CLUSTERING-AWARE STRUCTURE-CONSTRAINED LOW-RANK REPRESENTATION MODEL FOR LEARNING HUMAN ACTION ATTRIBUTES

Tong Wu[⋆] *Prudhvi Gurram*[†‡] *Raghuveer M. Rao*[‡] *Waheed U. Bajwa*[⋆]

[⋆] Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854
[†] Booz Allen Hamilton, McLean, VA 22102
[‡] U.S. Army Research Laboratory, Adelphi, MD 20783

ABSTRACT

This paper addresses the problem of learning meaningful human action attributes from high-dimensional video sequences based on union-of-subspaces (UoS) model. The model hypothesizes that each action attribute is represented by a subspace. It puts forth an extension of existing low-rank representation (LRR), termed the clustering-aware structure-constrained low-rank representation (CS-LRR) model, for unsupervised learning of human action attributes. The proposed CS-LRR model overcomes the shortcomings of existing techniques by its ability to handle disjoint subspaces, and by performing optimal spectral clustering of the subspaces. An efficient linear alternating direction method (LADM) is developed to solve the CS-LRR optimization problem. A human action or activity is represented as a sequence of transitions from one action attribute to another and can be uniquely represented by a subspace transition vector. These subspace transition vectors are used for human action recognition. The effectiveness of the proposed model is demonstrated through experiments on two real-world datasets for action recognition.

1. INTRODUCTION

A complex human activity or a high-level event can be considered to be a hierarchical model [1], consisting of a sequence of simpler human actions. Each action can further be divided into a sequence of movements of the human body, which we call action attributes [2]. In the past, a significant fraction of the video analytics literature has been devoted to the study of learning attributes in human actions [3, 4]. In this paper, we propose to represent human action attributes based on the *union of subspaces* (UoS) model [5, 6], which is motivated by the fact that high-dimensional video data usually lie in a union of low-dimensional subspaces instead of being uniformly distributed in the high-dimensional ambient space [5]. The hypothesis of this model is that each action attribute is represented by a subspace. A human action or activity can then be represented as a sequence of transitions from one attribute to another and, hence, can be represented by a subspace transition vector. Even though multiple actions can share action attributes, each action or activity can be uniquely represented by its subspace transition vector when the attributes are learned using the UoS model. Thus, these transition vectors can be used for human action recognition.

High-dimensional data can be clustered into subspaces by applying spectral clustering on a graph associated with the data. Recently, the authors in [5] developed sparse subspace clustering (SSC) technique, where spectral clustering is applied on an "ℓ_1 graph." This has been extended into a hierarchical structure to learn subspaces at multiple resolutions in [7]. To capture the global structure of the data, low-rank representation (LRR) models without and with sparsity constraints have been proposed in [6] and [8], respectively. It has been proved that LRR can achieve perfect subspace clustering results under the condition that the subspaces underlying the signals are independent [6, 9]. However, this condition is hard to satisfy in many real situations. To handle the case of disjoint subspaces, Tang *et al.* extended LRR by imposing restrictions on the structure of the solution. This method is called structure-constrained LRR (SC-LRR) in [10]. Existing LRR based subspace clustering techniques use spectral clustering as a post-processing step on the graph, which can lead to sub-optimal results [11]. To overcome this issue, we propose a clustering-aware structure-constrained LRR (CS-LRR) model to obtain an optimal clustering of human action attributes from a large collection of video sequences in an unsupervised manner by introducing spectral clustering into the optimization problem. In order to demonstrate effectiveness of the CS-LRR model and the proposed learning algorithm, we carry out numerical experiments using real video datasets. Results of these experiments demonstrate the effectiveness of our approach and its superiority over the state-of-the-art subspace clustering methods for human action recognition.

The following notation will be used throughout the paper. We use non-bold letters to represent scalars, bold lowercase letters to denote vectors, and bold uppercase letters to denote matrices. The i-th element of a vector $\mathbf{a}$ is denoted by $a_{(i)}$

This work was supported in part by the NSF under grants CCF-1218942 and CCF-1453073, by the Army Research Office under grant W911NF-14-1-0295, and by an Army Research Lab Robotics CTA subaward.

