Detecting Plumes in LWIR Using Robust Nonnegative Matrix Factorization with Graph-based Initialization

Jing Qin[⋆], Thomas Laurent[*], Kevin Bui[⋆], Ricardo Vicente R. Tan[*], Jasmine Dahilig[*], Shuyi Wang[⋆], Jared Rohe[†], Justin Sunu[‡], Andrea L. Bertozzi[⋆]

ABSTRACT

We consider the problem of identifying chemical plumes in hyperspectral imaging data, which is challenging due to the diffusivity of plumes and the presence of excessive noise. We propose a robust nonnegative matrix factorization (RNMF) method to segment hyperspectral images considering the low-rank structure of the noise-free data and sparsity of the noise. Because the optimization objective is highly non-convex, nonnegative matrix factorization is very sensitive to initialization. We address the issue by using the fast Nyström method and label propagation algorithm (LPA). Using the alternating direction method of multipliers (ADMM), RNMF provides high quality clustering results effectively. Experimental results on real single frame and multiframe hyperspectral data with chemical plumes show that the proposed approach is promising in terms of clustering quality and detection accuracy.

Keywords: hyperspectral images, label propagation, non-negative matrix factorization, Nyström extension, data analysis, image processing, robust principal component analysis, spectral clustering

1. INTRODUCTION

Hyperspectral imaging has been an active field of research recently and has provided a variety of applications including surveillance, astronomy, agriculture, and mineralogy. Airborne hyperspectral sensors capture a collection of spectral data with high wavelength resolution by utilizing the spectral radiance of different objects in a given scene. In particular, the long wavelength infrared (LWIR) hyperspectral imaging, which typically covers the spectral range from 7.8μm to 12μm, is widely used in defense and security to detect and identify chemical plumes present in the atmosphere. In industry production, gases are inevitably generated as part of the manufacturing process or as products themselves. The mass production of these chemicals can pose a risk in the rare occasion that safety precautions fail to prevent leaks. Likewise, mass production of gases as chemical weapons in terrorist states also poses a threat. As such, it is critically important to design an efficient and accurate method of detecting plumes from the background, which can be treated as a data clustering problem.

There are numerous data clustering and object detection algorithms for dealing with high-dimensional data. Given a set of data points $b_1, \ldots, b_n \in \mathbb{R}^m$ ($n \gg m$), the goal is to partition the entire data set into k clusters $S_1, \ldots, S_k$ according to a certain characteristic, e.g., the distance to the centroid of each cluster. The classical k-means method considers the problem

$$\min_{S_1,\ldots,S_k} \sum_{i=1}^{k} \sum_{b_j \in S_i} \|b_j - x_i\|_2^2 , \tag{1}$$

where x_i is the centroid of the ith cluster S_i. Let $B = [b_1, \ldots, b_n] \in \mathbb{R}^{m\times n}$ be the matrix whose columns are the data points, $X = [x_1, \ldots, x_k] \in \mathbb{R}^{m\times k}$ be the collection of all cluster centroids, and $Y \in \mathbb{R}^{k\times n}$ be the cluster indicator matrix, that is $Y_{ij} = 1$ if data point b_j belongs to cluster S_i and 0 otherwise. Note that since each

[⋆] Department of Mathematics, University of California Los Angeles, 520 Portola Plaza, Los Angeles, CA, USA
[*] Department of Mathematics, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA, USA
[†] Department of Mathematics, University of San Francisco, San Francisco, CA, USA
[‡] Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA, USA

data point belong to a single cluster, Y obviously satisfies that $\sum_{i=1}^{k} Y_{ij} = 1$ or equivalently $Y^T \mathbf{1_k} = \mathbf{1_n}$ where $\mathbf{1_k} \in \mathbb{R}^{\mathbf{k}}$ is a vector with all ones. Using these notations, the problem (1) can be recast as

$$\min_{X,\, Y^T \mathbf{1_k} = \mathbf{1_n},\, \mathbf{Y} \text{ is binary}} \|B - XY\|_F^2 , \tag{2}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix. Then the k-means method solves the above model by alternating the cluster assignment and centroid update iteratively. Although it is simple to implement, it suffers from the non-uniqueness of solution due to the non-convex nature of the objective function with respect to (X, Y). If the given data is noisy which causes oscillations in the distance, it is likely to generate inaccurate clustering result. To improve the k-means method, a large amount of variants are proposed, including moving k-means,[1] which modifies the k-means method to avoid inactive centroids, and fuzzy k-means clustering,[2] where each data point has a fuzzy degree in $[0, 1]$ of belonging to each cluster. However, if the given data is polluted by excessive noise or outliers, these methods become less robust and less efficient for large-scale data.

There is another category of data classification methods which treat the high-dimensional data as a graph G with vertices $b_1, \ldots, b_n$ and attempt to partition the graph into k clusters. The similarity (or weighted adjacency) matrix $W \in \mathbb{R}^{n \times n}$ of the graph G is defined by

$$W_{ij} = e^{-d(b_i, b_j)^2/\sigma^2}$$

which encodes how similar the data point b_i is from the data point b_j. Here $d(b_i, b_j)$ is usually one of the following two distance metrics:

- Euclidean distance: $d(b_i, b_j) = \|b_i - b_j\|_2$;
- Cosine similarity:[3] $d(b_i, b_j) = 1 - \frac{\langle b_i, b_j \rangle}{\|b_i\|_2 \|b_j\|_2}$.

