REINFORCING SELF-EXPRESSIVE REPRESENTATION WITH CONSTRAINT PROPAGATION FOR FACE CLUSTERING IN MOVIES

Krishna Somandepalli and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California
somandep@usc.edu, shri@ee.usc.edu

ABSTRACT

The ability to robustly cluster faces in movies is a necessary step in understanding media content representations of people along dimensions such as gender and age. Building upon the successes of sparse subspace clustering (SSC) in uncovering the underlying structure of the data, in this paper we propose an algorithm called Constraint Propagation Sparse Subspace Clustering (CP-SSC) for applications such as face clustering in videos where pairwise sample constraints (must-link and cannot-link sample pairs) are available in the processing pipeline since detected faces can be tracked locally in time. We learn the subspace structure while simultaneously incorporating the pairwise constraints to construct a similarity matrix needed for clustering. Our joint formulation uses low-rank matrix completion to propagate the initial pairwise constraints, that are used to reinforce the subspace representation during optimization. We evaluate CP-SSC for clustering faces in movies with pre-trained neural network embeddings as features. We first analyze CP-SSC with synthetic data and then show that it can be effectively used to cluster faces in movie videos. We evaluate our method for two movies annotated in-house and two benchmark movies released publicly. We also compare the performance of our algorithm with other clustering approaches that use pairwise constraint information.

***Index Terms*—** subspace clustering, self-expressive representation, constraint propagation, face clustering

1. INTRODUCTION

To address the emerging needs to analyze staggering volumes of rich and heterogeneous media data that are being generated, unsupervised learning approaches such as clustering offer an attractive solution. This paper is motivated by the application of clustering faces in videos, with a special focus on modeling representations along dimensions such gender and age in movies [1]. Such clustering typically follows the detecting and tracking faces that appear and disappear over time across shots and scenes in a movie.

A central step in clustering is the computation of a similarity matrix or a graph that encodes a measure of *distance* between the data points. But, this is often estimated from the input features. Predefined measures such as euclidean distance to obtain this similarity matrix may not always be optimal for the given data [2]. In cases where data points, (albeit high-dimensional) are drawn from a union of low-dimensional subspaces, the underlying graph structure can be inferred by expressing each data point as a linear combination of others that belong to the same subspace. This is commonly referred to as a *subspace-* or *self-expressive representation* [3] Subspace clustering concerns with learning this representation to partition data into clusters, where each cluster belongs to a unique subspace.

Thanks to Google, NSF and GDIGM.

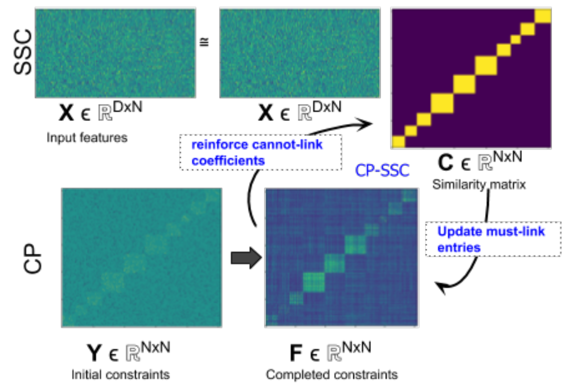

Fig. 1. Illustration of our proposed CP-SSC algorithm

Most subspace clustering methods are primarily based on two approaches: sparse subspace clustering (SSC) [3] and low rank representation (LRR) [4]. SSC learns the self expressive representation as a sparse coefficient matrix on the data points, while LRR exploits the low rank structure of the data matrix. A joint formulation involving SSC and LRR objectives has been proposed in [5] that demonstrates the benefit of combining these two methods for clustering tasks. Over the years, several variants of subspace clustering have been proposed. For a detailed review, see [6].

Subspace clustering has been successfully used for applications such as face clustering and motion segmentation, e.g., [3, 4]. Although face detection can be achieved with a high degree of precision, clustering faces for person recognition in videos is a challenging problem [7]. This is primarily due to the variation in quality (e.g., pose, illumination) of a person's face across the length of the movie [7]. This task is compounded with domain specific problems such as unknown number of clusters (e.g., movie characters) and outliers. A promising line of research is to leverage the temporal information in videos in the form of cannot-link and must-link associations between faces. In movies, these associations can be easily obtained. For example, two faces appearing in the same frame *cannot* belong to the same person, and all the faces in a track *must* belong to the same person. Such associations or *pairwise constraints* offer complementary information for clustering data.