978-1-5090-0746-2/16/$31.00

and the (i, j)-th element of a matrix $\mathbf{A}$ is denoted by a_{ij}. The i-th row and the j-th column of a matrix $\mathbf{A}$ are denoted by $\mathbf{a}^i$ and $\mathbf{a}_j$, respectively. The zero matrix is denoted by $\mathbf{0}$ and the identity matrix is denoted by $\mathbf{I}$. The only vector norm used in this paper is the ℓ_2 norm, which is represented by $\|\cdot\|_2$. We use a variety of different norms on matrices. The matrix ℓ_1 and $\ell_{2,1}$ norms are denoted by $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$ and $\|\mathbf{A}\|_{2,1} = \sum_j \|\mathbf{a}_j\|_2$, respectively. The ℓ_∞ norm is defined as $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$. The spectral norm of a matrix $\mathbf{A}$, i.e., the largest singular value of $\mathbf{A}$, is denoted by $\|\mathbf{A}\|$. The Frobenius norm and the nuclear norm (the sum of singular values) of a matrix $\mathbf{A}$ are denoted by $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_*$, respectively. Finally, the Euclidean inner product between two matrices is $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$, where $(\cdot)^T$ and $\mathrm{tr}(\cdot)$ denote transpose and trace operations, respectively.

2. PROBLEM FORMULATION

In this section, we give a brief review of low-rank representation (LRR) and structure-constrained LRR (SC-LRR), and formulate the problem studied in this paper. Consider a collection of N data points $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where each $\mathbf{x}_i \in \mathbb{R}^m$ is drawn from a union of L subspaces $\{\mathcal{S}_\ell\}_{\ell=1}^L$. The task of subspace clustering is to segment the data points according to their respective subspaces. Low-rank representation (LRR) is a recently proposed subspace clustering method [6], which aims to find the lowest-rank representation for the data using a dictionary to reveal the intrinsic geometric structure of the data. Mathematically, LRR can be formulated as the following optimization problem [6]:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_\iota \quad \text{s.t.} \quad \mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E}, \tag{1}$$

where $\mathbf{A}$ is a dictionary that linearly spans the data space, $\mathbf{Z}$ is the low-rank representation of the data over $\mathbf{A}$, and $\mathbf{E}$ is a matrix representing the reconstruction errors in the data. The nuclear norm is the convex relaxation of the rank function, $\|\cdot\|_\iota$ indicates a certain regularization strategy involving $\mathbf{E}$, and λ is a positive parameter which sets the tradeoff between low rankness of the matrix $\mathbf{Z}$ and the reconstruction error. To cluster the data points in $\mathbf{X}$, the data matrix $\mathbf{X}$ is used as the dictionary in (1) [9].

After obtaining the representation coefficient matrix $\mathbf{Z}$, one can define an affinity matrix $\mathbf{W}$ as $\mathbf{W} = \frac{|\mathbf{Z}|+|\mathbf{Z}^T|}{2}$, where $|\cdot|$ denotes the element-wise absolute value operation. Then we can apply spectral clustering [12] on $\mathbf{W}$ by solving the following problem:

$$\min_{\mathbf{F}} \mathrm{tr}(\mathbf{F}^T(\mathbf{D} - \mathbf{W})\mathbf{F}) \quad \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}, \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{N\times L}$ is the cluster indicator matrix and $\mathbf{D}$ is a diagonal matrix with diagonal elements $d_{ii} = \sum_j w_{ij}$. The solution $\mathbf{F}$ consists of the eigenvectors corresponding to the smallest L eigenvalues of the Laplacian matrix $\mathbf{M} = \mathbf{D} - \mathbf{W}$.

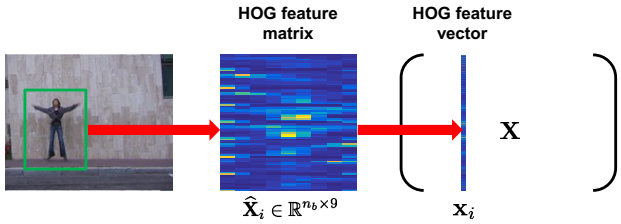

Fig. 1. Illustration of our approach to create the matrix $\mathbf{X}$.

