]:

$$\underset{\mathbf{C}}{\operatorname{argmin}} \ \ \lambda_1\|\mathbf{C}\|_1 + \frac{\lambda_2}{2}\|\mathbf{C}\|_F^2 \quad \text{s.t. } \mathbf{Y} = \mathbf{YC} \text{ and } c_{ii} = 0. \quad (5)$$

[1] www.cis.jhu.edu/~ehsan/Codes/SSC_ADMM_v1.1.zip

In (5), the matrix ℓ_1-norm is defined as $\|\mathbf{Z}\|_1 = \sum_i \sum_j |z_{ij}|$ and $\|\mathbf{Z}\|_F = \sqrt{\sum_i \sum_j z_{ij}^2}$ denotes the Frobenius norm. (5) is a combination of the matrix ℓ_1-norm and squared Frobenius norm. Accordingly, it is actually an extension of the vector elastic net regularizer [23] to matrices and admits nonzero entries for within-subspace affinities and zero entries for between-subspace affinities.

In practice, the assumption $\mathbf{Y} = \mathbf{Y}\mathbf{C}$ does not hold exactly, because the data are *approximately* drawn from unions of subspaces. This fact introduces deviations from the ideal modeling assumptions. The latter can be treated collectively as additive *noise* contaminating the ideal model i.e., $\mathbf{Y} = \mathbf{Y}\mathbf{C} + \mathbf{E}$. To account for the noise, a distortion term is inserted into (5) and a robust solution is sought for the following convex optimization problem:

$$\begin{aligned} &\underset{\mathbf{C},\mathbf{E}}{\text{argmin}} \quad \lambda_1\|\mathbf{C}\|_1 + \frac{\lambda_2}{2}\|\mathbf{C}\|_F^2 + \lambda_3\|\mathbf{E}\|_1 \\ &\text{s.t. } \mathbf{Y} = \mathbf{Y}\,\mathbf{C} + \mathbf{E} \text{ and } c_{ii} = 0 \end{aligned} \tag{6}$$

where $\lambda_3 > 0$ is a regularization parameter. To efficiently solve (6), the LADM [15] is employed, which is suitable for large scale optimization problems. By applying the LADM, one seeks to minimize the (partial) augmented Lagrangian function:

$$\begin{aligned} &\underset{\mathbf{C},\mathbf{E}}{\text{argmin}} \;\; \mathcal{L}(\mathbf{C},\mathbf{E},\boldsymbol{\Xi}) = \lambda_1\|\mathbf{C}\|_1 + \frac{\lambda_2}{2}\|\mathbf{C}\|_F^2 + \lambda_3\|\mathbf{E}\|_1 \\ &+\text{tr}\left(\boldsymbol{\Xi}^T(\mathbf{Y} - \mathbf{Y}\mathbf{C} - \mathbf{E})\right) + \frac{\mu}{2}\|\mathbf{Y} - \mathbf{Y}\mathbf{C} - \mathbf{E}\|_F^2, \\ &\text{s.t. } c_{ii} = 0 \end{aligned} \tag{7}$$

where $\boldsymbol{\Xi}$ gathers the Lagrange multipliers for the equality constraints in (6) and $\mu > 0$ is a penalty parameter. Let t denote the iteration index and σ be the largest singular value of $\mathbf{Y}$. Then, (7) is minimized with respect to each variable in an alternating fashion, as outlined in Algorithm 1.

Following [15], since (8) does not admit a closed-form solution, the smooth term in (7) is linearly approximated and a closed-form solution (9) has been derived [14]. The approximate solution (9) employs the shrinkage operator $\mathcal{S}_\tau[q] = \text{sgn}(q)\text{max}(|q| - \tau, 0)$ [24], which can be extended to matrices by applying it element-wise. Similarly, a closed-form solution for the optimization problem (11) is obtained by applying the shrinkage operator (12). The diagonal elements of $\mathbf{C}_{[t+1]}$ are set to zero in (10) in order to fulfil the constraint in (7). In Algorithm 1, the internal parameter θ is made data dependent, i.e., $\theta = 1.02\sigma^2$; the parameters λ_1, λ_2, λ_3 and ρ are set by a grid searching in the Extended Yale B and the Hollywood Human Actions datasets. Regarding the parameters related to the stoping conditions in Algorithm 1, $\epsilon_1 = 10^{-4}$ and $\epsilon_2 = 10^{-5}$ are typical choices [15]. The penalty parameter μ is updated in line 6 of Algorithm 1.