Spectral clustering algorithms (e.g. the normalized cut algorithm[4]) are among the most popular graph partitioning algorithms. They proceed by computing then post-processing the eigenvectors corresponding to the k smallest eigenvalues of the graph Laplacian matrix $L = W - D$. Here D is the degree matrix, that is the diagonal matrix whose ith diagonal entry is the degree of vertex i. Normalized variant of the graph Laplacian, such as the symmetric normalized Laplacian $L_{sym} = I - D^{-1/2} W D^{-1/2}$ and the random walk normalized Laplacian $L_{rw} = I - D^{-1} W$ are also often considered. More details can be found in the tutorial on spectral methods for graph partitioning.[5] Besides spectral methods, total variation based algorithms, such as Multiclass Total Variation algorithm (MTV),[6,7] have recently been demonstrated to provide high quality graph partitions. Finally, in a semi-supervised context in which some vertices are a priori known to belong to some cluster (we will refer to this vertices as being *labeled*), label propagation algorithms (LPA)[8–11] provide a simple and efficient way to obtain a partition by diffusing the labels along the graph.

Furthermore, large-scale data can be classified efficiently by using the nonnegative matrix factorization (NMF).[12] By relaxing the constraints in (2), NMF attempts to partition the data matrix B as a product of two matrices X and Y both with nonnegative entries, i.e.,

$$\min_{X,Y \geq 0} \|B - XY\|_F^2 . \tag{3}$$

There are some popular algorithms to solve (3), e.g., alternating least squares,[13] multiplicative update,[14] and the method[15] based on the alternating direction method of multipliers (ADMM). Introduced initially as positive matrix factorization[13] and later developed,[12] the NMF technique has been widely used and developed in a broad variety of applications, such as text mining, document clustering, computer vision, signal processing and many others. Some variants include symmetric NMF,[16] which attempt to find a factorization $W \approx Y^T Y$ of the similarity matrix of a graph, and NMF with various cost functions and regularizations[17] in the objective function. It has been shown that the symmetric NMF is equivalent to some variant of kernel k-means clustering and spectral clustering.[16] Similar to (2), the objective function in (3) is also not convex with respect to (X, Y), and therefore NMF is highly sensitive to the initialization.

Combining the powers of graph-based methods and NMF, we propose a robust NMF (RNMF) method with graph-based initialization in an attempt to classify the LWIR hyperspectral data polluted by excessive noise, which can be characterized by a sparse matrix. We start with the Nyström extension and LPA to obtain an initial guess, and then apply the RNMF to the original data to obtain a more refined classification result. Some related work using the same LWIR data includes simultaneous spectral analysis from multiple videos[18] and TV-based clustering methods.[3,19]

The paper is organized as follows. The graph based initialization method combining the Nyström method and LPA is presented in Section 2. Section 3 details the proposed RNMF model and the associated algorithm by using ADMM. The experimental results on the real data are shown in Section 4. We finalize the paper with concluding remarks in Section 5.

2. GRAPH BASED INITIALIZATION

In this section, we describe the graph-based initialization: at the first step we use the Nyström method to extract labels, then we propagate these labels along the graph using a random walk.

2.1 Label Extraction Using Nyström Method

In general, the spectral clustering methods involve the computation of eigenvalues and eigenvectors of large matrices and many distance calculations. To alleviate the computational burden, the spectral grouping based on the Nyström extension[20] was proposed to approximate the eigenpair (i.e., eigenvalue and its associated eigenvector) of the similarity matrix by using few random samples. Recently, Nyström method has been developed for the hyperspectral data,[18] and then was employed in more recent TV-based clustering.[3,19] Let W be the similarity matrix of the data matrix B, which can be rearranged and partitioned as the following block-matrix form

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}.$$

Here W_{11} is the similarity matrix of the sample points, $W_{12} = W_{21}^T$ is the similarity matrix of the sample points and the remaining points of the data set, and W_{22} is the similarity matrix of the remaining points. Suppose that the eigendecomposition of W_{11} is $U\Lambda U^T$ where U is unitary and Λ is diagonal. As a popular technique to solve integral eigenvalue problems, the Nyström extension yields

$$W \approx \bar{U}\Lambda\bar{U}^T := \hat{W} \quad \text{where} \quad \bar{U} = \begin{bmatrix} U \\ W_{21}U\Lambda^{-1} \end{bmatrix}.$$

By direct computation, we can see that W_{22} is approximated by $W_{21}W_{11}^{-1}W_{12}$. Since the columns of $\bar{U}$ are not orthogonal, we further orthogonalize $\bar{U}$.

Now that we have the eigenvector approximations for spectral clustering, we can extract the labels from them to initialize LPA. Each eigenvector can be reshaped to be a new image containing one or two specific features of the original image. If a hyperspectral image has k features, we select the k eigenvectors starting from the second one. Because we have no knowledge of what feature corresponds to an eigenvector, we threshold it to obtain a binary image. In our thresholding scheme, we set all values less than zero to -1 and all values greater than zero to 1. After visualizing the binary image, we are able to find the cluster corresponding to the desired feature. Finally, the obtained l labels are the l indices corresponding to the lth largest or the lth smallest components of the eigenvector representing that feature. The sampling scheme that we applied is detailed in Section 4.