To improve upon LRR for disjoint subspace clustering, Tang *et al.* [10] proposed SC-LRR model, which can be written as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{B} \odot \mathbf{Z}\|_1 + \lambda\|\mathbf{E}\|_{2,1} \quad \text{s.t.} \quad \mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E},$$

where α and λ are penalty parameters, $\mathbf{B}$ is a pre-defined weight matrix, and $\odot$ denotes the Hadamard product. The $\ell_{2,1}$ norm is used to model sample-specific corruptions and outliers. It has been shown in [10] that by designing an appropriate matrix $\mathbf{B}$, the optimal solution of $\mathbf{Z}$ is block-diagonal when the data are drawn from disjoint subspaces in the noiseless case. In [10], the (i, j)-th entry of $\mathbf{B}$ is defined as $b_{ij} = 1 - \exp(-\frac{1-|\mathbf{x}_i^T\mathbf{x}_j|}{\sigma})$, where σ is the mean of all $1 - |\mathbf{x}_i^T\mathbf{x}_j|$'s. The matrix $\mathbf{B}$ penalizes affinities between data samples from different clusters, while rewarding affinities between data samples from the same cluster. As discussed in [11], almost all the existing subspace clustering methods use spectral clustering as the post-processing step, whose solution may be sub-optimal. Instead, here, we propose a clustering-aware structure-constrained LRR (CS-LRR) model to address this issue by solving the following problem:

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{F},\mathbf{E}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{B} \odot \mathbf{Z}\|_1 + \beta\mathrm{tr}\Big(\mathbf{F}^T(\mathbf{D} - \frac{|\mathbf{Z}| + |\mathbf{Z}^T|}{2})\mathbf{F}\Big) \\ + \lambda\|\mathbf{E}\|_\iota \\ \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E},\ \mathbf{F}^T\mathbf{F} = \mathbf{I},\end{aligned} \tag{3}$$

where α, β and λ are penalty parameters. The number of clusters, L, is assumed to be known a priori in this paper, whereas the estimation of L will be considered as future work.

To perform action attribute clustering, we consider the action interest region in each frame of an action sequence determined by a bounding box and divide the action interest region into a grid of size $n_\sigma \times n_\sigma$. Then the HOG (histograms of oriented gradients) feature is extracted for each block [13], and the orientations are quantized into 9 bins. Therefore, the HOG feature of every frame can be stored into a 2-dimensional matrix $\widehat{\mathbf{X}}_i \in \mathbb{R}^{n_b\times 9}$, where n_b denotes the number of blocks in the action interest region, and the HOG feature vector of each block corresponds to a row in $\widehat{\mathbf{X}}_i$. We vectorize the HOG features and normalize each vector to unit ℓ_2 norm, forming individual data samples in the data matrix $\mathbf{X} \in \mathbb{R}^{m\times N}$, where

$m = n_b \times 9$, as shown in Figure 1. We set the error term $\|\mathbf{E}\|_\iota \equiv \sum_{i=1}^{N} \|\widehat{\mathbf{E}}_i\|_{2,1}$, to model HOG feature orientation-specific corruptions, where $\widehat{\mathbf{E}}_i$ denotes the reconstruction error with respect to $\widehat{\mathbf{X}}_i$.

3. OPTIMIZATION APPROACH

In this section, we propose an algorithm to solve (3) by using the linearized alternating direction method (LADM) [14]. By introducing an auxiliary variable $\mathbf{Q}$ to make the objective function of (3) separable, the problem (3) can be reformulated as

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{Q},\mathbf{F},\mathbf{E}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{B}\odot\mathbf{Q}\|_1 + \beta\mathrm{tr}\Big(\mathbf{F}^T(\mathbf{D} - \frac{|\mathbf{Q}|+|\mathbf{Q}^T|}{2})\mathbf{F}\Big) \\ + \lambda\sum_{i=1}^{N}\|\widehat{\mathbf{E}}_i\|_{2,1} \\ \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E},\ \mathbf{F}^T\mathbf{F} = \mathbf{I},\ \mathbf{Z} = \mathbf{Q}. \end{aligned} \tag{4}$$

