ReD-SFA: Relation Discovery Based Slow Feature Analysis for Trajectory Clustering

Zhang Zhang, Kaiqi Huang, Tieniu Tan, Peipei Yang
CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences
zzhang@nlpr.ia.ac.cn, kqhuang@nlpr.ia.ac.cn, tnt@nlpr.ia.ac.cn, ppyang@nlpr.ia.ac.cn

Jun Li
Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney
Jun.Li@uts.edu.au

Abstract

*For spectral embedding/clustering, it is still an open problem on how to construct an relation graph to reflect the intrinsic structures in data. In this paper, we proposed an approach, named **Re**lation **D**iscovery based **S**low **F**eature **A**nalysis (ReD-SFA), for feature learning and graph construction simultaneously. Given an initial graph with only a few nearest but most reliable pairwise relations, new reliable relations are discovered by an assumption of reliability preservation, i.e., the reliable relations will preserve their reliabilities in the learnt projection subspace. We formulate the idea as a cross entropy (CE) minimization problem to reduce the discrepancy between two Bernoulli distributions parameterized by the updated distances and the existing relation graph respectively. Furthermore, to overcome the imbalanced distribution of samples, a Boosting-like strategy is proposed to balance the discovered relations over all clusters. To evaluate the proposed method, extensive experiments are performed with various trajectory clustering tasks, including motion segmentation, time series clustering and crowd detection. The results demonstrate that ReD-SFA can discover reliable intra-cluster relations with high precision, and competitive clustering performance can be achieved in comparison with state-of-the-art.*

1. Introduction

Trajectory clustering provides benefits for many vision tasks, such as motion segmentation [24], object detection [4], action recognition [31] and scene modeling [32]. As trajectories often lie in a low dimensional subspace, spectral methods have been widely used, where trajectories will be firstly embedded into a low dimensional feature space before clustering. However, it is still an open problem to construct an appropriate relation graph by which the intrinsic structures in data can be well encoded.

With original (x, y) sequence representation of trajectories, the most straightforward way to construct relation graph is to calculate the distance (e.g., Euclidian distance) between all pairs of samples, and remain a number of pairwise relations with the smallest distances through K nearest neighbor (K-NN) criterion or ε-ball neighbor criterion. However, it is difficult to capture the true reliable relations by setting a constant k or ε for all samples and different datasets. A large k or ε may induce wrong relations due to the high dimensionality and variant distribution densities of samples, while a small one may miss true relations.

From another aspect, if efficient features can be extracted in advance, it will be much easier to construct reliable relation graph. For trajectory clustering, some hand-crafted features, such as velocity [8] and high-order derivatives [14], have been adopted before graph construction. Thus, it may be better to do feature learning and relation discovery jointly, instead of the one-step graph construction.

In this paper, given a small number of initial reliable relations which can be obtained by selecting the nearest neighborhood relations in original space X, we will discover new relations with ε-ball neighbor criterion in the low dimensional feature space Y, where the ε can be estimated by an assumption of reliability preservation, i.e., the reliable relations will preserve their reliabilities in the feature space.

The process is illustrated in Fig.1. The left two figures show the data samples of two clusters (denoted by blue and green stars) in original space X, and the dash line indicates the projection direction learnt by the current relations (red links). In projection space Y (right figures), some pairwise relations are enhanced greatly, as their distances in feature space Y are even smaller than the distances of some initial reliable relations. Thus, the enhanced relations can be discovered as new reliable relations based on which the projection direction can be further corrected (left bottom). With such progress, the samples of intra-cluster will be concen-

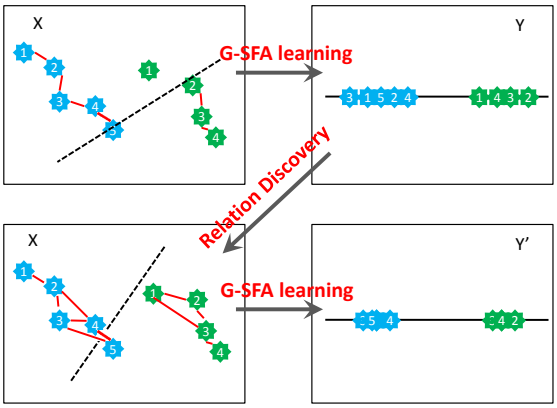


Figure 1. Illustration of the learning process of ReD-SFA.

trated more closely in the new space Y' (right bottom).

In this work, we proposed ReD-SFA to jointly construct relation graphs and learn features, where Slow Feature Analysis (SFA) [34] is chosen as the basic feature learning algorithm. In summary, the efforts of this paper include:

- The idea of joint graph construction and feature learning is formulated as a problem of cross entropy minimization, which aims to reduce the discrepancy between two Bernoulli distributions parameterized by the updated distances and the relation graph respectively.
- For efficient optimization, the original objective function is relaxed to its upper bound, so that the solution is well compatible with the GSFA. Furthermore, all parameters can be selected automatically.
- Furthermore, to overcome the problem of imbalance distribution of data samples, a Boosting-like strategy is proposed to improve the balance of discovered relations over different clusters.
- Extensive experiments are performed with various clustering tasks, e.g., motion segmentation, time series clustering and crowd detection. The competitive performance demonstrates the effectiveness of ReD-SFA.

2. Related Work

For spectral based trajectory clustering, the primary problem is how to construct the affinity graph to measure the pairwise relations between trajectories. In previous work, there are mainly two kinds of methods on the problem: similarity based methods and learning based methods.

For similarity based methods [4] [21], the similarities between all pairs of trajectories need to be calculated based on original trajectories or other trajectory features. Various hand-crafted features, e.g., velocity [4] [8] [38], high order motion models [14] [21], and Principle Components Analysis (PCA) coefficients [1], are proposed. Another important issue is to design a good similarity metric, especially to overcome the problem of trajectories with varying lengths and misalignment. The commonly used similarity metrics include Euclidean distance [1], Hausdorff distance [15], Dynamic Time Warping (DTW)[18], and Edit Distance on Real sequence (EDR) [7], etc. A comparison on different similarity metrics can be found in [36].