2.2 Propagating Labels With a Random Walk

Following a common approach in semi-supervised learning,[8–11] we then use a random walk in order to propagate the labels along the graph. Suppose that we have n_r labels of class r, for $r = 1, \ldots, k$. Let F be the $k \times n$ (normalized) indicator matrix of these labels:

$$F_{r,i} = \begin{cases} 1/n_r, & \text{if } b_i \text{ has been labeled to belong to cluster } r; \\ 0, & \text{otherwise.} \end{cases}$$

Note that the rows of F are normalized to sum to one, so that they can be interpreted as probability distribution on the vertices. We consider k random walkers on the graph, one for each class. The rth row of F is the initial probability distribution of the rth random walker, that is this random walker start uniformly at random on one of the label of class r. After ν steps of random walk, the probability distribution of each random walker is given by

$$\tilde{F} = M^{\nu} F.$$

Here $M = D^{-1}W$ is the random-walk Laplacian used as the transition matrix, that is M_{ij} is the probability to go from vertex i to vertex j. Each unlabeled vertex is then associated to the class of the random walker who is the most likely to visit it, that is

$$\text{class of vertex } i \ = \underset{r=1,\ldots,k}{\operatorname{argmax}} (M^{\nu} F)_{r,i}.$$

The algorithm is summarized in Algorithm 1.

Algorithm 1 Label Propagation Algorithm

Input: similarity matrix $W \in \mathbb{R}^{N\times N}$, degree matrix $D \in \mathbb{R}^{N\times N}$, number of clusters k, initial label indicator matrix $F \in \mathbb{R}^{k\times n}$
for $r = 1$ to R **do**
  $F \leftarrow F(D^{-1}W)$
end for
for $i = 1$ to n **do**
  $c_i = \operatorname{argmax}_{r=1,\ldots,k} F_{r,i}$
end for
return class assignment vector $(c_1, \ldots, c_n)$

3. PROPOSED ROBUST NONNEGATIVE MATRIX FACTORIZATION

Motivated by the ℓ_1-regularization in data analysis, e.g., Robust PCA,[21] we model the noise in LWIR images as sparse matrices. In fact, some denoising methods such as median filter,[22] have experimentally shown to fail to remove any significant outliers of the given data (see Section 4). Assume that the acquired data B is polluted by noise, which is characterized as a sparse matrix S, i.e.,

$$B = L + S, \qquad L = XY, \quad X, Y \geq 0.$$

Here L has low-rank structure characterizing the noise-free part of the data to be further factored out. Then we consider the following low-rank nonnegative matrix factorization model

$$\min_{X,Y\geq 0, L+S=B} \|L\|_* + \lambda \|S\|_1 + \frac{\rho}{2} \|L - XY\|_F^2 . \tag{4}$$

where $\|L\|_*$ is the nuclear norm of L, i.e., the sum of all singular values of L. Similar to the relaxation technique,[21] we consider the nuclear norm based regularization instead of rank so that the objective function in (4) is convex with respect to L. To solve the above optimization problem, we resort to ADMM which breaks the original bulky problem into small pieces, each of which is easier to solve. To start with, we construct the following augmented Lagrangian

$$\begin{aligned}\mathcal{L}(L, S, X, Y, U, V, \Gamma, \Pi, \Sigma) = \|L\|_* + \lambda \|S\|_1 + \frac{\rho}{2} \|L - XY\|_F^2 + \frac{\gamma_1}{2} \|X - U + \Gamma\|_F^2 \\ + \frac{\gamma_2}{2} \|Y - V + \Pi\|_F^2 + \frac{\gamma_3}{2} \|B - L - S + \Sigma\|_F^2 + \iota_+(X) + \iota_+(Y) + \iota_+(U) + \iota_+(V),\end{aligned} \tag{5}$$

where ι_+ is the indicator function defined by

$$\iota_+(U) = \begin{cases} 0, & U \geq 0; \\ \infty, & \text{otherwise.} \end{cases}$$

Then ADMM yields a sequence of subproblems for each iteration due to the separability of the objective function in (4). To solve subproblems explicitly, we first recall some relevant operators in convex optimization. Given a function f on $\mathbb{R}^{m\times n}$, the proximal operator of f is defined by

$$\text{prox}_{\mu f}(x) = \underset{y\in\mathbb{R}^{m\times n}}{\text{argmin}} \left\{ f(y) + \frac{\mu}{2}\|y-x\|_F^2 \right\}.$$

By direct calculations, there are closed-form solutions for the following two classical types of optimization problems

$$\underset{A}{\text{argmin}} \|A\|_1 + \mu/2\|A-B\|_F^2 = \text{prox}_{\mu\|\cdot\|_1}(B) := \mathcal{S}_{1/\mu}(B), \tag{6}$$

$$\underset{A}{\text{argmin}} \|A\|_* + \mu/2\|A-B\|_F^2 = \text{prox}_{\mu\|\cdot\|_*}(B) := \mathcal{D}_{1/\mu}(B). \tag{7}$$

Here $\mathcal{S}$ is called soft-thresholding or shrinkage defined by

$$\mathcal{S}_{1/\mu}(B) = \text{sign}(B)\odot\max\{|B|-1/\mu, 0\}$$

where $\text{sign}(B)$ is the entrywise signum function, and $\odot$ is entrywise multiplication. Additionally, the operator $\mathcal{D}$ corresponds to the singular value thresholding (SVT)[23] in matrix completion defined by

$$\mathcal{D}_{1/\mu}(B) = U\mathcal{S}_{1/\mu}(\Sigma)V^T \quad \text{where } B = U\Sigma V^T.$$

To comply with the form in (7), the L-subproblem can be reformulated as

$$\begin{aligned}&\underset{L}{\text{argmin}} \left\{ \|L\|_* + \frac{\rho}{2}\|L-XY\|_F^2 + \frac{\gamma_3}{2}\|L+S-B-\Sigma\|_F^2 \right\} \\ =& \underset{L}{\text{argmin}} \left\{ \|L\|_* + \frac{\rho+\gamma_3}{2}\left\| L - \frac{\rho XY + \gamma_3(B+\Sigma-S)}{\rho+\gamma_3} \right\|_F^2 \right\}\end{aligned}$$

In summary, the detailed description of the proposed algorithm is presented in Algorithm 2.