A number of methods have been proposed for face clustering in videos that incorporate pairwise constraints. For example, end-to-end methods for face tracking and clustering have been proposed in [8] and [9]. Other metric-learning based clustering approaches for movies have also been proposed in [10] and [11]. Although the performance of our approach on the benchmark data is comparable with these methods, the focus our paper is to examine subspace clustering in the presence of pairwise constraints. Thus, we primarily focus only on subspace clustering methods that use pairwise constraints.

One such method is weighted block-sparse low rank representation (WBSLRR). It extends the LRR formulation by using must-link constraints of all the faces in a track. WBSLRR has been shown to perform better than traditional clustering methods such as K-Means as well as several variants of SSC and LRR for the task of face clustering in videos [12]. A theoretical extension for SSC, similar to WBSLRR has been proposed in [13].

A key insight with respect to pairwise constraints is that the constraint matrix is *low-rank* when all its entries are known. This is because the number of clusters are fewer than the number of inputs. As shown in [14, 15], matrix completion approaches can be used to recover unknown constraints in the context of clustering. This is referred to as *constraint propagation* (CP).

The aforementioned works have either examined subspace clustering with constraints or CP individually. Complementary to these works, we propose a joint formulation, that uses CP to update or *reinforce* the self-expressive representation during optimization. The schematic of our proposed algorithm is shown in **Fig. 1**. The objective here is to use the updates must-link constraints from matrix completion to guide SSC, and to use SSC coefficients to guide matrix completion for CP. We show that our method can be effectively used for face clustering in movie videos and compare our results with two previous methods. The datasets we have created and our code are publicly available[1].

2. METHODS

2.1. Problem Setup

Subspace clustering can be formally described as follows: Let $\mathbf{X} \in \mathbb{R}^{D\times N}$ be the matrix of N data points with columns $\mathbf{x}_i \in \mathbb{R}^D$ as features of ambient dimension D. Each $\mathbf{x}_i$ belongs to a union of L linear (possibly affine) subspaces $\{S_l\}_{l=1}^{L}$ each of dimension d_l with N_l points from $\mathbf{X}$. Then $\mathbf{X}$ can be expressed as

$$\mathbf{X} \triangleq [\mathbf{x}_1, ..., \mathbf{x}_N] = [\mathbf{X}^{(1)}, ..., \mathbf{X}^{(L)}]\Gamma \tag{1}$$

where $\mathbf{X}^{(l)} \in \mathbb{R}^{D\times N_l}$ and Γ is an unknown permutation matrix. The objective of subspace segmentation is to find the set $\{\mathbf{S}_l | x_i \in \mathbf{S}_l, l = 1, ..., L\}$. The subspace dimensions D and N_l, and number of subspaces L are unknown.

Definition 1: Self-expressive representation. A coefficient matrix which allows one data point in $\mathbf{X}$ to be represented as a combination of other data points is referred to as a self expressive representation. The matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$ is said to be a such a representation [3] of $\mathbf{X}$ if every column $\mathbf{x}_i$ can be expressed as a sum of all the other points. Formally,

$$\mathbf{X} \quad = \quad \mathbf{XC} \quad s.t. \quad \text{diag}(\mathbf{C}) = 0 \tag{2}$$

$\text{diag}(\mathbf{C}) = 0$ is used to prevent a trivial solution of C.

Definition 2: SSC. Partitioning $\mathbf{X}$ into L groups using a sparse solution, $\mathbf{C}$ for (2) is referred to as SSC. Typically the similarity matrix for clustering is constructed as $\mathbf{A} = |\mathbf{C}| + |\mathbf{C}|^\top$.

Assuming the subspace dimension, $d_l < D \quad \forall l \in [1, L]$ and sufficient sampling density of points, ρ_l in each subspace, where $\rho_l = N_l/d_l, \rho_l > 1$ a sparsest solution for $\mathbf{C}$ in (2) can be obtained by minimizing the matrix l_0-norm as

$$\min_{\mathbf{C}} \|\mathbf{C}\|_0 \quad s.t. \quad \mathbf{X} = \mathbf{XC}, \text{diag}(\mathbf{C}) = 0 \tag{3}$$

This problem is known to be NP-hard. One practical solution is via convex relaxation that minimizes the l_1-norm instead. Theoretical analyses in [3, 16] have shown that the l_1-norm minimization problem achieves an optimal solution for (2) if the angles between the subspaces are sufficiently large and the subspaces are (preferably) disjoint[2]. SSC has also been shown to be robust to some overlap between subspaces [16].