For learning based methods, the subspace segmentation (clustering) [35] [11] [20] [24] have achieved impressive performance for point trajectory clustering. In their work, one step of subspace recovery is performed firstly, where each trajectory is reconstructed by other samples based on *self-expressiveness property*[11]. Then, the sparse coefficients are used to build the affinity matrix for subsequent spectral clustering. The main challenge is how to handle the noises in trajectories [20]. For this problem, variant algebraic regularization terms, such as affine projection [35], agglomerative lossy compression (ALC)[24], low rank [20], sparse [11], and epipolar constraint [17],has been proposed. Recently, Wang et.al, [33] combined sparse and low rank constraint to promote the robustness of subspace clustering.

Our work is close to the manifold clustering methods, such as LLMC[16] and SMCE[10], which try to find relations and a projection to a lower-dimensional space of data. Different with the work which adopted local linear assumption to learn neighbor weights in original space, we discover reliable relations in feature space iteratively. Our work is also related to information theoretic based subspace models [9][27] which adopted mutual information metric to measure the divergence between class labels and transformed features. While, this work focuses on unsupervised learning with a problem of cross entropy minimization.

In this work, SFA is adopted to learn trajectory features. SFA has been used for unsupervised invariance learning in visual neural cell modeling [34] [3] and feature extraction in pattern recognition, e.g., digit recognition [2], scene classification [26], behavior recognition [37] [25]. In [12] the SFA is generalized from temporal sequences to relation graphs for supervised dimensionality reduction. Here, we will extend the GSFA to unsupervised clustering based on joint relation discovery and feature learning.

3. ReD-SFA

3.1. Problem definition and initialization

Given a collection of data points $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ for $\mathbf{x}_i \in \mathbb{R}^d$, ReD-SFA will learn a projection matrix $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_c]$ and construct a relation graph composed of a Boolean matrix $\Delta = \{\delta_{ij}\}$ and a real value matrix $\mathbf{W} = \{\mathbf{w}_{ij}\}$ simultaneously, so that $\mathbf{P}^\top \mathbf{x}_i$ $(\in \mathbb{R}^c)$ is the intrinsic low-dim representation of $\mathbf{x}_i$, $\delta_{ij} \in \{1, 0\}$ indicates whether the relation of the ith sample and the jth one can be measured by $\mathbf{l}_{ij} = ||\mathbf{P}^\top \mathbf{x}_i - \mathbf{P}^\top \mathbf{x}_j||_2$ reliably or not, and $\mathbf{w}_{ij} \in [0, 1]$ denotes the confidence degree of δ_{ij}.

At initialization, only the first m nearest relations in original space are set to 1 in Δ^0, otherwise 0. $\mathbf{W}^0$ is set as follows, while the superscript 0 denotes the initial stage,

$$\mathbf{w}_{ij}^0 = \begin{cases} e^{-{\mathbf{l}_{ij}^0}^2/\mathbf{t}}, & \delta_{ij}^0 = 1, \\ 1 - e^{-{\mathbf{l}_{ij}^0}^2/\mathbf{t}}, & \delta_{ij}^0 = 0. \end{cases} \tag{1}$$

where $\mathbf{l}_{ij}^0 = ||\mathbf{x}_i - \mathbf{x}_j||_2$ and $\mathbf{t}$ is a scale factor.

3.2. Objective function

Assuming a Boolean variable $\mathbf{r}_{ij} \in \{1,0\}$ indicates the intra-cluster relation or not between $\mathbf{x}_i$ and $\mathbf{x}_j$, which can be modeled with a Bernoulli distribution described by $\mathbf{l}_{ij}$ in **feature space** as well as another Bernoulli distribution based on $\mathbf{w}_{ij}$ in **relation graph**, the ReD-SFA aims to reduce the discrepancy between the two distributions. Thus a cross entropy (CE) based objective function is adopted as follows.

$$\min \sum_{i,j,i\neq j}^{n} H\left(f_{\mathbf{w}_{ij}}, f_{\mathbf{l}_{ij}}\right) \tag{2}$$

The cross entropy $H\left(f_{\mathbf{w}_{ij}}, f_{\mathbf{l}_{ij}}\right)$ (abbreviated by H_{ij}) is calculated [19]:

$$H_{ij} = -\sum_{\mathbf{r}_{ij}\in\{1,0\}} f_{\mathbf{w}_{ij}}(\mathbf{r}_{ij}) \ln f_{\mathbf{l}_{ij}}(\mathbf{r}_{ij}), \tag{3}$$

where the Bernoulli distribution $f_{\mathbf{l}_{ij}}(\mathbf{r}_{ij})$ is defined as:

$$\begin{aligned} f_{\mathbf{l}_{ij}}(\mathbf{r}_{ij}) &= B_{\mathbf{l}_{ij}}{}^{\mathbf{r}_{ij}}(1-B_{\mathbf{l}_{ij}})^{1-\mathbf{r}_{ij}} \\ &= \left[\delta_{ij} Pr(\mathbf{r}_{ij}=1;\mathbf{l}_{ij},\delta_{ij}=1) + (1-\delta_{ij}) Pr(\mathbf{r}_{ij}=1;\mathbf{l}_{ij},\delta_{ij}=0)\right]^{\mathbf{r}_{ij}} \\ &\quad \left[\delta_{ij} Pr(\mathbf{r}_{ij}=0;\mathbf{l}_{ij},\delta_{ij}=1) + (1-\delta_{ij}) Pr(\mathbf{r}_{ij}=0;\mathbf{l}_{ij},\delta_{ij}=0)\right]^{1-\mathbf{r}_{ij}} \end{aligned} \tag{4}$$

We define $Pr(\mathbf{r}_{ij}=1;\mathbf{l}_{ij},\delta_{ij}=1) = \frac{1}{2} + \frac{1}{2}e^{-\mathbf{l}_{ij}^2}$ to be larger than $\frac{1}{2}$ with the assumption of *local consistency* [39], i.e.,nearby points are likely to have the same cluster label. Otherwise, $\delta_{ij}=0$ means the distance $\mathbf{l}_{ij}$ is too far to measure the reliability, thus $Pr(\mathbf{r}_{ij}=1;\mathbf{l}_{ij},\delta_{ij}=0)$ is decided by a threshold ε as $\frac{1}{2} - \frac{1}{2}e^{-\varepsilon^2}$ which is smaller than $\frac{1}{2}$.

