Fast Robust Non-Negative Matrix Factorization for Large-Scale Human Action Data Clustering

De Wang, Feiping Nie, Heng Huang*
Department of Computer Science and Engineering
University of Texas at Arlington
wangdelp@gmail.com, feipingnie@gmail.com, heng@uta.edu

Abstract

Human action recognition is important in improving human life in various aspects. However, the outliers and noise in data often bother the clustering tasks. Therefore, there is a great need for the robust data clustering techniques. Nonnegative matrix factorization (NMF) and Nonnegative Matrix Tri-Factorization (NMTF) methods have been widely researched these years and applied to many data clustering applications. With the presence of outliers, most previous NMF/NMTF models fail to achieve the optimal clustering performance. To address this challenge, in this paper, we propose three new NMF and NMTF models which are robust to outliers. Efficient algorithms are derived, which converge much faster than previous NMF methods and as fast as K-means algorithm, and scalable to large-scale data sets. Experimental results on both synthetic and real world data sets show that our methods outperform other NMF and NMTF methods in most cases, and in the meanwhile, take much less computational time.

1 Introduction

Human action recognition is very useful in improving human life in various aspects, like smart home applications, rehabilitation, human-computer interaction, *etc.* To recognize human action is helpful to know one's intent, such that an intelligent system could make corresponding reactions. However, in the real world life, the data are not as clean as we want, and there might be many outliers and noise. For example, it is common that part of the human body is shielded behind an obstacle, which makes the human action difficult to recognize. Therefore, robust data mining techniques should be developed in order to achieve good performance.

Non-negative matrix factorization (NMF) and non-negative matrix tri-factorization (NMTF) are models which aim to factorize a matrix into non-negative matrices with minimum reconstruction error. It has been widely researched and used in various kinds of applications, like document co-clustering [Li and Ding, 2006], computer vision [Lee and Seung, 1999], social network analysis [Huang *et al.*, 2013], bioinformatics [Wang *et al.*, 2012; 2014b], knowledge transfer [Wang *et al.*, 2011; 2015] and many others.

Different kinds of NMF and NMTF models are proposed these years. The standard NMF factorize a non-negative matrix into two non-negative matrices. [Ding *et al.*, 2005] discussed the equivalence relationship between K-means, NMF and spectral clustering. Later, [Ding *et al.*, 2010] proposed two NMF models: semi-NMF and convex NMF. In many situations, the data matrix may contain negative elements. In such situation, it is not suitable to enforce the non-negative constraints on NMF model factors. Different from standard NMF, semi-NMF relaxed the non-negative constraint on NMF models, allowed one factor to be mix-signed. In convex NMF, the author restricts that the centroid to be a linear combination of samples. Such model has better interpretability, however, the performance may not be as good as other NMF models [Li and Ding, 2006]. [Ding *et al.*, 2006] proposed to add orthogonality constraints in the cluster indicator factor G. This property will enforce the solution of G to have a clear cluster structure, which is desirable for using NMF for clustering. It is pointed out in [Gu *et al.*, 2011] that the orthogonal constraint can prevent trivial solution of NMF models. [Gu and Zhou, 2009] combines Laplacian graph based algorithms with NMF to enforce the smoothness on data manifolds. [Kong *et al.*, 2011; Gao *et al.*, 2015] proposed robust NMF models which are robust to outliers.

Most previous works can not achieve good clustering results when there exist outliers in the data. Presence of outliers is very common in real world applications. For example, in human action recognition, it is common that part of the human body is shielded behind an obstacle, which makes the human action difficult to recognize. Therefore, robust data mining techniques should be developed in order to achieve good performance.

In view of the above considerations, we propose three NMF and NMTF models which converge as fast as K-means, and are robust to outliers.

Notations: Given a matrix $X \in \mathbb{R}^{n\times m}$, its i-th row and j-th column are denoted by $X_{i\cdot}$, $X_{\cdot j}$, respectively. The $\ell_{r,p}$-norm of a matrix is defined as:

*To whom all correspondence should be addressed. This work was partially supported by US NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NIH R01 AG049371.