The augmented Lagrangian function of (4) is

$$\begin{aligned}&\mathcal{L}(\mathbf{Z},\mathbf{Q},\mathbf{F},\mathbf{E},\mathbf{\Gamma}_1,\mathbf{\Gamma}_2) \\ &= \|\mathbf{Z}\|_* + \alpha\|\mathbf{B}\odot\mathbf{Q}\|_1 + \beta\mathrm{tr}\Big(\mathbf{F}^T(\mathbf{D} - \frac{|\mathbf{Q}|+|\mathbf{Q}^T|}{2})\mathbf{F}\Big) \\ &\quad + \lambda\sum_{i=1}^{N}\|\widehat{\mathbf{E}}_i\|_{2,1} + \langle\mathbf{\Gamma}_1, \mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\rangle + \langle\mathbf{\Gamma}_2, \mathbf{Z}-\mathbf{Q}\rangle \\ &\quad + \frac{\mu}{2}(\|\mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\|_F^2 + \|\mathbf{Z}-\mathbf{Q}\|_F^2),\end{aligned} \tag{5}$$

where $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are Lagrangian multipliers and μ is a penalty parameter. The optimization of (5) can be solved iteratively by minimizing $\mathcal{L}$ with respect to $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{F}$ and $\mathbf{E}$ one at a time, with the other variables being fixed. The LADM algorithm of CS-LRR is outlined in Algorithm 1. The optimization problem in steps 2 and 4 can be solved by using singular value thresholding operator (SVT) [15] and eigendecomposition of $\mathbf{M}^{k+1}$, respectively. Here, we only address the details of the solutions of $\mathbf{Q}$ and $\mathbf{E}$ in the interest of available space.

Update $\mathbf{Q}$ *when fixing other variables:* When other variables are fixed, the problem of updating $\mathbf{Q}$ is

$$\begin{aligned}\min_{\mathbf{Q}} \alpha\|\mathbf{B}\odot\mathbf{Q}\|_1 + \beta\mathrm{tr}\Big((\mathbf{F}^k)^T(\mathbf{D} - \frac{|\mathbf{Q}|+|\mathbf{Q}^T|}{2})\mathbf{F}^k\Big) \\ + \frac{\mu^k}{2}\|\mathbf{Q} - (\mathbf{Z}^{k+1} + \frac{\mathbf{\Gamma}_2^k}{\mu^k})\|_F^2.\end{aligned} \tag{6}$$

Note that we can simplify $\mathrm{tr}((\mathbf{F}^k)^T(\mathbf{D} - \frac{|\mathbf{Q}|+|\mathbf{Q}^T|}{2})\mathbf{F}^k)$ as $\mathrm{tr}((\mathbf{F}^k)^T(\mathbf{D} - \frac{|\mathbf{Q}|+|\mathbf{Q}^T|}{2})\mathbf{F}^k) = \|\mathbf{\Theta}^k\odot\mathbf{Q}\|_1$, where $\theta_{ij}^k = \frac{1}{2}\|\mathbf{f}^{k,i} - \mathbf{f}^{k,j}\|_2^2$. Here, $\mathbf{f}^{k,i}$ and $\mathbf{f}^{k,j}$ denote the i-th and the j-th row of the matrix $\mathbf{F}^k$, respectively. Therefore, the problem (6) is equivalent to the following problem:

$$\min_{\mathbf{Q}} \alpha\|(\mathbf{B} + \frac{\beta}{\alpha}\mathbf{\Theta}^k)\odot\mathbf{Q}\|_1 + \frac{\mu^k}{2}\|\mathbf{Q} - (\mathbf{Z}^{k+1} + \frac{\mathbf{\Gamma}_2^k}{\mu^k})\|_F^2,$$

which has a closed-form solution given in [10, Proposition 3].

Algorithm 1: Solving Problem (4) by LADM

Input: The data matrix $\mathbf{X}$ and matrix $\mathbf{B}$, and parameters α, β and λ.

Initialize: $\mathbf{Z}^0 = \mathbf{Q}^0 = \mathbf{\Theta}^0 = \mathbf{\Gamma}_1^0 = \mathbf{\Gamma}_2^0 = \mathbf{0}$, $\rho = 1.1$, $\eta > \|\mathbf{X}\|^2$, $\mu^0 = 0.1$, $\mu_{\max} = 10^{30}$, $\epsilon = 10^{-8}$, $k = 0$.