Similarly, we can define $f_{\mathbf{w}_{ij}}(\mathbf{r}_{ij})$ with the type of Eq.4, where $Pr(\mathbf{r}_{ij}=1;\mathbf{w}_{ij},\delta_{ij}=1) = \frac{1}{2} + \frac{1}{2}\mathbf{w}_{ij}$ and $Pr(\mathbf{r}_{ij}=1;\mathbf{w}_{ij},\delta_{ij}=0) = \frac{1}{2} - \frac{1}{2}\mathbf{w}_{ij}$.

With the above definitions, Eq.2 can be summarized as a function of $\mathbf{P}$, Δ and $\mathbf{W}$. In the optimization, however, it is very complex to compute the derivative of $\mathbf{P}$ directly, due to the logarithmic term. For efficiency, an auxiliary variable $\eta < \frac{1}{2}$ is introduced, so that $\frac{1}{2} + \frac{1}{2}e^{-x^2}$ is approximated by $\frac{1}{2} + \left(\frac{1}{2} - \eta\right)e^{-x^2}$ and $\frac{1}{2} - \frac{1}{2}e^{-x^2}$ is replaced by $\frac{1}{2} - \left(\frac{1}{2} - \eta\right)e^{-x^2}$. Since $-\ln\left(\frac{1}{2} + \left(\frac{1}{2} - \eta\right)e^{-x^2}\right) \leq -\ln\left((1-\eta)e^{-x^2}\right)$ and $-\ln\left(\frac{1}{2} - \left(\frac{1}{2} - \eta\right)e^{-x^2}\right) \leq -\ln\left(\eta e^{-x^2}\right)$, H_{ij} in Eq.3 can be relaxed to the upper bound function J_{ij},

$$\begin{aligned} J_{ij} =& \delta_{ij}\mathbf{w}_{ij}||\mathbf{P}^\top\mathbf{x}_i - \mathbf{P}^\top\mathbf{x}_j||_2^2 + (1-\delta_{ij}\mathbf{w}_{ij})\varepsilon^2 \\ &+ \mathbf{w}_{ij}\ln\sqrt{\frac{\eta}{1-\eta}} - \ln\sqrt{\eta(1-\eta)}. \end{aligned} \tag{5}$$

Detailed derivations are shown in supplemental materials.

Additionally, one regularization term is added to avoid the divergence from initial $\mathbf{W}^0$, i.e., $J_{ij}^W = \left(\mathbf{w}_{ij} - \mathbf{w}_{ij}^0\right)^2$.

Thus, the final objective function is as follows:

$$\min_{\mathbf{P},\Delta,\mathbf{W}} \sum_{i,j,i\neq j}^{n} \left[\lambda J_{ij} + (1-\lambda) J_{ij}^W\right] \tag{6}$$

where λ is a trade-off parameter in $[0,1]$.

3.3. Optimization

Step 1: Given Δ and $\mathbf{W}$, Eq.6 can be transformed into

$$\min \sum_{i,j,i\neq j}^{n} \left[\lambda\delta_{ij}\mathbf{w}_{ij}\mathbf{Tr}\left(\left(\mathbf{P}^\top\mathbf{x}_i - \mathbf{P}^\top\mathbf{x}_j\right)\left(\mathbf{P}^\top\mathbf{x}_i - \mathbf{P}^\top\mathbf{x}_j\right)^\top\right)\right]. \tag{7}$$

As presented in GSFA [12], $\mathbf{P}$ can be optimized by solving the generalized eigenvalue problem,

$$A\mathbf{P} = \tau B\mathbf{P} \tag{8}$$

where the derivative covariance matrix A is calculated as:

$$A = \frac{1}{R}\sum_{i,j,i\neq j}^{n} \lambda\delta_{ij}\mathbf{w}_{ij}\left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \tag{9}$$

with $R = \sum_{i,j,i\neq j}^{n} \lambda\delta_{ij}\mathbf{w}_{ij}$. And the matrix B is as:

$$B = \frac{1}{V}\sum_{i=1}^{n} \lambda v_i\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^\top \tag{10}$$

where $V = \sum_{i=1}^{n} v_i$ with $v_i = \sum_{j=1}^{n} \lambda\delta_{ij}\mathbf{w}_{ij}$, and $\bar{\mathbf{x}}$ is the mean value of all data samples.

Step 2: Given $\mathbf{P}$ and $\mathbf{W}$, δ_{ij} can be simply updated with the following rule:

$$\delta_{ij} = \begin{cases} 1, & \mathbf{l}_{ij} \leq \varepsilon, \\ 0, & Otherwise. \end{cases} \tag{11}$$

where the ε should remain most initial relations $\delta_{ij}^0 = 1$ still reliable in the feature space. Meanwhile, it should be tight enough to avoid noise relations. The selection of ε will be discussed in the next section.

Step 3: Given Δ and $\mathbf{P}$, the optimization of $\mathbf{w}_{ij}$ can be obtained by calculating the first derivative for each $\mathbf{w}_{ij}$, then

let it equal to zero.

$$\mathbf{w}_{ij} = \begin{cases} \mathbf{w}_{ij}^0 + \frac{\lambda}{2(1-\lambda)} \left(\varepsilon^2 - \mathbf{l}_{ij}^2 + \ln \sqrt{\frac{1-\eta}{\eta}} \right), & \delta_{ij} = 1; \\ \mathbf{w}_{ij}^0 + \frac{\lambda}{2(1-\lambda)} \ln \sqrt{\frac{1-\eta}{\eta}}, & \delta_{ij} = 0. \end{cases} \tag{12}$$

Note, $\mathbf{w}_{ij}$ should be limited in the range of $[0,1]$.