Having found $\mathbf{C}$, its column space is useful for clustering. Let $\mathbf{C} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the singular value decomposition of $\mathbf{C}$ and $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{U}^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$. Then, an elastic net nonnegative symmetric affinity matrix $\mathbf{W} \in \mathbb{R}_+^{N\times N}$ has elements [12]:

$$w_{ij} = m_{ij}^2. \tag{13}$$

The segmentation of the columns of $\mathbf{Y}$ into K clusters is performed by applying spectral clustering to $\mathbf{W}$.

A more efficient solution is obtained by alternating between the solution of the problem with respect to $\mathbf{C}$, keeping $\mathbf{E}$ fixed:

$$\arg\min_{\mathbf{C};\mathbf{E}} ||\mathbf{Y} - \mathbf{Y}\mathbf{C} - \mathbf{E}||_F^2 + \lambda_2||\mathbf{C}||_F^2 + \lambda_1||\mathbf{C}||_1 \;\; \text{s.t. } c_{ii} = 0 \tag{14}$$

Algorithm 1 Solving (7) by the LADM method.

Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{M\times N}$ and the parameters λ_1, λ_2, λ_3, and ρ.

Output: Matrices $\mathbf{C} \in \mathbb{R}^{N\times N}$ and $\mathbf{E} \in \mathbb{R}^{M\times N}$.

```
1: Initialize: $\mathbf{C}_{[0]} = \mathbf{0}, \mathbf{E}_{[0]} = \mathbf{0}, \boldsymbol{\Xi}_{[0]} = \mathbf{0}$, $\mu_{[0]} = 10^{-6}$, $\theta = 1.02\sigma^2$, $\epsilon_1 = 10^{-4}$, and $\epsilon_2 = 10^{-5}$.
2: while not converged do
3:     Fix $\mathbf{E}_{[t]}$, and update $\mathbf{C}_{[t+1]}$ by
           $\mathbf{C}_{[t+1]} = \underset{\mathbf{C}_{[t]}}{\text{argmin}}\, \mathcal{L}(\mathbf{C}_{[t]}, \mathbf{E}_{[t]}, \boldsymbol{\Xi}_{[t]})$   (8)
           $\approx \mathcal{S}_{\frac{\lambda_1}{\theta\mu_{[t]}}}\Big[\mathbf{C}_{[t]} + \frac{1}{\theta}\Big(\mathbf{Y}^T(\mathbf{Y} - \mathbf{Y}\mathbf{C}_{[t]} - \mathbf{E}_{[t]} + \frac{1}{\mu_{[t]}}\boldsymbol{\Xi}_{[t]}) - \frac{\lambda_2}{\mu_{[t]}}\mathbf{C}_{[t]}\Big)\Big].$   (9)
           $c_{ii[t+1]} = 0.$   (10)
4:     Fix $\mathbf{C}_{[t+1]}$ and update $\mathbf{E}_{[t]}$ by
           $\mathbf{E}_{[t+1]} = \underset{\mathbf{E}_{[t]}}{\text{argmin}}\, \mathcal{L}(\mathbf{C}_{[t+1]}, \mathbf{E}_{[t]}, \boldsymbol{\Xi}_{[t]})$   (11)
           $= \mathcal{S}_{\frac{\lambda_3}{\mu_{[t]}}}\Big[\mathbf{Y} - \mathbf{Y}\mathbf{C}_{[t+1]} + \frac{1}{\mu_{[t]}}\boldsymbol{\Xi}_{[t]}\Big]$   (12)
5:     Update the Lagrange multiplier by
           $\boldsymbol{\Xi}_{[t+1]} = \boldsymbol{\Xi}_{[t]} + \mu_{[t]}(\mathbf{Y} - \mathbf{Y}\mathbf{C}_{[t+1]} - \mathbf{E}_{[t+1]})$.
6:     Update $\mu_{[t+1]}$ by $\mu_{[t+1]} \leftarrow \text{min}(\rho \cdot \mu_{[t]}, 10^{10})$.
7:     Check convergence conditions
           $\frac{\|\mathbf{Y} - \mathbf{Y}\mathbf{C}_{[t]} - \mathbf{E}_{[t]}\|_F}{\|\mathbf{Y}\|_F} \le \epsilon_1$ and
           $\text{max}\Big(\frac{\|\mathbf{E}_{[t]} - \mathbf{E}_{[t-1]}\|_F}{\|\mathbf{Y}\|_F}, \frac{\|\mathbf{C}_{[t]} - \mathbf{C}_{[t-1]}\|_F}{\|\mathbf{Y}\|_F}\Big) \le \epsilon_2.$
8:     $t \leftarrow t + 1$.
9: end while
```

and the solution with respect to $\mathbf{E}$, keeping $\mathbf{C}$ fixed:

$$\arg\min_{\mathbf{E};\mathbf{C}} ||(\mathbf{Y} - \mathbf{Y}\,\mathbf{C} - \mathbf{E}||_F^2 + \lambda_3||\mathbf{E}||_1. \tag{15}$$