In addition, to suppress noise more brutally, we could extend the proposed method by replacing the Frobenius norm in (4) by the ℓ_1-norm of $L-XY$ and then solve the modified model by ADMM. In that case, it involves more auxiliary variables and subproblems while the improvement in performance is limited, and therefore it is not useful in practice.

4. EXPERIMENTAL RESULTS

In this section, we conduct experiments on the selected frames from the two sets of LWIR hyperspectral data, called aa-12 and aa-13, provided by the Applied Physics Laboratory at Johns Hopkins University. The goal is to obtain four clusters, specifying atmosphere, mountain, foreground, and gas plume. The selection of the number of clusters is out of the scope for this paper. Each frame of hyperspectral data has the dimension of $128\times 320\times 129$, where the last dimension indicates the number of wavelengths. Each wavelength records a particular frequency from 7,820 nm to 11,700 nm, spaced 30 nm apart. The testing data is acquired from three LWIR spectrometers, each placed at a different location about two kilometers away from the release of a chemical plume at an elevation of around 1300 feet. One hyperspectral data cube is captured every five seconds.

Because the quality of a hyperspectral image is heavily affected by massive amount of noise, it becomes challenging to obtain a clear classification. To eliminate significant outliers, we use median filter[22] as a preprocessing step of classification, but unfortunately, a large amount of noise is still present. In an efficient manner, the proposed RNMF is able to suppress the noise while performing matrix factorization, which thereby yields more accurate clustering results in the presence of excessive noise. The parameters we select for Algorithm 2 are based on prior results.[15,24] We set parameters as $\lambda = 1/\sqrt{n}$, $\gamma_1 = m/k$, $\gamma_2 = \gamma_1/k$, $\gamma_3 = mn/4\|B^T\|_1$, and $\theta = 1.0$. It also has been empirically proven that it is more likely to achieve good performance when $\rho \ll 1$.

Algorithm 2 Robust Nonnegative Matrix Factorization (RNMF)

Input: data matrix B, number of clusters k, Y^0, maximal number of iterations N, tolerance ϵ, parameters $\rho, \lambda, \gamma_1, \gamma_2, \gamma_3 > 0$, and $\theta \in (0, (\sqrt{5}+1)/2)$.
Initialize $U^0, V^0, L^0, S^0, \Gamma^0, \Sigma^0$ as zero matrices.
for $k = 0$ to N **do**

$$L^{k+1} = \mathcal{D}_{1/(\rho+\gamma_3)}((\rho+\gamma_3)^{-1}(\rho X^k Y^k + \gamma_3(B + \Sigma^k - S^k)))$$
$$S^{k+1} = \mathcal{S}_{\lambda/\gamma_3}(B - L^{k+1} + \Sigma^k)$$
$$X^{k+1} = \mathcal{P}_+((\rho L^{k+1}(Y^k)^T - \gamma_1\Gamma^k + \gamma_1 U^k)(\rho Y^k (Y^k)^T + \gamma_1 I)^{-1})$$
$$Y^{k+1} = \mathcal{P}_+((\rho(X^{k+1})^T X^{k+1} + \gamma_2 I)^{-1}(\rho(X^{k+1})^T L^{k+1} - \gamma_2\Pi^k + \gamma_2 V^k))$$
$$U^{k+1} = \mathcal{P}_+(\Gamma^k + X^{k+1})$$
$$V^{k+1} = \mathcal{P}_+(\Pi^k + Y^{k+1})$$
$$\Gamma^{k+1} = \Gamma^k + \theta(X^{k+1} - U^{k+1})$$
$$\Pi^{k+1} = \Pi^k + \theta(Y^{k+1} - V^{k+1})$$
$$\Sigma^{k+1} = \Sigma^k + \theta(B - L^{k+1} - S^{k+1})$$
$$f^{k+1} = \frac{\rho}{2}\left\|X^{k+1}Y^{k+1} - L\right\|_F^2 + \frac{\gamma_3}{2}\left\|B - L^{k+1} - S^{k+1} + \Sigma^{k+1}\right\|_F^2$$

If $\frac{|f^{k+1}-f^k|}{|f^k|} \leq \epsilon$, then it stops.

end for
return X^{k+1} and Y^{k+1}.

We use the result from LPA as the initial guess for Y in Algorithm 2. To initialize LPA, we require labels to indicate which pixels are atmosphere, mountain, gas plume, or foreground. We use the label extraction method as outlined in Section 2.1, where we sample 100 points, and choose the cosine similarity as distance metric and $\sigma = 1$ in the Nyström method for constructing the similarity matrix. We obtain labels for atmosphere, mountain, and foreground by using a background image containing no gas plumes. Then we obtain the labels for plumes by using a reference frame with the desired gas plume. In both data sets aa-12 and aa-13, we select the second, the third, and the fifth eigenvectors of the similarity matrix because the second eigenvector corresponds to both atmosphere and background, the third eigenvector corresponds to the mountain, and the fifth eigenvector corresponds to the gas plume. The number of labels used for our experiments is 100 for each feature, so we have a total of 400 labels per data set. They are used to initialize Algorithm 1. The similarity matrix W in Algorithm 1 is a sparse matrix corresponding to a weighted kNN graph constructed by using the kd-tree query algorithm from the VLFeat Library with each vertex having 5 nearest neighbors with maximum comparison to 10 other vertices.[25]

4.1 Single Frame Analysis

We first perform the single frame analysis and test the proposed method on two specific frames of the hyperspectral data, i.e., aa-12 frame 378 and aa-13 frame 726. Then we compare it with the fuzzy k-means with cosine metric, NMF with random initialization, and NMF initialized by label propagation (LPA). Before carrying out any clustering algorithm, we reshape the testing hyperspectral frame into a matrix of 129×40960. Because fuzzy k-means and NMF are highly sensitive to noise, we preprocessed the data by applying the Robust PCA with suggested parameters.[24] Before running any NMF algorithms, including NMF and RNMF, we scale the input data matrix B in (3) or (4) so that the Frobenius norm of B is 5.0×10^6 as suggested.[15] For RNMF, we choose $\rho = 0.01$ in Algorithm 2. The results for each method are shown in Figures 1 and 2.