3.4. Selection of parameters

3.4.1 Selection of ε

In this work, ε indicates the threshold of reliable distances in feature space. As shown in Eq.11, a pairwise relation will be considered to be reliable, if the distance is smaller than ε. According to the assumption of *reliability preservation*, ε is estimated as the percentile of a very high percentage (e.g. 98th percentile), given the distribution of reliable relations.

Here, we find the reliable relation set $\mathbf{L} = \{\mathbf{l}_{ij} | \delta_{ij} = 1\}$ can be well modeled by alpha-stable distribution. Figure 2(a) shows one example of the histogram of distances on a trajectory set in Hopkins 155 dataset [28]. The distribution has a typical heavy right tail which cannot be modeled by common normal distribution.

Thus, we firstly estimate the alpha-stable distribution with the set $\mathbf{L}$, after the optimization of $\mathbf{P}$ at each round. Then, ε is calculate as the percentile value of a given very high percentage q, i.e.,

$$\varepsilon = Percentile(q|\alpha,\beta,\gamma,\mu) \tag{13}$$

where, α, β, γ and μ are the four parameters in alpha-stable distribution. $\alpha \in (0,2]$ describes the tail of the distribution. $\beta \in [-1,1]$ is the skewness. The last two parameters are the scale $\gamma > 0$, and the location $\mu \in R$, which are similar to the variance and mean in normal distribution. Here, a toolbox [29] is used for the estimation of alpha-stable distribution.

3.4.2 Selection of λ and η

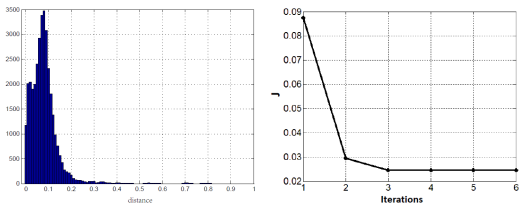
In Eq.6, λ controls the balance between J_{ij} and J_{ij}^W. And η controls the increment speed of $\mathbf{w}_{ij}$ in Eq.12. Here, instead of setting a constant value λ and η over all relations empirically, the two parameters are set for each relation individually. The priors come from the contextual information of a given sample pair. For the ith sample and the jth sample, as the increasing of the number of shared reliable neighborhoods, the relation $\mathbf{r}_{ij}$ is more likely to be reliable, so that it should favor more weight to J_{ij} and a higher increment speed of $\mathbf{w}_{ij}$. Based on the above idea, λ_{ij} and η_{ij} are determined as follows.

$$\lambda_{ij} = \frac{z_{ij}^s}{z_{ij}^t} e^{-\frac{\overline{z^t}}{z_{ij}^t}} \tag{14}$$

$$\eta_{ij} = \frac{1}{2} \left(1 - \frac{z_{ij}^s}{z_{ij}^t} \right) e^{-\frac{\overline{z^t}}{z_{ij}^t}} \tag{15}$$

where z_{ij}^s is the number of shared neighborhoods, i.e., $z_{ij}^s = \sum_{k,k\neq i,j} \delta_{ik}^0 \wedge \delta_{kj}^0$, and z_{ij}^t is the total number of reliable neighborhoods, i.e., $z_{ij}^t = \sum_{k,k\neq i,j} \delta_{ik}^0 \vee \delta_{kj}^0$, and $\overline{z^t}$ is the mean value over all pairs of samples. As $z_{ij}^t = 0$, a small value is set to λ_{ij} and η_{ij}, e.g., $\frac{1}{n}$.

Finally, the ReD-SFA is summarized in Alg.1. Figure 2(b) shows an instance on how the objective function changes with respect to the number of iterations. In our experiments, the ReD-SFA usually halts within 5 iterations.



(a) (b)

Figure 2. (a) An example to show the heavy-tailed distribution of L. (b)Objective function vs. the number of learning iterations.

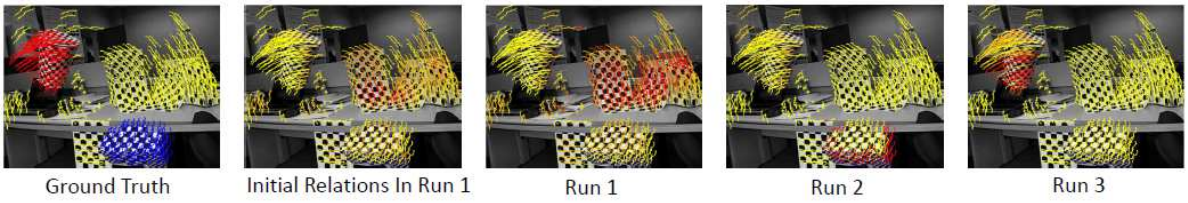
Algorithm 1 ReD-SFA Algorithm

Require: Δ^0, $\mathbf{W}^0$, $\mathbf{X}$
Ensure: $\mathbf{P}$, Δ, $\mathbf{W}$
1: $\Delta = \Delta^0$, $\mathbf{W} = \mathbf{W}^0$ and Calculate λ, η with Eq.14 and Eq.15
2: Calculate J with Eq.6
3: **while** 1 **do**
4: Update $\mathbf{P}$ with GSFA [33] to solve the problem in Eq. 7
5: Estimate ε with Eq.13
6: Update δ_{ij}, where $\delta_{ij} \in \Delta$ with Eq.11
7: Update $\mathbf{W}$ with Eq. 12
8: Calculate new J' with Eq.6
9: **if** $J < J'$ **then**
10: break;
11: **end if**
12: $J = J'$
13: **end while**

3.5. Boosting ReD-SFA

As presented above, ReD-SFA discovers reliable relations based on a few nearest neighborhoods. However, due to the imbalanced distribution of densities among different clusters, the initial reliable relations may be from only high-density clusters. In such case, the relations discovered by ReD-SFA will not cover all clusters.