Definition 3: Pairwise constraint matrix. The matrix $\mathbf{F} \in \{-1, +1\}^{N\times N}$ is a pairwise constraint matrix if $\mathbf{F}_{ij} = +1$ indicates that the pair of data points (x_i, x_j), belong to the same subspace, and $\mathbf{F}_{ij} = -1$ indicates that the pair of data points do not belong to the same subspace. This matrix is often incomplete for real data. For brevity, we refer to $\mathbf{F}$ as the constraint matrix, not be confused with optimization constraints.

Let $\Omega_{\mathcal{M}}$ and $\Omega_{\mathcal{C}}$ be known sets of pairs of data points in $\mathbf{X}$ that belong to the same subspace (must-link constraints, $\mathcal{M}$) and those that do not belong to the same subspace (cannot-link constraints, $\mathcal{C}$) respectively. The observed incomplete matrix $\mathbf{Y}$ is given by,

$$\mathbf{Y}_{ij} = \begin{cases} +1 & \text{if } (x_i, x_j) \in \Omega_{\mathcal{M}} \\ -1 & \text{if } (x_i, x_j) \in \Omega_{\mathcal{C}} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The objective is to complete $\mathbf{Y}$ to recover $\mathbf{F}$. The number of known entries in the matrix $\mathbf{Y}$ is denoted by ν – a parameter that controls the degrees of freedom (DOF) of $\mathbf{F}$ such that,

$$\nu = \frac{|\Omega_{\mathcal{C}} \cup \Omega_{\mathcal{M}}|}{\text{DOF}(\mathbf{F})} = \frac{|\Omega|}{r(2N - r)} \tag{5}$$

By definition, $\mathbf{F}$ is low-rank with $\text{rank}(\mathbf{F}) \approx r < D$, where $r = L$ in the absence of outliers (else $r \approx L+1$). The column entries of the completed constraint matrix indicate the ground-truth membership of $\mathbf{x}_i$ to the underlying subspaces.

2.2. Constraint Propagation SSC (CP-SSC)

Our joint formulation of CP-SSC involves recovering the entries of the incomplete constraint matrix, $\mathbf{Y}$ to *reinforce* the self-expressive representation $\mathbf{C}$ with the updated cannot-link constraints. For this objective, we solve the following convex optimization problem. If the data matrix $\mathbf{X}$ is noise-free:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{F}} \quad & \lambda_1\|\mathbf{C}\|_1 + \lambda_2\|[\mathbf{XC}; \mathbf{F}]\|_* \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{XC}, \quad \text{diag}(\mathbf{C}) = 0, \quad \Psi_{\mathbf{F}}(\mathbf{C}) = 0 \\ & \mathbf{F}_{ij} = \mathbf{Y}_{ij} \quad \forall(\mathbf{x}_i, \mathbf{x}_j) \in \Omega \end{aligned} \tag{6}$$

For a noisy $\mathbf{X}$ such that $\mathbf{X} = \mathbf{XC} + \mathbf{Z}$:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{F}} \quad & \frac{\alpha}{2}\|\mathbf{X} - \mathbf{XC}\|_{\mathbf{F}}^2 + \lambda_1\|\mathbf{C}\|_1 + \lambda_2\|[\mathbf{XC}; \mathbf{F}]\|_* \\ \text{s.t.} \quad & \text{diag}(\mathbf{C}) = 0, \quad \Psi_{\mathbf{F}}(\mathbf{C}) = 0 \\ & \mathbf{F}_{ij} = \mathbf{Y}_{ij} \quad \forall(\mathbf{x}_i, \mathbf{x}_j) \in \Omega \end{aligned} \tag{7}$$

where $\|\cdot\|_*$ is the matrix nuclear norm. We define a linear operation, $\Psi_{\mathbf{F}}(\mathbf{C})$ to set the entries of $\mathbf{C}$ corresponding to the cannot-link entries in $\mathbf{F}$ to zero i.e., $\mathbf{C}_{ij} = 0 \quad \forall \mathbf{F}_{ij} \leq -1$.