Let $F(\mathbf{C})$ denote the objective function in (14). $F(\mathbf{C})$ can be decomposed as $F(\mathbf{C}) = f(\mathbf{C}) + g(\mathbf{C})$, where $f(\mathbf{C}) = ||\mathbf{Y} - \mathbf{Y}\,\mathbf{C} - \mathbf{E}||_F^2 + \lambda_2||\mathbf{C}||_F^2$ and $g(\mathbf{C}) = \lambda_1||\mathbf{C}||_1$. It is seen that $f(\mathbf{C}) : \mathbb{R}^{N\times N} \to \mathbb{R}$ is a smooth convex function continuously differentiable with Lipschitz continuous gradient and Lipschitz constant $L(f)$, while $g(\mathbf{C}) : \mathbb{R}^{N\times N} \to \mathbb{R}$ is a continuous convex function. Following similar lines to [16], it can be shown that for $L > 0$ the optimization problem $\min F(\mathbf{C})$ admits the unique minimizer [16]

$$\begin{aligned} p_L(\mathbf{Z}) = \arg\min_{\mathbf{C}} \Bigg\{ g(\mathbf{C}) + \frac{L}{2}\text{tr}\Bigg[\Bigg(\mathbf{C} \\ -\Big(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\Big)\Bigg)^T \Big(\mathbf{C} - \Big(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\Big)\Big)\Bigg]\Bigg\} \end{aligned} \tag{16}$$

Algorithm 2 Solving (7) with the FISTA.

```
Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{M \times N}$ and the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $L = L(f)$.
Output: Matrices $\mathbf{C} \in \mathbb{R}^{N \times N}$ and $\mathbf{E} \in \mathbb{R}^{M \times N}$.

Initialize: $t = 1$, $\zeta_{[1]} = 1$, $\mathbf{C}_{[0]} = \mathbf{0}$, $\mathbf{E}_{[0]} = \mathbf{0}$, $\mathbf{Z}_{[1]} = \mathbf{C}_{[0]}$.
while not converged do
    $\mathbf{C}_{[t]} = \mathcal{S}_{\frac{\lambda_1}{L}}\left[\mathbf{Z}_{[t]} + \frac{2}{L}\Big(\mathbf{Y}^T(\mathbf{Y} - \mathbf{Y}\mathbf{Z}_{[t]} - \mathbf{E}_{[t-1]}) - \lambda_2 \mathbf{Z}_{[t]}\Big)\right]$
    $\mathbf{E}_{[t]} = \mathcal{S}_{\frac{\lambda_3}{2}}\left[\mathbf{Y} - \mathbf{Y}\mathbf{C}_{[t]}\right] \qquad (18)$
    $\zeta_{[t+1]} = \frac{1+\sqrt{1+4\zeta_{[t]}^2}}{2}$
    $\mathbf{Z}_{[t+1]} = \mathbf{C}_{[t]} + \frac{\zeta_{[t]}-1}{\zeta_{[t+1]}}\left(\mathbf{C}_{[t]} - \mathbf{C}_{[t-1]}\right)$
    $t \leftarrow t + 1$.
end while
```

where $\nabla f(\mathbf{Z}) = 2(\mathbf{Y}^T\mathbf{Y} + \lambda_2 \mathbf{I})\mathbf{Z} - 2\mathbf{Y}^T(\mathbf{Y} - \mathbf{E})$ and $\mathbf{I}$ is the identity matrix of compatible dimensions. The Lipschitz constant of ∇f is $L(f) = 2(\eta_{\max}(\mathbf{Y}^T\mathbf{Y}) + \lambda_2) = 2(\sigma^2 + \lambda_2) \leq 2(\mathrm{tr}(\mathbf{Y}^T\mathbf{Y}) + \lambda_2)$, because for the maximum eigenvalue of $\mathbf{Y}^T\mathbf{Y}$, $\eta_{\max}(\mathbf{Y}^T\mathbf{Y})$, it holds $\eta_{\max}(\mathbf{Y}^T\mathbf{Y}) = \sigma^2$. Accordingly, the iterative shrinkage-thresholding algorithm asserts [16]