From the results, we can see that the proposed method performs best since it is able to identify the gas plume successfully and distinguish clusters more accurately than the other methods for both aa-12 and aa-13. RNMF produces better soft clustering results since each feature is identified in its own frame while the other methods misidentify one feature as part of another. For example, NMF with random initialization identifies the

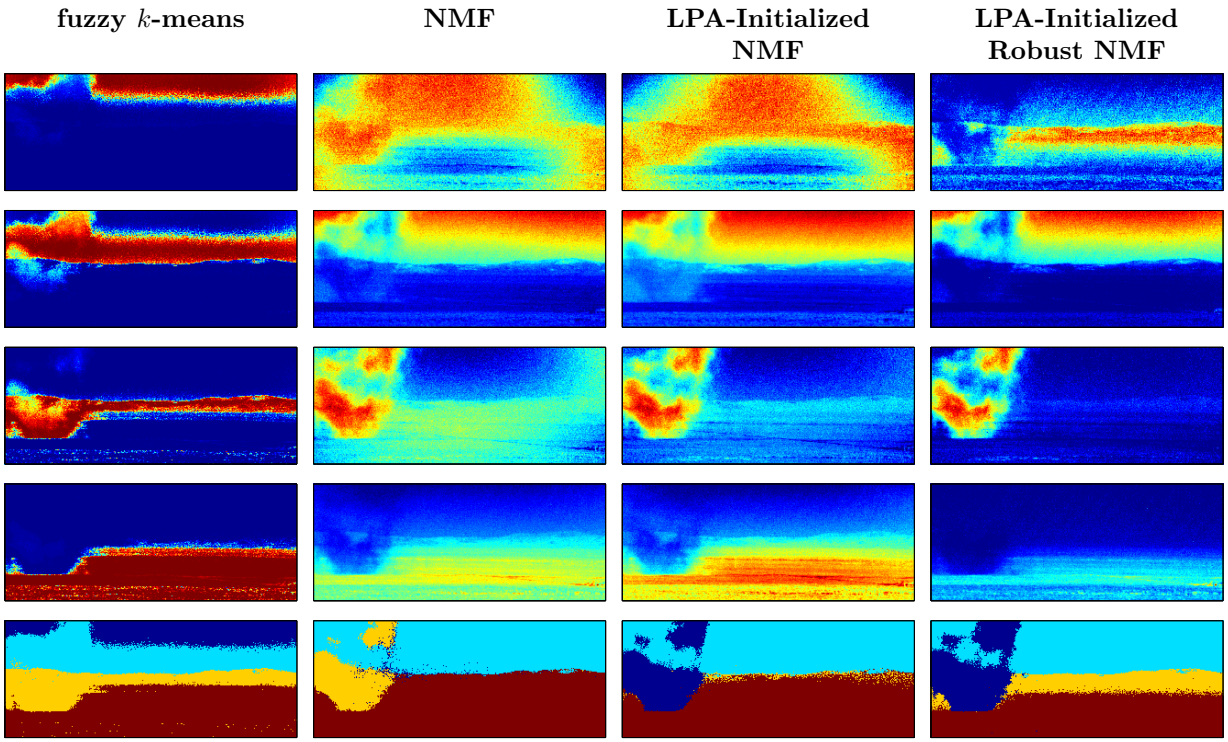

Figure 1: Comparison of performance for aa-12 frame 378. The images in row 1-4 are obtained by reshaping each row of Y in Algorithm 2 to a matrix. The last row displays the hard clustering results for each method.