To address this problem, we adopt the idea of boosting learning to run the ReD-SFA in several rounds to discover relations from all clusters. At each new round of ReD-SFA, the data samples uncovered in previous rounds are selected out to form a candidate set Φ. Then, the ReD-SFA is implemented only over the pairwise relations in Φ. Finally, all the discovered relations are gathered to form the final relation



	Total	True	C 1	C 2	C 3	Prec.	Rec.
Run 1	32122	32122	0	172	31950	100%	26.6%
Run 2	33702	33624	6	1648	31970	99.8%	27.8%
Run 3	34152	34074	456	1648	31970	99.8%	28.2%
Ground Truth		120868	2450	10506	107912	-	-

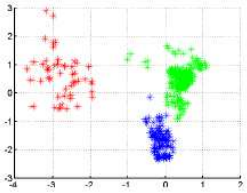


Figure 3. Illustration of Boosting-like ReD-SFA. The table shows the distributions of the discovered relations over different clusters in the three runs of ReD-SFA. The "ground truth" denotes the number of true intra-cluster relations in each clusters.

graph and the GSFA is performed again to obtain the final feature representation.

Figure 3 shows an example of the implementation process of boosting ReD-SFA, where the trajectories covered by the discovered relations in each run of ReD-SFA (the last three sub-figures) are labeled as red color. Besides, the statistics of the discovered relations in each round are presented in the table. The high precisions (Prec. = True/Total) indicate that most discovered relations are intra-cluster relations. The imbalance distributions of the relation discovery over the three clusters are also shown in the last row (Ground Truth), where the cluster "C1" has a much small proportion (only 2%) over the total intra-cluster relations, while the cluster "C3" is the most prevalent one. As shown in the figure, the first run of ReD-SFA concentrates on the cluster "C3", because the initial relations are mainly on this cluster (the second sub-figure from left), the second round is mainly for the cluster "C2", and so on. The multiple runs of ReD-SFA improve the balance of the discovered relations. The right bottom figure shows the final feature representation, where the well-separated clusters can be easily grouped by simple clustering algorithms, e.g., K-means.

4. Experimental Results

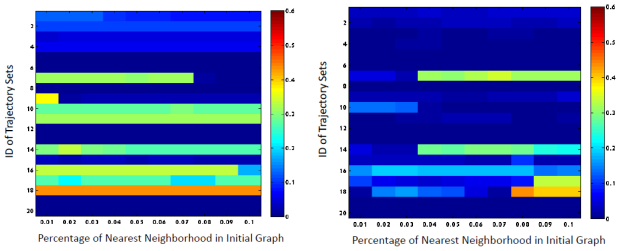
4.1. Motion segmentation on Hopkins 155 dataset

The Hopkins 155 dataset [28] has become a standard benchmark to evaluate trajectory clustering algorithms. The dataset provides 155 trajectories sets with 2 or 3 motion categories as well as the ground truth of all trajectories.

Both parameters of ReD-SFA are percentage type. One is m which denotes the first m nearest relations are selected in initial graph, the other is q indicating the percentile threshold ε for relation discovery. In this work, m ranges from 1% to 10%, q ranges from 92% to 99%. Furthermore, PCA is performed as preprocessing, like in SFA [34] and GSFA [12]. For all sets, the dimensionality of PCA is set to 9, and the final is further reduced to 3 after ReD-SFA.

To better understand the properties of the ReD-SFA on relation discovery and feature learning, we pick out 20 trajectory sets with three motions for a set of controlled tests. On this set, GSFA and ReD-SFA are performed with the same initial relation graph in terms of the first m nearest pairwise relations. Fig.4 shows the clustering errors of GSFA and ReD-SFA over the 20 sets along with different m values, where q is fixed to 95%. From the two figures, it is not surprised that the overall errors of ReD-SFA are much lower than that of GSFA, where the blue colors denote low errors (it is better shown in color). Furthermore, for GSFA, the necessary percentages of initial relations to obtain low errors are also variant. For example, the 7th set needs at least 8% to obtain a low error rate, and the 16th set needs above 10%. Thus, the sensitivity to m makes GSFA difficult to set a constant value over all sets. While for ReD-SFA, it achieves lower errors only using small m values with the helps of relation discovery.

We find on the 18th set, the clustering errors of GSFA over all m values are very high (around 40% displayed by orange color), while for ReD-SFA, the errors are very low (below 10%) with only small number of initial relations. Thus, we further analyze the discovered relations on the set. Based on the ground truth (labels) of trajectories, we check whether each relation discovered by the ReD-SFA is intra-



(a) GSFA (b) ReD-SFA

Figure 4. The error rates with variant initial parameters m over the 20 sets in Hopkins Dataset. This graphic should be seen in color.

	Total D. Num.	Prec./Rec.(TP Num.)	Rec. C1 (TP Num.)	Rec. C2 (TP Num.)	Rec. C3 (TP Num.)
Per. of Nearest Neighbors in Initial Graph: **1%**					
Initial	1232	100% / 2.6% (1232)	0.27% (8)	2.79% (2490)	2.73% (3562)
ReD-SFA	9478	99.68% / 19.91% (9448)	89.36% (2654)	11.78% (1855)	17.19% (4939)
Per. of Nearest Neighbors in Initial Graph: **3%**					
Initial	3696	100% / 7.79% (3696)	2.09% (62)	8.93% (1406)	7.75% (2228)
ReD-SFA	16946	95.82% / 34.22% (16238)	67.31% (1999)	43.71% (6885)	25.6% (7354)
Per. of Nearest Neighbors in Initial Graph: **5%**					
Initial	6160	100% / 12.98% (6160)	3.64% (108)	15.81% (2490)	12.4% (3562)
ReD-SFA	21473	94.74% / 42.87% (20344)	65.56% (1947)	56.37% (8879)	33.13% (9518)
Ground Truth		47450	2970	15750	28730

Table 1. Results of Statistical Analysis on Pairwise Relation Discovery in the 18th set "2RT3RCT_A".

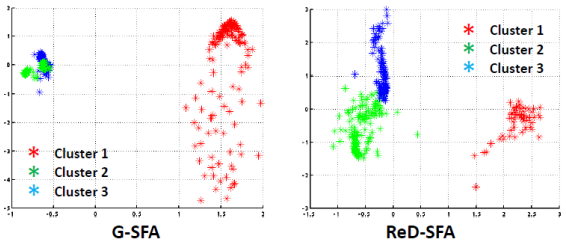


Figure 5. The features learnt by GSFA and ReD-SFA.