$\|X\|_{r,p} = (\sum_{i=1}^{n}(\sum_{j=1}^{m}|x_{ij}|^r)^{\frac{p}{r}})^{\frac{1}{p}}$. According to this definition, we can get the following norms:

$$\|X\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}x_{ij}^2} = \sum_{i=1}^{n}\left\|x^i\right\|_2 \quad (1)$$

$$\|X\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{m}|x_{ij}| \quad (2)$$

$$\|X\|_F = \|X\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}^2} \quad (3)$$

2 New Fast Algorithms for Robust NMF and NMTF Models

2.1 Motivations

NMF models often have the following form of objective functions, subjecting to different kinds of constraints.

$$\min_{F,G}\left\|X - FG^T\right\|_F^2 \qquad s.t. \quad F \geq 0, G \geq 0 \quad (4)$$

where $X \in \mathbb{R}_+^{d\times n}$ is a data matrix with d features and n samples. Such models have close relation to K-means clustering as pointed out in [Ding *et al.*, 2005]. $F \in \mathbb{R}_+^{d\times c}$ can be viewed as cluster centroids, and $G \in \mathbb{R}_+^{n\times c}$ can be viewed as clustering indicator matrix.

Many previous NMF work [Ding *et al.*, 2010; 2006; Lee and Seung, 2001; Kong *et al.*, 2011] use a soft indicator matrix G, *i.e.*, elements in G are continuous values. Algorithms for those models are based on matrix multiplicative updating rules. However, these algorithms converge very slowly, often need hundreds or even thousands of iterations before convergence. Considering the computational cost of large matrix multiplication, such algorithms usually take a long time to converge. In addition, an additional post-processing step need to be performed to get the final clustering labels using the soft indicator matrix G.

On the other hand, outliers are very common in real world data problems. For example, in image recognition problems, there maybe different extent of noise due to different conditions of illuminations, viewpoints, wearings, *etc*. Also, human measurement and recording errors may also incur many outliers. Previous NMF models using Frobenius norm as loss measurement can not get good clustering results if there exist outliers/noise in the data. This is because the loss term will be dominated by outliers, thus the loss of other normal data points will be disregarded. We will show this using a synthetic data set in the experimental section. Therefore, developing robust models to handle data with outliers is important.

To address the above challenges, in this paper, we propose three new fast robust NMF and NMTF models.

2.2 New Fast Robust NMF and NMTF Models

To make the NMF/NMTF models robust to outliers, instead of using Frobenius norm, we propose to use the $\ell_{2,1}$-norm and ℓ_1-norm as loss measurements. The new robust and fast NMF models aim to minimize the following objective functions:

$$\min_{F\geq 0, G\in Ind}\left\|X - FG^T\right\|_1 \quad (5)$$

$$\min_{F\geq 0, G\in Ind}\left\|X - FG^T\right\|_{2,1} \quad (6)$$

$$\min_{F\in Ind, G\in Ind, S\geq 0}\left\|X - FSG^T\right\|_1 \quad (7)$$

where $G \in Ind$ or $F \in Ind$ indicates that G and F are indicator matrices, *i.e.* $g_{ij} = 1$ if x_i belongs to class j, and $g_{ij} = 0$ otherwise. There is only one element can be non-zero in each row of binary indicator matrix. Note that in this way, the constraint $G^TG = I$ in orthogonal NMF [Ding *et al.*, 2006] is automatically satisfied. In addition, clustering labels are directly obtained without the need for further post-processing as in previous NMF works. While $F \geq 0$ indicates that F is a non-negative matrix with all elements greater or equal to zero.

In the tri-factorization model (7), since both F and G are restricted to be binary indicators, S is introduced to absorb the magnitude in the original data matrix X. Since our methods are robust and fast NMF/NMTF models, we call the three models as RFNMF_L1, RFNMF, and RFNMTF, respectively.

$\ell_{2,1}$-norm and ℓ_1-norm are often used for enforcing sparsity when applied to regularize parameter matrix [Nie *et al.*, 2010; Wang *et al.*, 2014a; 2013], and achieving robustness to outliers when applied to loss function [Kong *et al.*, 2011].