1: **while** not converged **do**

2: Fix other variables and update $\mathbf{Z}$:
$\mathbf{Z}^{k+1} = \arg\min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \frac{\eta\mu^k}{2}\|\mathbf{Z} - \mathbf{Z}^k + (\mathbf{X}^T(\mathbf{X}\mathbf{Z} - \mathbf{X} + \mathbf{E}^k - \frac{\mathbf{\Gamma}_1^k}{\mu^k}) + (\mathbf{Z} - \mathbf{Q}^k + \frac{\mathbf{\Gamma}_2^k}{\mu^k}))/\eta\|_F^2$.

3: Fix other variables and update $\mathbf{Q}$:
$\mathbf{Q}^{k+1} = \arg\min_{\mathbf{Q}} \alpha\|(\mathbf{B} + \frac{\beta}{\alpha}\mathbf{\Theta}^k)\odot\mathbf{Q}\|_1 + \frac{\mu^k}{2}\|\mathbf{Q} - (\mathbf{Z}^{k+1} + \frac{\mathbf{\Gamma}_2^k}{\mu^k})\|_F^2$.

4: Compute the Laplacian matrix:
$\mathbf{M}^{k+1} = \mathbf{D}^{k+1} - \frac{|\mathbf{Q}^{k+1}|+|(\mathbf{Q}^{k+1})^T|}{2}$ and update $\mathbf{F}$ by
$\mathbf{F}^{k+1} = \arg\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \mathrm{tr}(\mathbf{F}^T\mathbf{M}^{k+1}\mathbf{F})$.

5: Fix other variables and update $\mathbf{E}$:
$\mathbf{E}^{k+1} =$
$\arg\min_{\mathbf{E}} \lambda\|\mathbf{E}\|_\iota + \frac{\mu^k}{2}\|\mathbf{E} - (\mathbf{X} - \mathbf{X}\mathbf{Z}^{k+1} + \frac{\mathbf{\Gamma}_1^k}{\mu^k})\|_F^2$.

6: Update Lagrange multiplier:
$\mathbf{\Gamma}_1^{k+1} = \mathbf{\Gamma}_1^k + \mu^k(\mathbf{X} - \mathbf{X}\mathbf{Z}^{k+1} - \mathbf{E}^{k+1})$,
$\mathbf{\Gamma}_2^{k+1} = \mathbf{\Gamma}_2^k + \mu^k(\mathbf{Z}^{k+1} - \mathbf{Q}^{k+1})$.

7: Update the parameter μ^{k+1} by
$\mu^{k+1} = \min(\mu_{\max}, \rho\mu^k)$.

8: Check the convergence conditions and break if
$\|\mathbf{X} - \mathbf{X}\mathbf{Z}^{k+1} - \mathbf{E}^{k+1}\|_\infty \le \epsilon$, $\|\mathbf{Z}^{k+1} - \mathbf{Q}^{k+1}\|_\infty \le \epsilon$.

9: Update k by $k = k + 1$.

10: **end while**

Output: The optimal low-rank representation $\mathbf{Z}^*$.

Update $\mathbf{E}$ *when fixing other variables:* When other variables are fixed, the problem for updating $\mathbf{E}$ can be written as

$$\min_{\mathbf{E}} \lambda\sum_{i=1}^{N}\|\widehat{\mathbf{E}}_i\|_{2,1} + \frac{\mu^k}{2}\|\mathbf{X} - \mathbf{X}\mathbf{Z}^{k+1} - \mathbf{E} + \frac{\mathbf{\Gamma}_1^k}{\mu^k}\|_F^2.$$

For the sake of presentation, we define $\mathbf{C}^{k+1} = \mathbf{X} - \mathbf{X}\mathbf{Z}^{k+1} + \frac{\mathbf{\Gamma}_1^k}{\mu^k}$. This problem can be decomposed into N independent subproblems, where we update the i-th column of $\mathbf{E}$ (i.e., $\mathbf{e}_i$) by solving the following problem:

$$\widehat{\mathbf{E}}_i^{k+1} = \arg\min_{\widehat{\mathbf{E}}_i} \lambda\|\widehat{\mathbf{E}}_i\|_{2,1} + \frac{\mu^k}{2}\|\widehat{\mathbf{E}}_i - \widehat{\mathbf{C}}_i^{k+1}\|_F^2, \tag{7}$$

where $\widehat{\mathbf{C}}_i^{k+1} \in \mathbb{R}^{n_b\times 9}$ is the reshaped "image" of the vector $\mathbf{c}_i^{k+1}$. The problem (7) can be solved by the $\ell_{2,1}$ minimization operator [6].