Following the analysis in [3, 5, 14], the weights, $\alpha_1 = \min_i \max_{j\neq i} \|\mathbf{x}_i^\top \mathbf{x}_j\|_1$ and $\alpha_2 = \|(sgn([\mathbf{X}; \mathbf{Y}]))\|$ where $\|\cdot\|$ is the spectral norm. A single parameter λ is tuned to control the terms that minimize the l_1-norm and nuclear norm in (7); $\lambda_1 = \frac{\lambda}{\alpha_1(1+\lambda)}, \lambda_2 = \frac{1}{\alpha_2(1+\lambda)}$. We study the effect of λ with experiments

[1] github.com/usc-sail/mica-face-clustering

[2] A set of subspaces is said to be disjoint if every pair of subspaces intersect only at origin

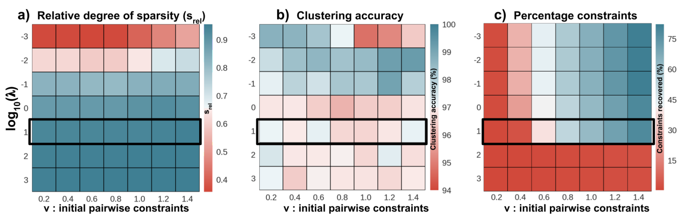

Fig. 2. Performance evaluation on synthetic data as a function of sparsity-matrix completion tradeoff λ and initial constraints ν

on synthetic data as described in Section 3.2. We perform low-rank matrix completion on the stacked thin matrix, $[\mathbf{XC}; \mathbf{F}]$ instead of the $\mathbf{F}$ matrix only [14].

Because the problem in (7) is generally solved at large-scale, off the shelf solvers are often slow, hence we derive a numerical algorithm using alternating direction methods of multipliers (ADMM). To rewrite the constraints, we introduce auxiliary variables, $\mathbf{J}, \mathbf{B}, \mathbf{M}, \mathbf{Q}$ such that

$$\mathbf{J} = \mathbf{C} - \text{diag}(\mathbf{C}), \qquad \mathbf{J} \in \mathbb{R}^{N \times N} \tag{8}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{XJ} \\ \mathbf{F} \end{bmatrix}; \mathbf{M} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}; \mathbf{Q} = \begin{bmatrix} \mathbf{XJ} \\ \mathbf{Y} \end{bmatrix} \text{ s.t. } \mathbf{B} + \mathbf{M} = \mathbf{Q} \tag{9}$$

Here, $\mathbf{R}$ encodes the unknown entries of $\mathbf{Y}$. For simplicity, we set the upper part of the corresponding stacked matrix, $\mathbf{M}$ to $\mathbf{0}$. The nuclear norm term in (7) is now rewritten as $\mathbf{B}$. The input, known matrix is $\mathbf{Q}$. In order to ensure that known entries of $\mathbf{Y}$ are not altered through iterations, we define a linear operation, Φ that resets all the known entries of $\mathbf{Y}$ in $\mathbf{R}$ i.e., $\Phi_\Omega(R) = 0$. Now, the modified problem to be solved is:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{C},\mathbf{J}} \quad & \frac{\alpha}{2}\|\mathbf{X} - \mathbf{XJ}\|_{\mathbf{F}}^2 + \lambda_1\|\mathbf{C}\|_1 + \lambda_2\|\mathbf{B}\|_* \\ \text{s.t.} \quad & \mathbf{J} = \mathbf{C} - \text{diag}(\mathbf{C}), \quad \Psi_{\mathbf{F}}(\mathbf{J}) = 0 \\ & \mathbf{B} + \mathbf{M} = \mathbf{Q}, \quad \Phi_\Omega(\mathbf{R}) = 0 \end{aligned} \tag{10}$$

The derivation of ADMM for (10) is straight-forward by introducing primal variables [17]. l_1-norm and nuclear-norm minimization is achieved by shrinkage-thresholding singular-value thresholding operations. We omit the details of the derivation due to lack of space.

3. EXPERIMENTS

We evaluated the performance of CP-SSC along three dimensions: 1) the degree to which a sparse solution is achieved for SSC, 2) clustering accuracy because the application of our method is for clustering data, and 3) percentage of pairwise-constraints recovered. We prepared synthetic data by varying different parameters of the subspace assumption and pairwise constraints for this purpose. Finally, we performed clustering on face tracks from movies and compare with others that use pairwise constraints for subspace clustering.