$$\mathbf{C}_{[t+1]} = p_L(\mathbf{C}_{[t]}) = \arg\min_{\mathbf{C}} \Big\{ g(\mathbf{C}) + \frac{L}{2}||\mathbf{C} - \Big(\mathbf{C}_{[t]} - \frac{1}{L}\nabla f(\mathbf{C}_{[t]})\Big)||_F^2 \Big\} = \mathcal{S}_{\frac{\lambda_1}{L}}\Big[\mathbf{C}_{[t]} + \frac{2}{L}\Big(\mathbf{Y}^T(\mathbf{Y} - \mathbf{Y}\mathbf{C}_{[t]} - \mathbf{E}) - \lambda_2 \mathbf{C}_{[t]}\Big)\Big]. \qquad (17)$$

The optimization with respect to $\mathbf{E}$ in (15) yields the updating equation (18). The multistep version of an accelerated gradient-like method proposed in [25] can be exploited to come up with the fast iterative-shrinkage algorithm for solving the elastic net representation summarized in Algorithm 2.

Algorithm 2 is more efficient than Algorithm 1, because there is no need to update any Lagrange multipliers.

3. EXPERIMENTAL RESULTS

The ENSC algorithm was applied to face clustering on two publicly available datasets, namely the Extended Yale B dataset [18] and the Hollywood Human Actions dataset [19]. The implementation in Algorithm 1 was used. The performance of ENSC was compared to that of the SSC in both datasets. The SSC achieved the top performance in the former dataset [11].

The Extended Yale B dataset consists of 192 × 168 pixel cropped face images of $K = 38$ subjects. There are $N_k = 64$, $k = 1, 2, \ldots K$ frontal face images for each subject acquired under various lighting conditions. All images were downsampled to 48 × 42 pixels and treated as column vectors of size $M = 2016$. It has been found that the face images lie close to a union of 9-dimensional subspaces [11]. To study the effect of the number of subjects in the clustering performance of the ENSC and the SSC, we adhered to the experimental setting used in [11]. That is, the 38 subjects were divided into 4 groups, where the first three groups corresponded to subjects 1 to 10, 11 to 20, 21 to 30, and the fourth group included subjects 31 to 38. For each of the first three groups, all choices of $K \in \{2, 3, 5, 8, 10\}$ subjects were considered. For the last group, the choices $K \in \{2, 3, 5, 8\}$ were taken into account. For each trial (i.e., set of K subjects), both clustering algorithms were tested. Table 1 summarizes the clustering error, defined as the ratio of the number of misclassified face vectors over the total number of face vectors N, as in [11]. The data points were made zero-mean by centering and normalized to unit-norm, prior to the ENSC. To reduce the computational time of the ENSC, the size M of the data points was reduced to one third by random projections (ENSC-R) prior to centering and unit-norm normalization. The parameters for the ENSC were set as follows: $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$, and $\rho = 1.5$. The elastic net affinity matrix was further post-processed by applying a two-dimensional Gabor filter with angle $\pi/4$ in order to enhance any diagonal structures in it. It is seen that the deterioration in the clustering error is insignificant. That is, less than 1% for subjects fewer than 5, approximately 1% for 8 subjects, and 2.24% for 10 subjects. The performance quoted for SSC is that in [11, Table 5]. The ENSC without employing the random projections is ranked first for 8 and 10 subjects, second-best for 3 and 5 subjects, and third-best after the SSC and the Low-Rank Recovery with heuristic (LLR-H) [11].

Table 1. Clustering error (%) of different algorithms on the Extended Yale B dataset.