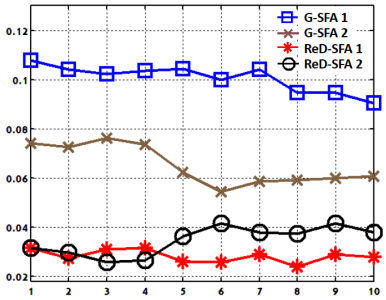


Figure 6. The average errors of the GSFA and the ReD-SFA over all 155 sets with different m values, when q is fixed to 95%.

cluster relation or not. The statistical results are presented in Table 1, where the precision (Prec.) and recall (Rec.) are defined as

$$Prec. = \frac{TP.Num.}{Total.D.Num}, \tag{16}$$

$$Rec. = \frac{TP.Num.}{GroundTruth}, \tag{17}$$

where $TP.Num$ means the number of true positives, $Total.D.Num$ is the total number of discovered relations. Besides the overall recall values, the recall values on the three clusters are also reported indivally. Since we do not distinguish the cluster labels of discovered relations, we only show the overall precisions.

From Table 1, we find that all initial relations are reliable intra-cluster relations in the three initial settings with $m = 1\%, 3\%, 5\%$ (Precision is 100%). However, the distribution of initial relations is very imbalance on the three clusters. The very little amount of relations in the 1st cluster (the recall is only 2.09% with $m = 3\%$) will make the features of this cluster disperse in a large scale after the GSFA learning, shown with red points in the left of Figure 5. It explains the high errors of GSFA While since the Boosting-like relation discovery, the recall values of all clusters can be improved greatly, especially on the 1st cluster (from 2.09% to 67.31% as $m = 3\%$), so that the features learnt by ReD-SFA are more centralized according to individual clusters (right in Fig. 5).

Figure 6 presents the average errors of GSFA and ReD-SFA over all 155 sets with different m and q values. Besides ε-ball initialization, denoted by *1*, we also test k-NN initialization denoted by *2*. From the figure, the performance of the GSFA with k-NN initialization is clearly better than ε-ball initialization, because the k-NN initialization provides more balanced distribution of initial relations. However, k-NN initialization trends to introduce noise relations (inter-cluster relations) as the increasing of m. The initial noise relations will lead to a larger ε in relation discovery, which may introduce more noise relations in later stage. Thus, the performance of ReD-SFA with k-NN initialization (black circle) is worse than that of ε-ball initialization (red star).

Compared to state-of-the-art methods, Table 2 shows the average errors over the entire dataset. The compared methods include GPCA [30], ALC [24], SSC [11], SCC [6], L-RR [20], RV [17], LLMC [16] and SMCE [10]. From the table, the recent work on the dataset has achieved very low errors (below to 1%). However, these results are achieved with specifical designs for 3D affine motion models or some additional pre/post-processing. Overall, ReD-SFA is still competitive with previous methods.

We also test the computational times on the Hopkins 155 dataset. On average, there are 296 trajectory samples per set. The average number of frames per trajectory is 29. The average time costs per trajectory of ReD-SFA as well as several classical approaches are shown in Table 1. The results of other approaches are taken from the work [20][13]. ReD-

	GPCA'05	ALC'10	SCC'09	SSC'13	LRR'13	RV'14	LLMC'07	SMCE'11	ReD-SFA
total	10.02%	3.37%	2.70%	2.41%	1.59%	0.77%	4.80%	3.25%	2.19%
three	28.66%	6.69%	5.89%	4.40%	-	1.88%	8.85%	7.03%	3.98%
two	4.59%	2.4%	1.77%	1.83%	-	0.44%	3.62%	2.15%	1.67%

Table 2. Clustering errors (%) on the entire Hopkins 155 dataset.

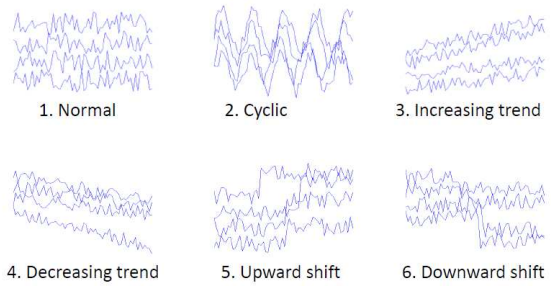


Figure 7. Some samples from the CC Dataset.

SFA is implemented on a PC with Intel i7 2.1GHz. We can see that the ReD-SFA achieves the lowest time cost because of the close form solution in ReD-SFA, while other methods needs more computing iterations in optimization.

Algorithm	SSC [11]	LRR [20]	ICLM[13]	ReD-SFA
Avg. time (s)	94	1.9	1.4	1.2

Table 3. Average time cost (seconds) for the Hopkins 155 dataset.

4.2. Time series clustering on CC Dataset

In this section, we will test ReD-SFA with a more general time series clustering task. The synthetic control charts (CC) dataset contains 600 1D samples [22] for time series classification and indexing. There are six classes: normal, cyclic, increasing trend, decreasing trend, upward shift, and downward shift. Each class has 100 samples, and the dimensionality of each sample is 60. Some samples are illustrated in Figure 7, where misalignment widely occurs in intra-class trajectories. We use the first 2, 3, 4, 5 and all 6 classes data to do clustering 5 times. For comparison, we implement a DTW based spectral clustering method [36], two subspace clustering algorithms, i.e., LRR [20] and SSC [11], and a manifold clustering algorithm, SMCE [10]. For ReD-SFA, LRR and SSC, PCA is firstly performed and the first 9 principle components are remained. For SMCE, we run the algorithm with the parameter setting of $k = 10$ (maximum neighbor size), $\lambda = 10$ (trade-off parameter).