3.1. Performance Measures

We define a relative degree of sparsity,s_{rel} similar to [5] as,

$$s_{rel} = \exp\left(- \frac{\sum_{(i,j)\notin\mathcal{M}} |\mathbf{C}|_{i,j}}{\sum_{(i,j)\in\mathcal{M}} |\mathbf{C}|_{i,j}} \right) \tag{11}$$

where $\mathcal{M}$ is the set of indices corresponding to all the must-link constraints (see **Definition 3**). A perfect sparse solution, $\mathbf{C}$ would have a $s_{rel} = 1$ and values closer to 0 would indicate a non-sparse and noisy result.

A solution with $s_{rel} = 1$ could be extremely sparse with respect to intra-cluster connections. Thus, a high s_{rel} does not necessarily mean that we have obtained a similarity matrix optimal for clustering. Hence, we also measure clustering accuracy as a percentage of the total number of points that were correctly classified. In order to evaluate the performance of matrix completion, we report the number of matrix entries correctly recovered as a fraction of all the matrix entries ($N^2 - N$).

3.2. Experiments with synthetic Data

The goal of these simulations was to study the feasibility of CP-SSC for different subspace settings used to generate $\mathbf{X}$ and the number of initial constraints in $\mathbf{Y}$. We consider a varying number clusters, L which is controlled by a minimum subspace dimension, d_{min}. Thus, the number of clusters $L = 2D/d_{min}$. We set the ambient dimension $D = 128$ and $d_{min} = \{6, 8, 10, 15, ..., 40, 45\}$. The basis dimensions, d_i for the L subspaces are also varied by randomly picking an integer in the range $[d_{min}, D/2]$, in this case: $[d_{min}, 64]$. The L bases are generated such that $\{\mathbf{U}_i \in \mathbb{R}^{128\times d_i}\}_{i=1}^L$; rank$([\mathbf{U}_1, ..., \mathbf{U}_L]) = D$. The sampling density ρ_l (see Definition 2) is picked randomly from $\{2, 3, 4\}$. Together, ρ_l, L and d_i determine N, the total number of data points in $\mathbf{X}$. A uniformly distributed noise is then added to the data matrix $\mathbf{X}$, where each column is normalized to have a unit l_2-norm. These settings were chosen to obtain a feasible SSC solution [18], at the same time make the problem harder by ensuring that the data points in a given subspace can be reconstructed as a linear combination of data points from other subspaces as well. The incomplete constraint matrix, $\mathbf{Y}$ was simulated by picking $\nu \cdot \text{DOF}(\mathbf{F})$ entries uniformly distributed across $\mathbf{Y}$.

In order to assess the performance of our algorithm under different settings, we varied two parameters in creating the synthetic data: 1) $\nu = \{0.2, 0.4, ..., 1.2, 1.4\}$ per eqn. (5), and 2) $d_{min} = \{4, 6, 8, 10, 15, 20\}$ which controls the subspace dimensions. We obtained the performance measures by tuning over $\lambda = \{1e-3, 1e-2, 0.1, 1, 10, 100, 1000\}$. This parameter controls the weights on the terms minimizing l_1-norm and nuclear norm in (7). All performance measures are averaged across the different number of clusters, which is controlled by d_{min}.

3.3. Face Clustering Experiments

We evaluated our algorithm on four datasets to assess its performance for face clustering in movies. We used two popular benchmark datasets, the movie *Notting-Hill* (**NH**) [12], and the *BF-05-02* (**BFF**), an episode from the TV series, *Buffy the Vampire Slayer* [7]. In addition, we created labeled face data for two 2014 movies, *Dumb and Dumber To* (**DD2**) and *Maleficent* (**MT**).

Face detection and tracking was performed with Google's API for face recognition [3] to obtain a homogeneous, local sequence of faces, or *face-tracks*. Pose and facial landmarks were available with face tracking. All faces were aligned in-plane to an average face using the landmarks for eyes and nose. We then divided the face-tracks into smaller segments or *tracklets* based on the variation in pose. Tracklets ensure that the similarity matrix is robust to pose variation within a track, as well as naturally bootstrap must-link constraints. Cannot-link constraints were obtained by querying all pairs of tracklets that consist of at least one common frame with a face from two different tracks. *OpenFace* [19] features of dimension $D = 128$

[3]Neven Vision fR™ API

Table 1. Description of the in-house and benchmark dataset

Dataset	Tracks	Tracklets	ν	Faces	Clusters
NH	225	320	0.51	13657	7
BFF	229	625	0.73	16963	6
DD2	1120	2350	0.65	108962	7
MT	524	1350	0.79	42265	10