Algorithm	SSC	ENSC	ENSC-R
2 Subjects			
Mean	1.86	3.41	3.61
Median	0	2.34	1.56
3 Subjects			
Mean	3.10	4.06	4.31
Median	1.04	3.65	3.65
5 Subjects			
Mean	4.31	4.58	4.73
Median	2.50	4.06	3.91
8 Subjects			
Mean	5.85	5.57	6.59
Median	4.49	4.30	3.91
10 Subjects			
Mean	10.94	6.67	8.91
Median	5.63	4.22	3.91

The Hollywood Human Actions dataset [19] consists of 32 movie clips. Only 23 of the movies were used, as in [5]. Face images were retrieved by means of a face detector and a face tracker in each movie clip. Actors' faces are depicted at varying poses and illumination. The faces are not always aligned to the camera. All images were downsampled to 60 × 60 pixels and treated as column vectors of size $M = 3600$. The number of face images in each movie clip is indicated in Table 2. Clustering was performed by setting K equal to the ground truth value. The number of different actors tracked in each movie (i.e., K) varies between 2 and 10 individuals. The parameter `alpha` in the SSC code[1], which controls the penalty parameters λ_1, λ_2, and λ_3 in the LADM solving the SSC optimization problem, was varying among the movies. Its values are listed in Table 2. The data points were made zero-mean by centering and normalized to unit-norm before the application of the ENSC. The parameters for the ENSC were set as follows: $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, $\lambda_3 = 0.1$, and $\rho = 1.1$ for all face image subsets (i.e., the face images detected and tracked in each movie clip). These parameters were determined by grid searching in the first 4 face image subsets. Both spectral graph clustering variants [20, 21] were applied to the affinity matrix of the EN. The spectral graph clustering variant yielding he smallest clustering error was chosen. The clustering error and the average F_1 measure, $\overline{F_1}$, were used as figures of merit. Let $\mathcal{G}_k$ denote a class according to the ground truth and $\hat{\mathcal{G}}_k$ be the cluster created by an algorithm, such that $k = \mathrm{map}(\kappa)$, where the right-hand side refers to the permutation mapping function that maps each cluster label κ to the equivalent ground truth label k. The precision and the recall for the class k are given by

$$P(k) = \frac{|\mathcal{G}_k \cap \hat{\mathcal{G}}_k|}{|\hat{\mathcal{G}}_k|} \tag{19}$$

$$R(k) = \frac{|\mathcal{G}_k \cap \hat{\mathcal{G}}_k|}{N_k} \tag{20}$$

where | | denotes set cardinality. Then, $\overline{F_1}$ is defined as

$$\overline{F_1} = \sum_{k=1}^{K} \frac{N_k}{N} F_1(k) = 2 \sum_{k=1}^{K} \left(\frac{N_k}{N}\right) \left(\frac{P(k) + R(k)}{P(k)\, R(k)}\right). \tag{21}$$

On average, the ENSC with the same set of parameters achieved an equally descriptive clustering with the SSC whose parameters were tuned for each image subset, taking into account $\overline{F_1}$. In particular, the ENSC performs better than the SSC in 11 image subsets among the 23. In another 2 image subsets, the performance of ENSC matches that of the SSC. It is worth mentioning that the ENSC outperforms the SSC for the 7 image subsets including more than 90 face images. This is attributed to the ability of the ENSC to handle better the collinear atoms in face clustering than the SSC.

4. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper, the performance of the ENSC was assessed in face clustering on two publicly available datasets, namely the Extended Yale B and the Hollywood Human Actions datasets. Its performance was compared to that of the SSC. Encouraging results have been demonstrated without paying any specific effort to either tune the algorithm parameters or estimate the number of clusters in the data matrix. The latter has the top priority in the future research. For example, Robust Principal Component Analysis could be exploited toward this direction. Although a limited set of figures of merit has been used to assess the clustering performance, in order to enable comparisons with existing related works, the set of figures of merit could be much broader, including pairwise recall, pairwise precision, and pairwise F measure, cluster purity, person purity, the Rand index, the mutual information, or the conditional entropy for over-segmentation and under-segmentation. The extended set of figures of merit makes sense, when the number of clusters is estimated. The concept of correntropy, and particularly the correntropy induced metric or the maximum correntropy criterion, as a generalized similarity measure between two random vectors, already exploited for low-rank representations in [26], will be employed to re-formulate the EN representation.

5. REFERENCES

[1] P. Antonopoulos, N. Nikolaidis, and I. Pitas, "Hierarchical face clustering using sift image features," in *Proc. IEEE Symp. Computational Intelligence for Image and Signal Processing*, Honolulu, Hawai, USA, 2007, pp. 325–329.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] C. Chrysouli, N. Vretos, and I. Pitas, "Face clustering in videos based on spectral clustering techniques," in *Proc. Asian Conference Pattern Recognition*, Beijing, China, 2011, pp. 130–134.

[4] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movie content analysis," *Image and Vision Computing*, vol. 29, no. 10, pp. 693–705, July 2011.