Table 4 presents the errors of different clustering methods. The two subspace clustering methods cannot achieve satisfying results. LRR fails to recover correct subspaces in all clustering tasks. SSC is much better than LRR in 2 and 3 classes, however it still cannot achieve good performance, as more classes are involved. That is mainly because the assumption of global linear relationship among data, i.e., "*self-expressiveness*" property, widely adopted in subspace clustering is restrictive to the large misalignment. While, the SMCE and the ReD-SFA based on local neighbor relationships are more reliable in general case and thus obtain better performance. As more classes (5 and 6 classes), ReD-SFA achieves lower errors than SCME due to the high precision of relation discovery. DTW shows impressive clustering results, due to its superior capability to correct the misalignment. Compared to these methods, ReD-SFA achieves much better results than subspace learning based methods, and similar with DTW, which demonstrates the effectiveness of ReD-SFA on the general time series clustering task.

Num. C	LRR	SSC	SMCE	DTW	ReD-SFA
2	41.5%	1.5%	0%	0%	0%
3	49.7%	2.33%	0%	0%	0%
4	48.5%	38%	0%	0.3%	0%
5	51.6%	34.8%	28.4%	4%	6.6%
6	60.5%	44%	43.2%	24.7%	22.8%

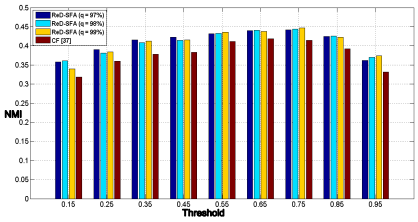
Table 4. Clustering errors on the CC dataset.

4.3. Crowd detection on CUHK Crowd Dataset

In this section, we test ReD-SFA with crow detection task in numerous and varied scenes. The CUHK crowd dataset [23] includes 474 crowd videos with various densities and perspective scales from many different surveillance environments. In [38] [23], coherent crowd motion patterns are detected by tracklets clustering, where tracklets from 300 video clips are manually annotated into groups for evaluation. Some samples are illustrated in the last column of Fig.9, where the tracklets indicated by the same color and marker form one crowd group and the outliers that do not belong to any groups are indicated by red plus.

For each frame, a set of tracklets with the same length is selected by a sliding window. Then, ReD-SFA is performed to discover the reliable intra-cluster relations among tracklets. Finally, a number of crowd groups are identified as the connected components in the relation matrix Δ, where the groups with size of less than 3 are deemed as outliers.

Here, to reduce the serious effects of outliers often occurring in real world scenes, we adopt a priori on crowd motion, *coherent neighbor invariance (CNI)* in [38], and restrict the relation discovery of ReD-SFA only from the coherent neighbors. With the prior, coherent neighbors a-



Methods	NMI
DTM [5]	0.30
HC [14]	0.27
CF [37]	0.42
CT [22]	0.48
ReD-SFA	0.45

Figure 8. Quantitative comparison on CUHK Dataset. The left figure shows the NMI with varying parameter settings. The right is the comparison results with other state-of-the-art methods.

mong tracklets are those spatial neighborhood relationships persevered in several consecutive frames. The details on *CNI* can refer to [38]. In their work, crowd groups are determined through coherent filtering (CF) which removes the unreliable relations with low velocity correlations in coherent neighbors. Different from CF [38] which explicitly calculates the velocity correlation distance with strong priors on crowd groups, ReD-SFA learns tracklet features automatically based on original x-y coordinates.

As the number of crowd groups is assumed unknown in [38] [23], Normalized Mutual Information (NMI) is adopted to measure the results quantitatively. The left of Fig.8 shows the comparison results of ReD-SFA ($q = 96\%$ and $c = 4$) and CF with varying thresholds of velocity correlations for CF, which are also used for the initialization of ReD-SFA. Additionally, the results of ReD-SFA without using CNI priori (blue square marker) are also presented. From the results, the ReD-SFA with CNI priori (red circle marker) achieves higher clustering results than the CF consistently at all settings. Compared to other methods, including mixture of dynamic texture (DTM) [5], hierarchical clustering (HC) [15], coherent filtering(CF) [38] and collective transition (CT) [23], the results reported in [23] are also presented at the right of Fig.8. The result of ReD-SFA is better than other methods except for the CT method which adopts an additional Markov chain model to refine the clustering results of CF with more dynamic information.

Some examples on the crowd detection results of CF and ReD-SFA are presented in Fig. 9. As shown in the figure, ReD-SFA obtain more accurate motion boundaries than CF. It is worthy noting: Firstly, in the ground truth, the members in the same group only need have a common goal and coherent motion direction, even they are moving at different locations with far spatial distance. However, the relations discovered by ReD-SFA encodes the spatial affinity based on original tracklet representation. Thus, ReD-SFA tends to obtain smaller crowd groups as spatial gaps exist between group members. In some cases, e.g., the 3rd row and the 4th row, such detection results are more consistent with common perception. Secondly, some outliers in ground truth (red plus) are not well labeled. In the last row, the tracklets located at the lady with white coat are mixed with her back person, and labeled as outliers. In such cases, ReD-SFA obtains more accurate segmentation results. In the first and second row of Fig.9, ReD-SFA merges some outlier points (red plus in ground truth) with its around group. However, it seems more coherent than the ground truth.

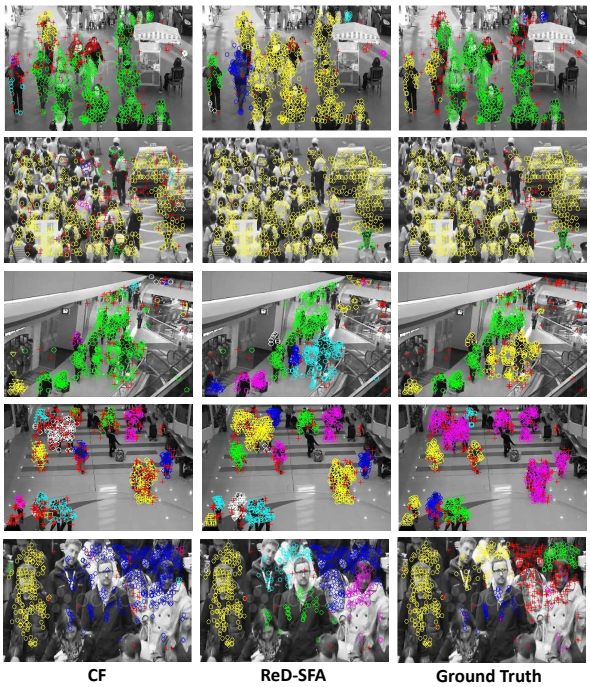


Figure 9. Some crowd detection results on CUHK Dataset. Note, the groups with the same color can be further distinguished by the markers. This graphic should be seen in color for complete clarity.