Laplacian Noise Interpretation for RFNMF_L1: We show the probabilistic motivation of our model from Laplacian noise distribution. Suppose x_i is an observed data point contaminated by an additional noise σ_i:

$$x_i = \alpha_i + \sigma_i \quad (8)$$

where α_i is the unobservable true data, in NMF clustering it is the clustering centroid, *i.e.* $\alpha_i = FG_{i.}^T, G_{i.} \in Ind.$ σ_i is the noise. Suppose the noise is drawn from Laplacian distribution with zero mean, then we have:

$$p(x_i|\alpha_i) = \frac{1}{2b}exp(-\frac{\|x_i - \alpha_i\|_1}{b}) \quad (9)$$

where b is the scale parameter of Laplacian distribution. Suppose we have an observed data set $X = [x_1, x_2, ..., x_N]$, the maximum likelihood estimate (MLE) of α_i should be:

$$\max_{\alpha_i} log \prod_{i=1}^{N} p(x_i|\alpha_i) \Rightarrow \max_{\alpha_i} -\tfrac{1}{b}\sum_{i=1}^{N}\|x_i - \alpha_i\|_1$$

$$\Rightarrow \quad \min_{\alpha_i}\sum_{i=1}^{N}\|x_i - \alpha_i\|_1 \Rightarrow \min_{F,G_{i.}\in Ind}\sum_{i=1}^{N}\|x_i - FG_{i.}\|_1$$

$$\Rightarrow \quad \min_{F,G\in Ind}\sum_{i=1}^{N}\|X - FG\|_1 \quad (10)$$

Therefore, the MLE under Laplacian noise assumption is equivalent to the RFNMF_L1 model. In the algorithm section, we will show that the solution of cluster centroids F is exactly finding the sample medians, which coincides with the MLE of the location parameter of Laplacian distribution.

Similar to the Laplacian noise interpretation of RFNMF_L1, it is also not difficult to see that: the RFNMTF

model can also be interpreted from a Laplacian distributed noise, except α is replaced by $FSG_{i.}$ instead of $FG_{i.}$; the RFNMF model can be interpreted from a Gaussian distributed noise. For space reason, we do not show the detail derivation.

The constraints in the objective functions involves combinatorial search over the solution space of F and G, which is very hard to optimize. Previous works always solve the relaxed version to make G have continuous values. In the following section, efficient algorithms are proposed for solving the objective functions with binary indicator constraints.

3 Optimization Algorithms

We first introduce a lemma which will be used in the following optimization algorithms.

Lemma 1. *Considering the following objective function:*

$$\min_{z} \sum_{i} |z - a_i| \tag{11}$$

The optimal solution of z is the median value of a_i.

This lemma can be easily obtained by setting the derivative of Eq. 11 to zero:

$$\sum_{i} sgn(z - a_i) = 0 \tag{12}$$

where $sgn(x) = 1$ if $x > 0$, $sgn(x) = -1$ if $x < 0$, and $sgn(x) = 0$ if $x = 0$. This condition can be satisfied if and only if z takes the median value of a_i.

3.1 Efficient Algorithm for Problem (5)

Given an initial guess of F and G, we iteratively update the model parameters. When G is fixed, problem (5) becomes:

$$\min_{F \geq 0} \left\| X - FG^T \right\|_1 \tag{13}$$

Since ℓ_1-norm can be decoupled through rows and columns, the above problem can be reduced to:

$$\begin{aligned} &\Rightarrow \quad \min_{F \geq 0} \sum_{i} \left\| X_{i.} - \sum_{k} F_{ik} G_{.k}^T \right\|_1 \\ &\Rightarrow \quad \min_{F \geq 0} \sum_{i} \sum_{k} \sum_{G_{jk}=1} |X_{ij} - F_{ik}| \end{aligned}$$

The above problem can be decoupled as: for $\forall i, k$, solving:

$$\min_{F_{ik}} \sum_{G_{jk}=1} |X_{ij} - F_{ik}| \tag{14}$$

According to Lemma 1, the optimal solution of F_{iK} can be efficiently obtained by finding the median values of samples belong to the k-th cluster.

When F is fixed, G can be efficiently optimized by assigning label to the cluster with nearest centroid, *i.e.*:

$$g_{ij} = \begin{cases} 1 & j = arg\min_{k} \|X_{.i} - F_{.k}\|_1 \\ 0 & otherwise \end{cases} \tag{15}$$

3.2 Efficient Algorithm for Problem (6)

When G is fixed, using reweighted method, problem (6) can be reduced to:

$$\min_{F} Tr((X - FG^T)D(X - FG^T)^T) \tag{16}$$

where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements are:

$$D_{ii} = \frac{1}{2\left\| X_{.i} - FG_{i.}^T \right\|} \tag{17}$$

Setting the derivative *w.r.t.* F to zero, we get:

$$F = XDG(G^T DG)^{-1} \tag{18}$$

When F is fixed, G can be obtained by assigning labels to the cluster with the nearest centroid:

$$g_{ij} = \begin{cases} 1 & j = arg\min_{k} \|X_{.i} - F_{.k}\|_2 \\ 0 & otherwise \end{cases} \tag{19}$$