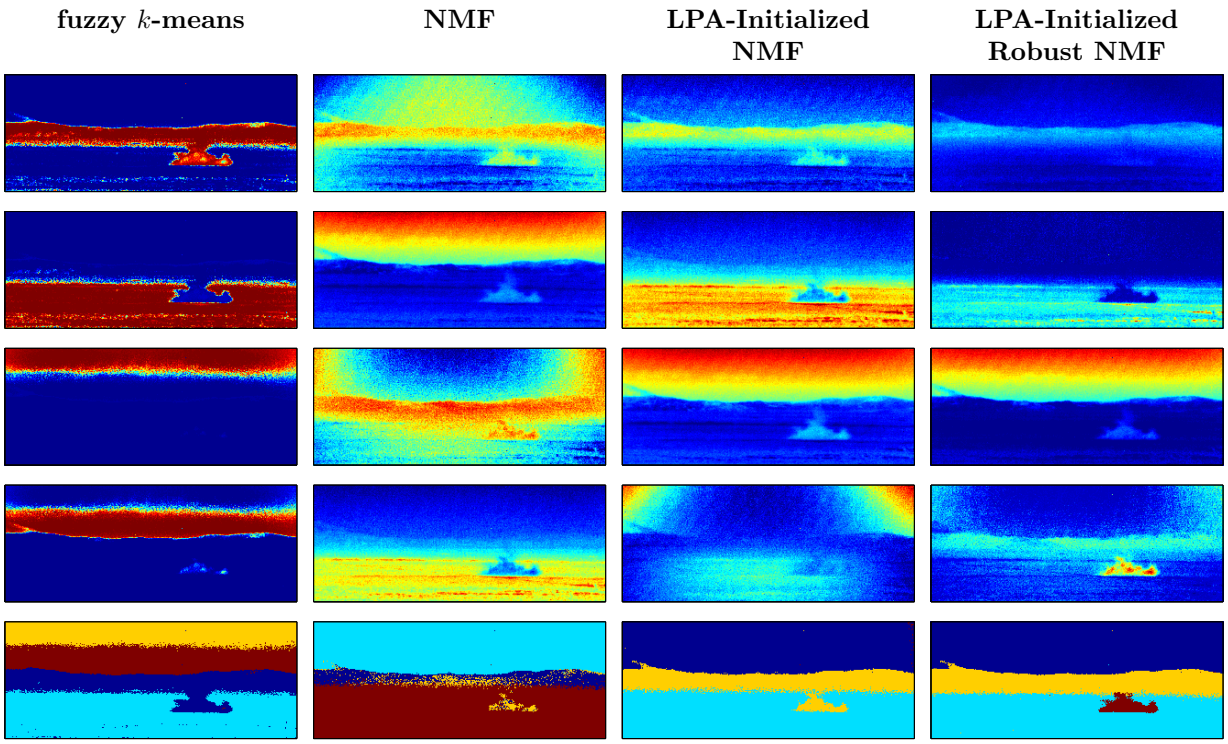

Figure 2: Comparison of performance for aa-13 frame 726. The images in row 1-4 are obtained by reshaping each row of Y in Algorithm 2 to a matrix. The last row displays the hard clustering results for each method.

gas plume as part of the foreground in Figure 1 and as part of the mountain in Figure 2. Since LPA-Initialized Robust NMF produces the best soft clustering results out of the four methods, it thereby produces the best hard clustering result.

We compare the hard clustering performance of the proposed method with other hard clustering techniques with results shown in Figure 3. Although AMSD[26] is not a hard clustering technique since it does only binary classification, we include its results as a baseline. Again, we can see that our result is clearer than the others. Moving k-means with cosine metric fails to detect the gas plume from the mountain or atmosphere. Multiclass total variation with Euclidean distance is able to segment the gas plume for aa-12, but it fails in aa-13. LPA, on the other hand, succeeds to obtain four clusters in both aa-12 and aa-13, but each cluster still contains certain amount of noise and the boundary of two clusters is also corrupted by noise.