[5] G. Orfanidis, N. Nikolaidis, and I. Pitas, "Facial image clustering in single channel and stereo video content," in *Proc. 2013 Int. Workshop Biometrics and Forensics*, Lisbon, Portugal, 2013.

[6] C. Zhu, F. Wen, and J. Sun, "A rank-order distance based clustering algorithm for face tagging," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, 2011, pp. 481–488.

[7] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 52–68, March 2011.

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

[9] G. Chen and G. Lerman, "Spectral curvature clustering," *Int. J. Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, 2011, pp. 2790–2797.

[11] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theorem and applications," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 2765-2781, pp. 35, 2013, to appear.

[12] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Computer Vision*, Barcelona, Spain, 2011, pp. 1615–1622.

[13] V. Zografos, L. Ellis, and R. Mester, "Discriminative subspace clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 2107–2114.

[14] Y. Panagakis and C. Kotropoulos, "Elastic net subspace clustering applied to por/rock music structure analysis," *Pattern Recognition Letters*, vol. 38, no. 1, pp. 46–53, 2014.

[15] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. 2011 Neural Information Processing Systems Conf.*, Granada, Spain, 2011, pp. 612–620.

[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

Table 2. Clustering error (%) and $\overline{F_1}$ of different algorithms on the Hollywood Human Actions Dataset.

Image Subsets		SSC			ENSC	
Movie title	Number of face images	`alpha`	Clustering error	$\overline{F_1}$	Clustering error	$\overline{F_1}$
American Beauty	91	18	19.78	0.8352	4.39	0.9521
As Good As It Gets	109	5	30.28	0.7197	24.77	0.6986
Being John Malkovich	133	18	15.79	0.8662	15.78	0.8311
Big Fish	243	15	32.10	0.7765	22.64	0.8238
Great Lebowski	47	15	12.77	0.8814	27.66	0.6936
Bringing Out The Dead	60	20	8.47	0.9190	20.33	0.7834
Butterfly Effect	70	20	7.14	0.9313	10	0.8975
Crying Game	53	20	5.66	0.9422	28.30	0.7635
Dead Poets Society	222	5	34.23	0.6167	36.48	0.6168
Erin Brockovitch	236	20	52.12	0.6069	47.45	0.6719
Forest Gamp	184	20	44.02	0.6311	23.91	0.8260
Gandhi	235	5	26.38	0.7701	33.19	0.5711
The Graduate	125	5	48	0.6599	48	0.6599
I Am Sam	79	5	11.39	0.8622	11.39	0.8523
Indiana Jones And The Last Crusade	52	5	9.62	0.9253	23.07	0.8311
Kids	26	5	34.62	0.6941	42.30	0.6529
LOTR-Fellowship Of The Ring	63	18	22.22	0.8290	20.63	0.8055
Lost Highway	150	18	47.33	0.5728	37.33	0.6988
Mission To Mars	109	5	33.94	0.6924	38.53	0.6683
The Pianist	91	18	47.25	0.6052	38.46	0.6744
Pulp Fiction	67	5	14.93	0.8560	34.32	0.5715
The Godfather	30	5	30	0.7687	26.66	0.7302
Two Weeks Notice	940	18	42.87	0.5934	32.87	0.6788
Mean			27.43	0.7633	28.19	0.7371
Median			30	0.7701	27.26	0.6988

[17] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: Applications to image and text data,” in *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Fransisco, CA, 2001, pp. 245–250.

[18] K.-C. Lee, J. Ho, and D. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[19] I. Laptev, M. Marszkek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.

[20] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[21] A. Ng, Y. Weiss, and M. Jordan, “On spectral clustering: Analysis and an algorithm,” in *Proc. Neural Information Processing Systems*, 2001, pp. 849–856.

[22] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symposium Math. Stat. and Prob.*, Berkeley, CA, 1967, vol. I, pp. 281–297.

[23] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal R. Stat. Soc., Series B*, vol. 67, pp. 301–320, 2005.

[24] E. J. Candes, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” *Journal of ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[25] Y. E. Nesterov, “Gradient methods for minimizing composite objective function,” Tech. Rep., CORE report, 2007, `http://www.ecore.be/DPs/dp_1191313936.pdf`.

[26] Y. Zhang, Z. Sun, R. He, and T. Tan, “Low-rank representation via correntropy,” in *Proc. 2nd IAPR Asian Conf. Pattern Recognition*, Okinawa, Japan, 2013.