3.3 Efficient Algorithm for Problem (7)

When F and G are fixed, problem Eq. (7) becomes:

$$\begin{aligned} &\Rightarrow \quad \min_{S} \left\| X - FSG^T \right\|_1 \\ &\Rightarrow \quad \min_{S} \left\| \sum_{k,l} S_{kl} F_{.k} G_{.l}^T - X \right\|_1 \\ &\Rightarrow \quad \min_{S} \sum_{k,l} \sum_{F_{ik}=1, G_{jl}=1} |S_{kl} - X_{ij}| \end{aligned}$$

This problem can be decoupled as: $\forall k, l$, solving:

$$\Rightarrow \quad \min_{S_{kl}} \sum_{F_{ik}=1, G_{jl}=1} |S_{kl} - X_{ij}| \tag{20}$$

According to Lemma 1, S can be efficiently optimized by finding the median values.

When S and F are fixed, G can be obtained by:

$$g_{ij} = \begin{cases} 1 & j = arg\min_{k} \|X_{.i} - FS_{.k}\|_1 \\ 0 & otherwise \end{cases} \tag{21}$$

When S and G are fixed, F can be obtained by:

$$f_{ij} = \begin{cases} 1 & j = arg\min_{k} \left\| X_{i.} - S_{k.} G^T \right\|_1 \\ 0 & otherwise \end{cases} \tag{22}$$

Therefore, all the three models can be efficiently solved by iteratively update F and G.

Unlike previous works which need to initialize G using the clustering results of K-means, our algorithms works good with random initialization of G. In practice, we can repeat the algorithm many times using different random initializations, and then take the result with the minimum objective value.

Note that our algorithms use binary indicator matrix instead of soft indicator matrix. Therefore, the clustering labels can be obtained directly by assigning labels to the cluster with nearest centroid. Therefore, our algorithms avoid the intensive computation of matrix multiplication. More importantly, previous NMF/NMTF works converges slowly, often needs hundreds or even thousands iterations before convergence. Our algorithms converge very fast, usually in just tens of iterations, and take much less computational time.

3.4 Convergence Analysis

In this section, we prove the convergence of our algorithms. To begin with, following are some simple definitions that will be used.

$$J(f, f') = \sum_i |x_i - f'_{\mathfrak{s}_i(f)}| \tag{23}$$

where x_i is the i-th data, f is the cluster centroid of the previous iteration, f' is the new cluster centroid, $\mathfrak{s}_i(f)$ denotes the clustering assignment of the i-th data using the cluster centroid of previous step. Obviously, when $f = f' = f_t$, $E(f_t) = J(f_t, f_t)$ is exactly the objective value of RFNMF_L1 in Eq. (5) in the t-th iteration.

According to Lemma 1: the optimum solution of f' is the median value of samples, and we have $J(f, f') \leq J(f, f)$.

In addition, since by definition $\mathfrak{s}_i(f')$ is the best cluster assignment using cluster centroid f', we have:

$$\begin{aligned} & E(f') - J(f, f') \\ = \; & J(f', f') - J(f, f') \\ = \; & \sum_i |x_i - f'_{\mathfrak{s}_i(f')}| - \sum_i |x_i - f'_{\mathfrak{s}_i(f)}| \leq 0 \end{aligned} \tag{24}$$

So we have:

$$\begin{aligned} & E(f') - E(f) \\ = \; & E(f') - J(f, f') + J(f, f') - J(f, f) \\ \leq \; & J(f, f') - J(f, f) \leq 0 \end{aligned} \tag{25}$$

Therefore, our algorithm for solving problem (5) monotonically decreases the objective value at each iteration until it reaches the optimum solution where $f'_* = f_*$. In addition, the KKT condition is satisfied. So the algorithm converges to a local minimum.

The convergence proof of the other two algorithms follows a similar procedure. For space reasons, we omit the detailed proof.

4 Experimental Results

4.1 Experiment Settings

In order to validate the effectiveness of the proposed NMF and NMTF methods, we compare our methods with some related methods as following:

(1) Standard NMF (NMF) [Lee and Seung, 2001]: solves the objective function in Eq. (4).