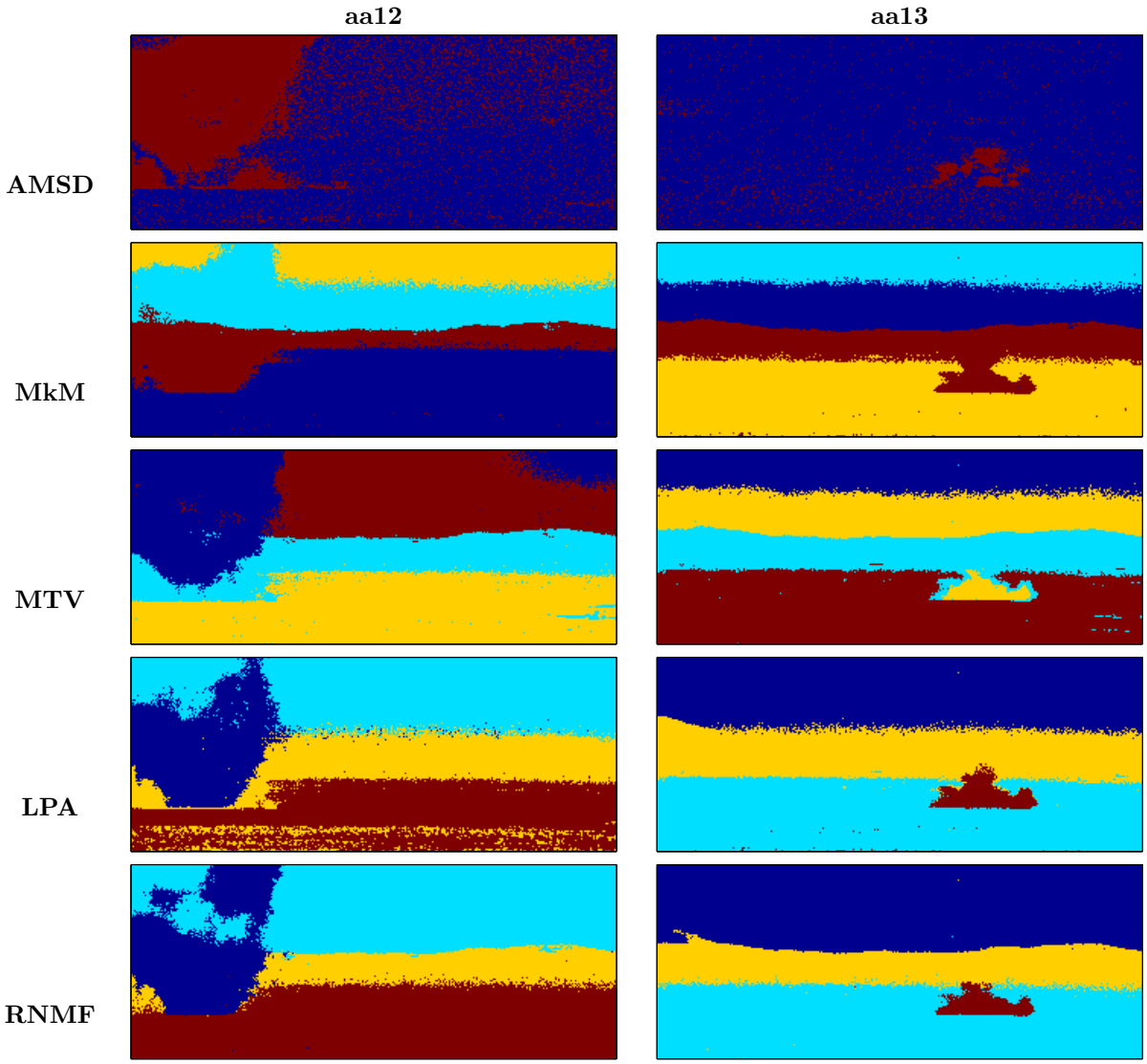

Figure 3: Comparison of hard clustering results for aa-12 frame 378 and aa-13 frame 726.

4.2 Multiframe Analysis

In this subsection, we incorporate temporal information and apply the proposed method for multiframe analysis. We examine two series of hyperspectral data each of which contains 20 frames, more specifically the data set

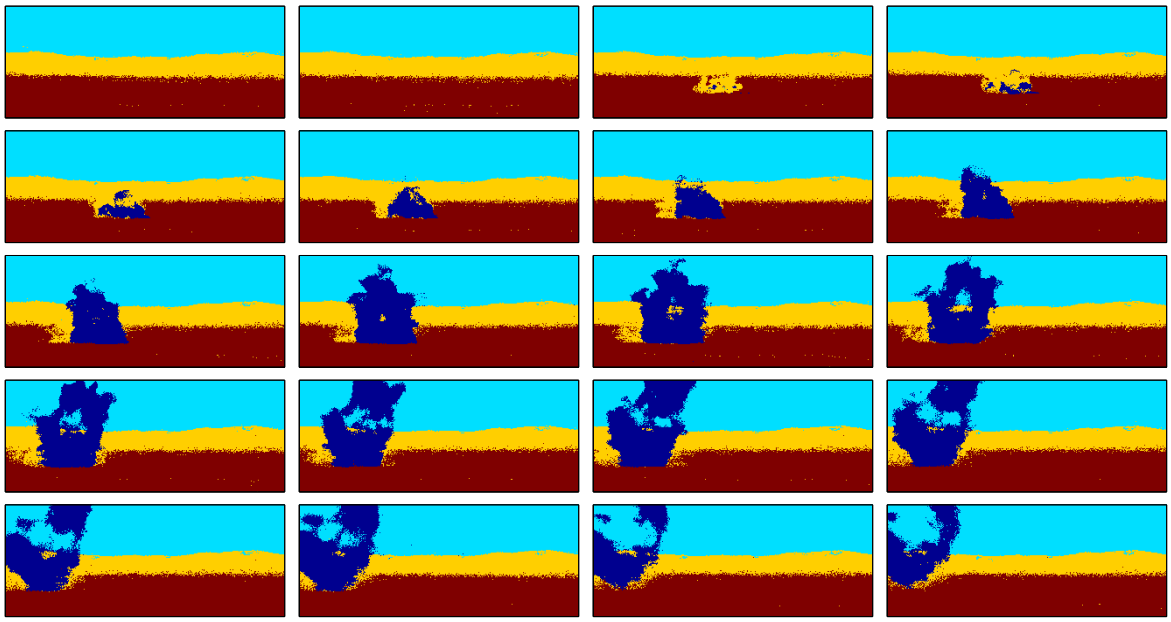

Figure 4: Proposed hard clustering results for aa12 frame 362-381.

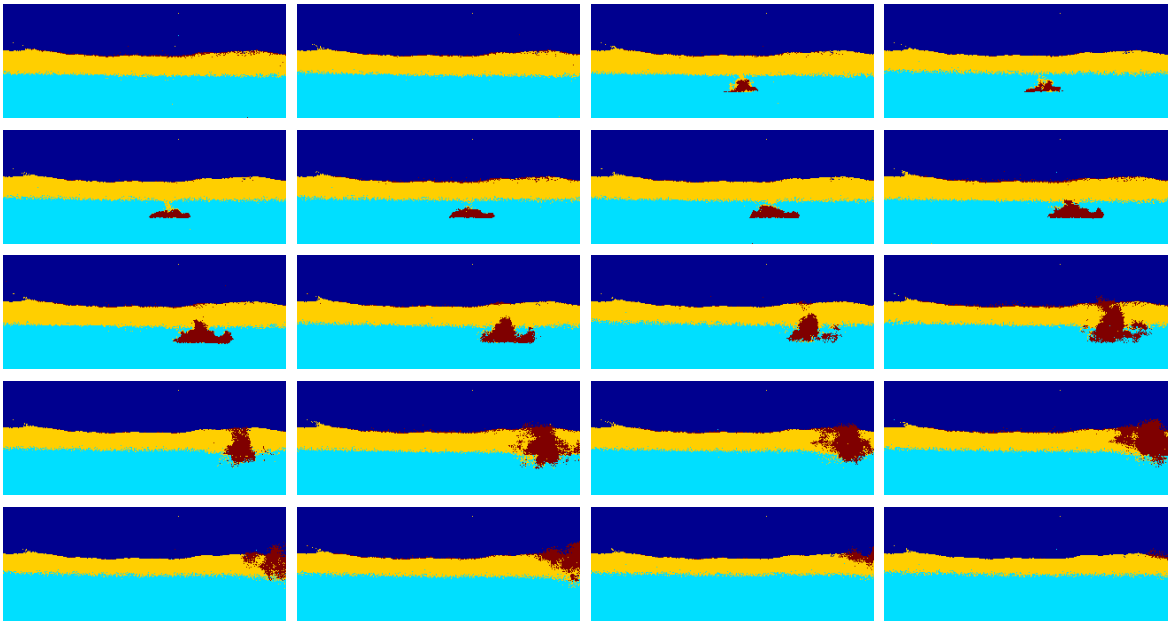

Figure 5: Proposed hard clustering results for aa13 frame 719-738.