After obtaining $\mathbf{Z}^*$, we set the coefficients below a given threshold to be zeros, and we denote the final representation matrix by $\widehat{\mathbf{Z}}$. By defining the affinity matrix $\mathbf{W} = \frac{|\widehat{\mathbf{Z}}|+|\widehat{\mathbf{Z}}^T|}{2}$, we apply spectral clustering [12] to obtain the

subspace clustering result, with the final clusters denoted by $\{\mathbf{X}_\ell \in \mathbb{R}^{m\times N_\ell}\}_{\ell=1}^{L}$. Finally, we estimate the subspaces $\mathcal{S}_\ell$'s underlying $\mathbf{X}_\ell$'s by identifying their orthonormal bases $\mathbf{P}_\ell$'s using the signals in each cluster. To be specific, we obtain eigendecomposition of the covariance matrix $\mathbf{C}_\ell = \mathbf{X}_\ell\mathbf{X}_\ell^T$ as $\mathbf{C}_\ell = \mathbf{U}_\ell\mathbf{\Sigma}_\ell\mathbf{U}_\ell^T$, where $\mathbf{\Sigma}_\ell = \mathrm{diag}(\lambda_1, \ldots, \lambda_{N_\ell})$ is a diagonal matrix ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N_\ell}$) and $\mathbf{U}_\ell = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_\ell}]$. Then the dimension of the subspace $\mathcal{S}_\ell$, denoted by d_ℓ, is estimated based on the energy threshold, i.e., $d_\ell = \arg\min_d \frac{\sum_{q=1}^{d}\lambda_q}{\sum_{q=1}^{N_\ell}\lambda_q} \geq \gamma$, where γ is a predefined threshold and is set close to 1 for better representation. In this paper, we set $\gamma = 0.98$. The orthonormal basis of $\mathcal{S}_\ell$ can then be written as $\mathbf{P}_\ell = [\mathbf{u}_1, \ldots, \mathbf{u}_{d_\ell}]$.

4. ACTION RECOGNITION USING CS-LRR

In this section, we describe the classification procedure to perform action recognition using the subspaces learned from CS-LRR. We first compute the distances between every training and test sequence. Let $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{|Y|}]$ and $\widehat{\mathbf{\Phi}} = [\widehat{\boldsymbol{\phi}}_1, \widehat{\boldsymbol{\phi}}_2, \ldots, \widehat{\boldsymbol{\phi}}_{|\widehat{Y}|}]$ denote two video sequences of lengths $|Y|$ and $|\widehat{Y}|$, respectively. We first apply Dynamic Time Warping (DTW) [16] to align the two action sequences using the HOG feature vectors and define $\mathbb{P}_{\mathbf{\Phi},\widehat{\mathbf{\Phi}}} = \{\boldsymbol{\phi}_{i_v}, \widehat{\boldsymbol{\phi}}_{j_v}\}_{v=1}^{V}$ to be the optimal alignment path computed from DTW, where V is the length of the optimal path. Then we assign every vector in these two sequences to the closest subspace in $\mathcal{S}_\ell$'s and we use $\boldsymbol{\psi} \in \mathbb{R}^{|Y|}$ and $\widehat{\boldsymbol{\psi}} \in \mathbb{R}^{|\widehat{Y}|}$ to denote the vectors which contain the resulting subspace assignment indexes of $\mathbf{\Phi}$ and $\widehat{\mathbf{\Phi}}$, respectively. Based on the optimal alignment path $\mathbb{P}$, the distance $\mathrm{dist}(\mathbf{\Phi}, \widehat{\mathbf{\Phi}})$ between these two action sequences is defined as the average of the normalized subspace distances on the alignment path: $\mathrm{dist}(\mathbf{\Phi}, \widehat{\mathbf{\Phi}}) = \frac{\sum_{v=1}^{V} d_u(\mathcal{S}_{\psi_{(i_v)}}, \mathcal{S}_{\widehat{\psi}_{(j_v)}})}{V}$, where $d_u(\mathcal{S}_\ell, \mathcal{S}_{\widehat{\ell}}) = \sqrt{1 - \frac{\mathrm{tr}(\mathbf{P}_\ell^T\mathbf{P}_{\widehat{\ell}}\mathbf{P}_{\widehat{\ell}}^T\mathbf{P}_\ell)}{\max(d_\ell, d_{\widehat{\ell}})}}$ [17]. Finally, we use the k nearest neighbor (k-NN) classifier to recognize actions based on sequence-to-sequence distances, i.e., a test sequence is declared to be in the class for which the average distance between the test sequence and the k nearest training sequences is the smallest.