(2) Orthogonal NMTF (OrthNMF) [Ding *et al.*, 2006]: factorizes a matrix into three non-negative components, and each column of the soft indicator matrices (F and G) are required to be orthogonal.

(3) SemiNMF [Ding *et al.*, 2010]: allows the basis matrix F in standard NMF to be mix-signed

(4) Convex NMF (ConvNMF) [Ding *et al.*, 2010]: restricts the basis matrix F into a linear combination of original data points.

(5) Robust NMF (RNMF) [Kong *et al.*, 2011]: replaces the loss measurement in standard NMF from Frobenius norm to $\ell_{2,1}$-norm, which makes the model robust to outliers. The difference between our methods and RNMF is that: our methods use a hard indicator, which is more difficult to solve, but has clearer clustering interpretation since they reduce to directly assign cluster labels to samples at each updates of indicator matrix. This approach is better for clustering tasks. In addition, the algorithm is also much faster as we will show empirically in the experiment section. All the above models are solved using a multiplicative updating rules, which converges slowly and involves intensive matrix computations.

Table 1: Average distance from the centroids for normal data (blue and red points in figure 1 (a)), outliers, and all data.

	normal data	outliers	all data
NMF	6.02	10.82	6.09
Our methods	1.27	12.96	1.45

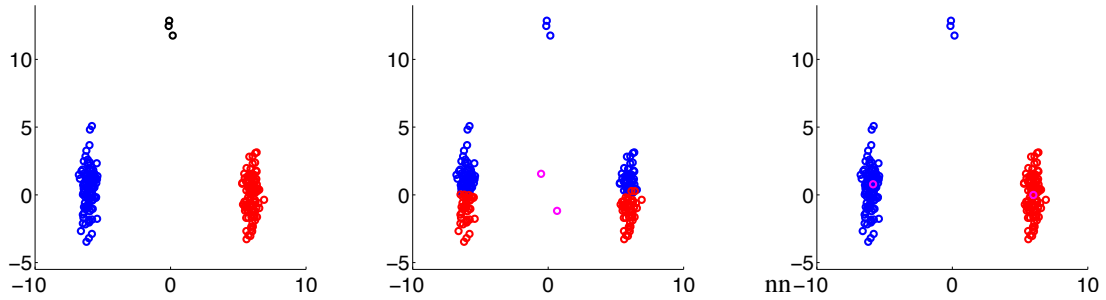
K-means serves as a baseline in the clustering performance comparison. The above comparison methods need to be initialized with the clustering results of K-means, since the objective functions are non-convex, and the final clustering results are sensitive to initializations. A little perturbation is needed for preventing these NMF models from sticking at the K-means solutions. As suggested in previous researches [Kong *et al.*, 2011; Ding *et al.*, 2006], the initialization of indicator matrix is set as $G0 = G_k + 0.2$, where G_k is the result of the K-means algorithm. The solution of these comparison methods is then get by iteratively update F and G using the multiplicative rule in Table 1 in [Li and Ding, 2006]. The iteration number is set as 500.

As for our methods, the initialization is set as the following: for RFNMF and RFNMF_L1, rather than using the result of K-means, we just randomly initialize the cluster indicator matrix G in case of the model stuck at the solution of K-means. We can see in the experiments that our objective converges good on a local minimum even with random initialization. For RFNMTF, the initialization of G and F is using the clustering results of the data dimension and feature dimension, similar to the strategy used in traditional NMF methods.