5. Conclusion

In this work, we proposed ReD-SFA for joint feature learning and relations discovery for trajectory clustering. Here, ReD-SFA is formulated as an problem of cross entropy minimization which continuously reduces the discrepancy between the updated pairwise similarities and the existing affinity graph, so that more reliable relations can be discovered. Furthermore, a Boosting-like strategy is proposed to tackle the imbalance of the discovered relations. Benefitting from the relation discovery, competitive results on different trajectory clustering tasks validate the proposed method. In future, we will extend the ReD-SFA to other data sources, e.g., images or videos. And a scalable ReD-SFA will be designed for large scale feature learning.

6. Acknowledgement

This work was supported by the National Basic Research Program of China under Grant 2012CB316302 and the National Natural Science Foundation of China under Grants 61473290, 61322209, and 61403388. This work was also supported by the Australian Research Council under Project Project DP-140102164.

References

[1] F. Bashir, A. Khokhar, and D. Schonfeld. Real-time motion trajectory-based indexing and retrieval of video sequences. *IEEE Transactions on Multimedia*, 9(1):58–65, 2007.

[2] P. Berkes. Handwritten digit recognition with nonlinear fisher discriminant analysis. *Artificial Neural Networks: Formal Models and Their Applications ICANN 2005*, page 285C287, 2005.

[3] P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, pages 579–602, 2005.

[4] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. *Proc. ECCV*, 6315(1):282–295, 2010.

[5] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE TPAMI*, 2008.

[6] G. Chen and G. Lerman. Motion segmentation by scc on the hopkins 155 database. *Proc. ICCV Workshops*, 2009.

[7] L. Chen, M. O. Tamer, and V. Oria. Robust and fast similarity search for moving object trajectories. *Proc. ACM SIGMOD International Conference on Management of Data*, pages 491–502, 2005.

[8] A. Cheriyadat and R. Radke. Non-negative matrix factorization of partial track data for motion segmentation. *Proc. ICCV*, pages 865–C872, 2009.

[9] Y. Deng, Y. Li, Y. Qian, X. Ji, and Q. Dai. Visual words assignment via information-theoretic manifold embedding. *IEEE T. CYBERNETICS*, 44, 2014.

[10] E. Elhamifar. Sparse modeling for high-dimensional multi-manifold data analysis. *Thesis*, 2012.

[11] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE T-PAMI*, (11):2765–2781, 2013.

[12] A. Escalante-B and L. Wiskott. How to solve classification and regression problems on high-dimensional data with a supervised extension of slow feature analysis. *Journal of Machine Learning Research*, (14):3683–3719, 2013.

[13] F. Flores-Mangas and A.Jepson. Fast rigid motion segmentation via incrementally-complex local model. *CVPR*, 2013.

[14] M. Fradet, P. Robert, and P. Perez. Clustering point trajectories with various lifespans. *European Conference on Visual Media Production (ECVMP)*, 2009.

[15] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE TPAMI*, 2012.

[16] A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. *Proc. CVPR*, 2007.

[17] H. Jung, J. Ju, and J. Kim. Rigid motion segmentation using randomized voting. *Proc. CVPR*, 2014.

[18] E. Keogh and M. Pazzani. Scaling up dynamic time warping for datamining application. *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 285–289, 2000.

[19] M. Kevin. *Machine Learning: A Probabilistic Perspective*. 2012.

[20] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE T-PAMI*, (1):171C–184, 2013.

[21] P. Ochs and T. Brox. Higher order motion models and spectral clustering. *Proc. CVPR*, pages 614C–621, 2012.

[22] R.J.Alcock and Y.Manolopoulos. Time-series similarity queries employing a feature-based approach. *Proc. Hellenic Conference on Informatics*, 1999.

[23] J. Shao, C. C. Loy, and X. Wang. Scene-independent group profiling in crowd. *Proc. CVPR*, 2014.

[24] S.Rao, R.Tron, R.Vidal, and Y.Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE. T-PAMI*, (10):1832–1845, 2010.

[25] L. Sun, K. Jia, T. Chan, Y. Fang, G. Wang, and S. Yan. Dl-sfa: Deeply-learned slow feature analysis for action recognition. *Proc. CVPR*, pages 2625–2632, 2014.

[26] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2603–2610, 2013.

[27] K. Torkkola. Feature extraction by non parametric mutual information maximization. *JMLR*, 3:1415C–1438, 2003.

[28] R. Tron and R. Vidal. A benchmark for the comparison of 3d motion segmentation algorithms. *Proc. CVPR*, 2007.

[29] M. Veillette. Alpha-stable distributions in matlab. *http://math.bu.edu/people/mveillet/html/alphastablepub.html*.

[30] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE T-PAMI*, (12):1945–C1959, 2005.

[31] M. Vrigkas, V. Karavasilis, C. Nikou, and I. Kakadiaris. Matching mixtures of trajectories for human action recognition. *Computer Vision and Image Understanding (CVIU)*, 19:27–40, 2014.

[32] X. Wang, K. T. Ma, G. Ng, and E. Grimson. Trajectory analysis and semantic region modeling using nonparametric bayesian models. *International Journal of Computer Vision (IJCV)*, 96:287–321, 2011.

[33] Y. Wang, H. Xu, and C. Leng. Provable subspace clustering:whe n lrr meets ssc. *Proc. NIPS*, 2013.

[34] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, (4):715–770, 2002.

[35] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. *Proc. ECCV*, 3954:94–106, 2006.

[36] Z. Zhang, K. Huang, and T. Tan. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. *Proc. ICPR*, 2006.

[37] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE T-PAMI*, (3):436–450, 2012.

[38] B. Zhou, X. Tang, and X. Wang. Coherent filtering: Detecting coherent motions from crowd clutters. *Proc. ECCV*, 2012.

[39] D. Zhou. Learning with local and global consistency. *NIPS*, 2003.