aa-12 frames 362-380 and the data set aa-13 frames 719-738. Each raw data set is first reshaped to a matrix of size 129×829200 by concatenating all single frames of size 129×40960. We set $\rho = 10^{-3}, 10^{-1}$ in Algorithm 2 for aa-12 and aa-13, respectively, which achieve the best results in our experiments. In fact, we can see that the gas plume is mixed with the mountain as ρ becomes smaller while it is mixed with the atmosphere as ρ becomes larger. It is sufficient to run RNMF for 20 iterations for each data set. From the results shown in Figures 4 and 5, we can see that the proposed method has great potential in detecting gas plumes robustly and efficiently for hyperspectral videos.

5. CONCLUSIONS

In this paper, we proposed the robust nonnegative matrix factorization method with graph-based initialization to segment hyperspectral images and video sequences in a computationally efficient manner. To address the non-convexity of the objective function, we apply the Nyström method and LPA to obtain a reliable initial guess in a extremely fast way. Taking into account the excessive noise present in the hyperspectral data, we decompose the original data into a sparse matrix specifying the noise, and a low-rank matrix representing the noise-free part to be further factored into two nonnegative matrices. The resultant label information is retrieved by using one factor matrix. Numerical experiments on real LWIR hyperspectral data have shown that the proposed approach is able to detect gas plumes from the noisy background more reliably than some state-of-the-art methods. Furthermore, since the Nyström method is based on random sampling which may fail to provide good initial guess sometimes, fast unsupervised graph-based methods could be incorporated in our framework as future work.

6. ACKNOWLEDGMENTS

The authors would like to thank the Johns Hopkins Applied Physics Laboratory for providing the raw data sets and the associated supplementary documents. This project was funded by NSF grants DMS-1417674, DMS-1118971, DMS-1045536, and DMS-1414396, Office of Naval Research (ONR) grant N000141210838, and UC Lab Fees Research grant 12-LR-236660. Jared Rohe was supported by NSF grant DMS-1312361.

REFERENCES

[1] M. Y. Mashor. Hybrid Training Algorithm for RBF Network. *International Journal of Computer, Internet and Management*, 8(2):50–65, 2000.

[2] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*, volume 20. 2007.

[3] H. Hu, J. Sunu, and A. L. Bertozzi. Multi-class Graph Mumford-Shah Model for Plume Detection Using the MBO scheme. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 209–222. Springer, 2015.

[4] J. Shi and J. Malik. Motion Segmentation and Tracking Using Normalized Cuts. In *Computer Vision, 1998. Sixth International Conference on*, pages 1154–1160, 1998.

[5] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[6] X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht. Multiclass Total Variation Clustering. *Advances in Neural Information Processing Systems*, pages 1421–1429, 2013.

[7] X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht. Convergence and energy landscape for Cheeger cut clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1394–1402, 2012.

[8] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107*, 2002.

[9] J. Xie and B. K. Szymanski. LabelRank: a stabilized label propagation algorithm for community detection in networks. In *Proc. IEEE Network Science Workshop*, pages 138 – 143, 2013.

[10] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E.*, 76(3), 2007.

[11] S. M. Van Dongen. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, 2001.

[12] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[13] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[14] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[15] Y. Zhang. An Alternating Direction Algorithm for Nonnegative Matrix Factorization. Technical Report TR10-03, Rice University, January 2010.

[16] C. H. Ding, X. He, and H. D. Simon. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In *SDM*, volume 5, pages 606–610. SIAM, 2005.

[17] R. Rajabi and H. Ghassemian. Unmixing of hyperspectral data using robust statistics-based NMF. In *Telecommunications (IST), 2012 Sixth International Symposium on*, pages 1157–1160. IEEE, 2012.

[18] J. Sunu, J. M. Chang, and A. L. Bertozzi. Simultaneous spectral analysis of multiple video sequence data for LWIR gas plumes. *SPIE Defense and Security Conference, Baltimore, MD*, pages 90880T–90880T, 2014.

[19] E. Murkerjev, J. Sunu, and A. L. Bertozzi. Graph MBO method for multiclass segmentation of hyperspectral stand-off detection video. In *Proc. Int. Conf. Image Proc. (ICIP)*, pages 689–693, Paris, 2014.

[20] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):214–225, 2004.

[21] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *In Intl. Workshop on Comp. Adv. in Multi-Sensor Adapt. Processing, Aruba, Dutch Antilles*, page 61, 2009.

[22] T. Gerhart, J. Sunu, L. Lieu, E. Merkurjev, J.M. Chang, J. Gilles, and A. L. Bertozzi. Detection and Tracking of Gas Plumes in LWIR Hyperspectral Video Sequence Data. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIX*, pages 87430J–87430J. Proc. SPIE 8743, 2013.

[23] J. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[24] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[25] A. Vedaldi and B. Fulkerson. Vlfeat: An Open and Portable Library of Computer Vision Algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

[26] D. Manolakis, C. Siracusa, and G. Shaw. Adaptive matched subspace detectors for hyperspectral imaging applications. In *2001 IEEE International Conference*, volume 5, pages 3153–3156, 2001.