4.2 Experiments on Synthetic Data

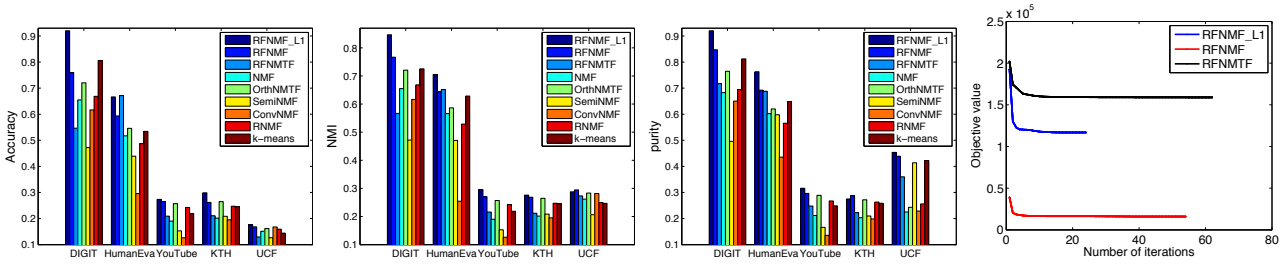
We compare our methods with standard NMF on a synthetic data set with outliers. As shown in Figure 1 (a), the data set contains two normal clusters (blue points and red points) drawn from two gaussian distributions with mean (-6, 0) and (6,0), respectively. Each cluster contains 100 data points. Three outliers (black points) were added far away from these two clusters. Figure 1 (b) shows the clustering results using standard NMF algorithm. Because standard NMF is not robust to outliers, the original blue/red points in one cluster are split into two clusters. Other comparison methods follow the same pattern as standard NMF, *i.e.* split the original cluster into two clusters. Figure 1 (c) shows the clustering results using the proposed three methods. Our methods are robust to outliers, therefore, the correct clustering structure is recovered.

Table 1 reports the average distance of data points from the corresponding cluster centroids. We can see that the distance of standard NMF algorithm for outliers is a little smaller than



(a) Original data with outliers (b) Clustering results with standard NMF (c) Clustering results with our three methods

Figure 1: Clustering performance on synthetic data. Blue points and red points are normal data drawn from two gaussian distributions. Black points are outliers. Magenta points are computed cluster centroids.



(a) Accuracy (b) NMI (c) purity (d) DIGIT

Figure 2: (a)-(c): Clustering performance comparison (d): objective value versus number of iterations.

Table 2: Computational time (in seconds) comparison. Averaged over 10 repetitions.

	RFNMF_L1	RFNMF	RFNMTF	NMF	OrthNMTF	SemiNMF	ConvNMF	RNMF	K-means
DIGIT	**3.57**	5.98	12.13	54.89	243.27	46.98	55.47	264.22	**1.72**
HumanEva	**5.79**	49.32	23.63	67.64	2181.26	12.60	775.23	1626.69	**2.23**
YouTube	**3.70**	**7.86**	20.32	366.91	553.55	43.89	37.68	203.30	28.25
KTH	**8.04**	**12.51**	26.41	300.72	676.98	64.71	65.04	499.08	27.83
UCF	**246.19**	251.20	278.06	1974.57	3891.82	349.75	639.00	1580.65	**163.20**

our methods, since it pays more attention to outliers. The distances of our methods for normal data and for all data are much more smaller than standard NMF.

4.3 Data Set Descriptions

In this section, we will present empirical results to evaluate the proposed NMF and NMTF approaches. Five real world data sets are used to evaluate the effectiveness of our method.

Digit data set is a public data set hosted in UCI Machine Learning Repository [1]. This data set consists of handwritten digits from 0 to 9, and each digit is a class.

HumanEva data set [2] contains 10000 samples from two subjects with 5000 samples per subject. There are five types of motions: boxing, walking, throw-catch, jogging and gesturing.

YouTube data set [3] contains 11 human activities: soccer juggling, swinging, tennis swinging, to name a few. Figure 3 show some sample images with bounding action labels in YouTube data set.

UCF data set [4] is an extension to the YouTube data sets. It contains 50 actions consisting of realistic videos in YouTube. This data set is very hard to recognize due to large variations in camera position, pose, illumination conditions, viewpoint, and cluttered background.

KTH data set [5] contains six type of human actions: walking, jogging, running, boxing, hand waving and hand clapping. These actions are performed by 25 subjects in four different scenarios: outdoor, outdoor with scale variations, outdoor with different clothes, and indoors. All samples are taken in homogeneous background, so that it is more close to natural environments. This setting add noise in the sample, thus, it is more difficult to recognize the actions.

[1] https://archive.ics.uci.edu/ml/datasets/Multiple+Features
[2] http://vision.cs.brown.edu/humaneva/index.html
[3] http://www.cs.ucf.edu/~liujg/YouTube_Action_dataset.html
[4] http://crcv.ucf.edu/data/UCF50.php
[5] http://www.nada.kth.se/cvap/actions/

Figure 3: Sample images from the website of the YouTube data set.