Table 2. CP-SSC performance evaluation (clustering accuracy%)

Method\Movie	NH	BFF	DD2	MT
uLDML [10]	43.83	49.29	-	-
cHRMF [11]	47.95	61.87	-	-
SSC+C	49.71	52.17	51.32	47.10
WBSLRR [12]	58.32	62.10	62.31	60.57
CP-SSC [ours]	54.27	65.24	71.31	74.63

were obtained for all the faces, and averaged across each tracklet, and normalized to have a unit l_2-norm. We chose *OpenFace* features because they have been shown to offer robust face representation under varying conditions [19]. The number of tracklets, constraints and cluster information for the four datasets are shown in **Table 1**.

We compared our method with four baseline algorithms: 1) **SSC+C**: This offers a solution of SSC and CP performed separately rather than jointly. Spectral clustering was performed on the similarity matrix from SSC [3] after *kernelizing* with the recovered constraints as described in [14]. 2) **WBSLRR**: This approach enforces a block sparse structure of the face tracks to use pairwise constraints in a LRR formulation. We chose this as a baseline because it was specifically designed for face clustering in videos. Additionally, as described in Section 1, WBSLRR gives one of the best results among the subspace clustering methods that use constraint information for movie data. We implemented this using the code released by the authors [12]. For the benchmark datasets, we also compare our method with two unsupervised learning baselines: 3) **uLDML:** unsupervised logistic discriminative metric learning [10], and 4) a recently proposed constrained clustering approach, **cHMRF:** coupled hidden Markov random field [20]. We chose these methods as they are unsupervised approaches that exploit some form of pairwise-constraints.

4. RESULTS AND DISCUSSION

Our objective of using synthetic data was to empirically understand the behavior of our algorithm subject to different parameters of the subspace assumptions. As such, we did not compare with the baseline methods for the synthetic data. A detailed theoretical analysis would be a subject of our future work. The parameter λ that we use to set the weights λ_1, λ_2 in (7) controls the shrinkage value for l_1-norm and nuclear norm minimization respectively. As λ decreases, λ_1 decreases, thus stepping slowly for shrinkage in the l_1-norm objective. Simultaneously, the singular value thresholding for nuclear-norm minimization is more aggressive which results in a faster minimization of the rank for the constraint matrix. λ captures a trade-off between sparsity and matrix completion in our joint formulation. The results for synthetic data as a function of λ and ν are shown in **Fig. 2**. A higher λ produces a more sparse solution with s_{rel} (11) closer to 1 (**Fig. 2a**). However, for a sufficient ν, on increasing λ, fewer entries in the constraint matrix were recovered (**Fig. 2c**). Because of the disjoint subspaces assumption used to create synthetic data, we obtain clustering accuracy of over 94% in all cases (**Fig. 2b**). Although, this assumption is not entirely satisfied for real data, the SSC has been shown to be robust [18] to overlap between subspaces. Based on these simulations, we set $\lambda = 10$ (See bold box in **Fig. 2**) for face clustering experiments.

As shown in **Fig. 2c**, not surprisingly, about 90% of the constraints can be successfully recovered if the number of entries is at least as big as the DOF of the matrix. This is consistent with matrix completion theory [21] and CP experiments in [14]. Our simulations also showed that over 70% of the entries can be successfully recovered if at least $0.7 \cdot \text{DOF}(\mathbf{Y})$ entries are given (See $\nu > 0.7$ in **Fig. 2c**). The underlying assumption that the given entries are uniformly distributed is satisfied in our simulations. However, the constraints available from real data may not be uniformly distributed. Hence, analysis of measures of coherence [21] in the case of general sampling distribution [22] is necessary to understand the theoretical limitations of our approach. This will be a focus of our future work.

The results for face clustering experiments comparing our method with the baseline experiments are shown in **Table 2**. CP-SSC outperformed SSC+C for all the datasets. This indicates that the joint formulation is beneficial to performing the two steps consecutively for this task. This was consistent with our experiments on synthetic data as well. Additionally, methods that incorporate pairwise constraints outperformed unsupervised learning methods uLDML [10] and cHRMF [11] that were developed for face clustering in TV videos. As explained in Section 3.3. We used *OpenFace* features averaged across all faces in a tracklet and the clustering was performed at the tracklet level. For all datasets except **NH**, CP-SSC performed better than WBSLRR. This suggests that having fewer number of samples as compared to the ambient dimension (320 data points vs. 128) is detrimental to our system performance. This lowers the sampling density of some of the subspaces which makes clustering challenging. This observation is consistent with the *sampling condition theorem* from [16] which establishes a lower bound on the sampling density failing which the SSC solution is not ideal for clustering. Furthermore, we had fewer pairwise constraints for **NH** compared to other videos.