5. PERFORMANCE EVALUATION

In this section, we evaluate the performance of our proposed method for human action recognition and compare our approach with several union-of-subspaces learning methods on Weizmann dataset [18] and Ballet dataset [3]. We compare the performance of CS-LRR with Low-Rank Representation (LRR) [6], Sparse Subspace Clustering (SSC) [5], Least Square Regression (LSR) [19], and Structure-Constrained LRR (SC-LRR) [10]. For these methods, we tune the parameters to achieve their best performance.

The Weizmann dataset consists of 10 different actions: walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack, and skip. Each action is performed by 9 subjects. We evaluate all the subspace/attribute learning approaches based on a leave-one-subject-out experiment. The original resolution of the frames is 180×144. We align all the sequences and crop them into 88×64 frames, then the HOG features are extracted with grid of size $n_\sigma = 8$, resulting in $m = 792$. We set $L = 20$ in the experiment as we expect every action to be associated with 2 attributes. We perform CS-LRR with parameters $\alpha = 1.4$, $\beta = 0.2$, and $\lambda = 0.8$. The results are listed in Table 1. We can see that by representing the human actions using the attributes learned by CS-LRR, we are able to recognize the actions at a superior rate compared to other techniques.

Table 1. Recognition results (%) on different datasets

Data ↓ Method →	CS-LRR	LRR	SSC	LSR	SC-LRR
Weizmann [18]	**83.33**	73.33	62.22	72.22	75.56
Ballet [3]	**61.02**	54.24	47.46	40.68	55.93

The Ballet dataset contains 44 video sequences of 8 unique actions. The eight actions are left-to-right hand opening, right-to-left hand opening, standing hand opening, leg swinging, jumping, turning, hopping, and standing still. Each video may contain several actions. Finally, we have 59 video clips and each video clip contains only one action. We crop all the video frames into 288×288 pixels and extract HOG descriptor of every frame with grid of size $n_\sigma = 32$; hence $m = 729$. Since there is no significant motion between consecutive frames, instead of using HOG features of each frame separately, we take the sum of the HOG features of two adjacent frames at a time and the sum is used as the feature of two adjacent frames. We perform CS-LRR with parameters $\alpha = 0.6$, $\beta = 0.1$, and $\lambda = 0.2$. We evaluate the subspace learning approaches based on a leave-one-sequence-out experiment, and we set $L = 10$ for all union-of-subspaces learning methods. The classification results are listed in Table 1. We can see that our proposed method again outperforms all other subspace clustering methods.

6. CONCLUSION

The proposed clustering-aware structure-constrained low-rank representation (CS-LRR) model overcomes the limitations of existing techniques by its ability to handle disjoint subspaces and perform optimal spectral clustering of the subspaces. The human action attributes generated by the proposed approach are a better representation of the movements of the human body than those generated by existing techniques. Thus, CS-LRR model produces better action recognition performance than other techniques when applied on real-world datasets that were used in our experiments.

7. REFERENCES

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimed. Inf. Retr.*, vol. 2, no. 2, pp. 73–101, 2013.

[2] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 3337–3344.

[3] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[4] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 461–468.

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

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

[7] T. Wu, P. Gurram, R. M. Rao, and W. U. Bajwa, "Hierarchical union-of-subspaces model for human activity summarization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2015, pp. 1053–1061.

[8] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2328–2335.

[9] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 663–670.

[10] K. Tang, R. Liu, Z. Su, and J. Zhang, "Structure-constrained low-rank representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2167–2179, 2014.

[11] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4238–4246.

[12] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 849–856.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[14] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 612–620.

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

[16] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

[17] L. Wang, X. Wang, and J. Feng, "Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition," *Pattern Recognit.*, vol. 39, no. 3, pp. 456–464, 2006.

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[19] C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 347–360.