4.4 Clustering Performance and Convergence Speed Comparison

We use accuracy, NMI (normalized mutual information), and Purity as measurements of the clustering performance. Since all the methods are sensitive to initializations, we repeat each algorithm 10 times, and the clustering results with the minimum objective value is recorded in Figure 2. We can see that: RFNMF_L1 and RFNMF achieve the best performance on most data sets in terms of both accuracy, NMI and purity. The performance of OrthNMTF and RNMF is also pretty good on some data sets.

Figure 2 (d) shows the convergence of the proposed algorithms. For space limitation, only one of the five data sets are shown in the paper. We can see that all the three algorithms converge fast, usually in about 50 iterations.

Table 2 shows the comparison of computational time. The algorithms are run on a Dell desktop with double i7 Cores and 16GB memory. For RFNMF and RFNMF_L1, we just use random initialization in the experiments. For all of the other methods, K-means result is used as initialization since it is suggested by previous research. The time consumption of K-means initialization is not considered for all of these methods. We can see that the RFNMF_L1 algorithm converges very fast, which is often comparable or even faster than K-means. Thus, our algorithm is scalable to large data. The computational time of RFNMF and RFNMTF is also much less than other compared NMF and NMTF methods.

5 Conclusions

We proposed three new NMF and NMTF models which are robust to outliers to improve the human action clustering tasks. Efficient algorithms are derived, which converge as fast as the standard K-means algorithm, and thus are scalable to large-scale data sets. Experimental results on both synthetic and real world data sets show that our methods outperform other existing NMF and NMTF methods in most cases, and in the meanwhile, take much less computational time.

References

[Ding *et al.*, 2005] C. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, pages 606–610, 2005.

[Ding *et al.*, 2006] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

[Ding *et al.*, 2010] Chris Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010.

[Gao *et al.*, 2015] Hongchang Gao, Feiping Nie, Weidong Cai, and Heng Huang. Robust capped norm nonnegative matrix factorization. *24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, pages 871–880, 2015.

[Gu and Zhou, 2009] Quanquan Gu and Jie Zhou. Neighborhood preserving nonnegative matrix factorization. In *Proceedings of the British machine vision conference*, 2009.

[Gu *et al.*, 2011] Quanquan Gu, Chris Ding, and Jiawei Han. On trivial solution and scale transfer problems in graph regularized nmf. In *Proceedings of the international joint conference on artificial intelligence*, 2011.

[Huang *et al.*, 2013] Jin Huang, Feiping Nie, Heng Huang, Yicheng Tu, and Yu Lei. Social trust prediction using heterogeneous networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(4):17:1–17:21, 2013.

[Kong *et al.*, 2011] Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using l2,1-norm. In *ACM Conference on Information and Knowledge Management (CIKM 2011)*, 2011.

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

[Lee and Seung, 2001] D.D. Lee and H.S. Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562, 2001.

[Li and Ding, 2006] Tao Li and Chris Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 362–371. IEEE, 2006.

[Nie *et al.*, 2010] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. *Advances in Neural Information Processing Systems*, 23:1813–1821, 2010.

[Wang *et al.*, 2011] Hua Wang, Heng Huang, and Chris Ding. Cross-language web page classification via joint nonnegative matrix tri-factorization based dyadic knowledge transfer. In *Annual ACM SIGIR Conference 2011*, pages 933–942, 2011.

[Wang *et al.*, 2012] Hua Wang, Heng Huang, Chris Ding, and Feiping Nie. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *The 16th International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 314–325, 2012.

[Wang *et al.*, 2013] De Wang, Feiping Nie, Heng Huang, Jingwen Yan, Shannon L Risacher, Andrew J Saykin, and Li Shen. Structural brain network constrained neuroimaging marker identification for predicting cognitive functions. In *Information Processing in Medical Imaging*, pages 536–547. Springer Berlin Heidelberg, 2013.

[Wang *et al.*, 2014a] De Wang, Feiping Nie, and Heng Huang. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In *Machine Learning and Knowledge Discovery in Databases*, 2014.

[Wang *et al.*, 2014b] Hua Wang, Heng Huang, and Chris Ding. Correlated protein function prediction via maximization of data-knowledge consistency. *The 18th International Conference on Research in Computational Molecular Biology (RECOMB 2014)*, pages 311–325, 2014.

[Wang *et al.*, 2015] Hua Wang, Feiping Nie, and Heng Huang. Large-scale cross-language web page classification via dual knowledge transfer using fast nonnegative matrix tri-factorization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1), 2015.