Across all these experiments, complementary measures that quantify clustering performance: completeness, homogeneity, adjusted mutual information and adjust rand index [23] showed a consistent trend similar to that of the clustering accuracy (data not shown). For all experiments, we assumed the true number of clusters to be known and no outlier face-tracks.

5. CONCLUSION

In this paper, we proposed the CP-SSC algorithm to leverage pairwise constraints in subspace clustering for applications such as face clustering in videos. Our simulations suggest that CP-SSC is feasible for a wide range of subspace dimensions and the number of initial pairwise constraints. We showed that our method can be used for face clustering in movie videos using benchmark data for movie faces and data labeled in-house. Given sufficient sampling density, CP-SSC performed better than a state-of-the-art subspace clustering method that was designed specifically for face clustering in videos in the presence of pairwise constraints.

A promising direction for this work is to formalize a unified framework for the concepts of subspace clustering and matrix completion. This can be used for unsupervised learning when constraint information based on domain-knowledge is available.

6. REFERENCES

[1] Tanaya Guha, Che-Wei Huang, Naveen Kumar, Yan Zhu, and Shrikanth S Narayanan, "Gender representation in cinematic content: A multimodal approach," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 31–34.

[2] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PLOS ONE*, vol. 10, no. 12, pp. 1–20, 12 2015.

[3] Ehsan Elhamifar and Rene Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[4] Guangcan Liu, Zhouchen Lin, and Yong Yu, "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, USA, 2010, ICML'10, pp. 663–670.

[5] Yu-Xiang Wang, Huan Xu, and Chenlei Leng, "Provable subspace clustering: When lrr meets ssc," in *Advances in Neural Information Processing Systems 26*, pp. 64–72. 2013.

[6] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, March 2011.

[7] Mark Everingham, Josef Sivic, and Andrew Zisserman, "Hello! my name is... buffy–automatic naming of characters in tv video," 2006.

[8] SouYoung Jin, Hang Su, Chris Stauffer, and Erik Learned-Miller, "End-to-end face detection and cast grouping in movies using erdos-rényi clustering," *arXiv preprint arXiv:1709.02458*, 2017.

[9] B. Wu, S. Lyu, B. G. Hu, and Q. Ji, "Simultaneous clustering and tracklet linking for multi-face tracking in videos," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 2856–2863.

[10] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Unsupervised metric learning for face identification in tv video," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1559–1566.

[11] B. Wu, Y. Zhang, B. G. Hu, and Q. Ji, "Constrained clustering and its application to face clustering in videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3507–3514.

[12] Shijie Xiao, Mingkui Tan, and Dong Xu, *Weighted Block-Sparse Low Rank Representation for Face Clustering in Videos*, pp. 123–138, Springer International Publishing, Cham, 2014.

[13] D. S. Pham, S. Budhaditya, D. Phung, and S. Venkatesh, "Improved subspace clustering via exploitation of spatial constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 550–557.

[14] Zhenyong Fu, "Pairwise constraint propagation via low-rank matrix recovery," *Computational Visual Media*, vol. 1, no. 3, pp. 211–220, Sep 2015.

[15] Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," 2015, AAAI'15, pp. 2778–2784.

[16] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès, "Robust subspace clustering," *Ann. Statist.*, vol. 42, no. 2, pp. 669–699, 04 2014.

[17] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, pp. 1–122, Jan. 2011.

[18] Mahdi Soltanolkotabi and Emmanuel J. Candès, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, 08 2012.

[19] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Tech. Rep., CMU-CS-16-118, CMU School of Computer Science, 2016.

[20] Baoyuan Wu, Bao-Gang Hu, and Qiang Ji, "A coupled hidden markov random field model for simultaneous face clustering and tracking in videos," *Pattern Recognition*, vol. 64, pp. 361–373, 2017.

[21] Emmanuel J. Candès and Benjamin Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717, Apr 2009.

[22] Olga Klopp et al., "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[23] Nguyen Xuan Vinh, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 1073–1080, ACM.