Provable Self-Representation Based Outlier Detection in a Union of Subspaces

Chong You[†], Daniel P. Robinson[‡], René Vidal[†]

[†]Center for Imaging Science, Johns Hopkins University
[‡]Applied Mathematics and Statistics, Johns Hopkins University

Big data in computer vision

20,500 terabytes
of image data

300 hours
of video/minute

3720 terabytes
of photos & videos

136,000 photos
every minute

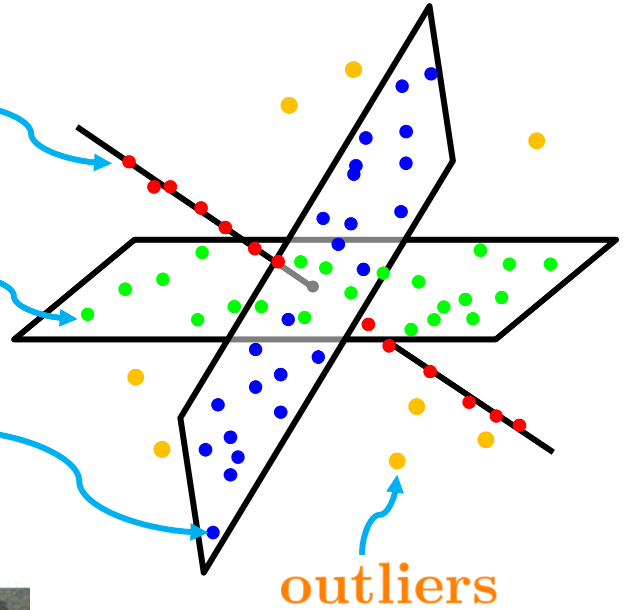
Most collected data are unlabeled/weakly labeled

http://mashable.com/2012/08/22/google-maps-facts/#1qlZlTRk9ZqS
https://fortunelords.com/youtube-statistics/
https://googlesystem.blogspot.com/2015/10/google-photos-stats.html#gsc.tab=0
https://zephoria.com/top-15-valuable-facebook-statistics/

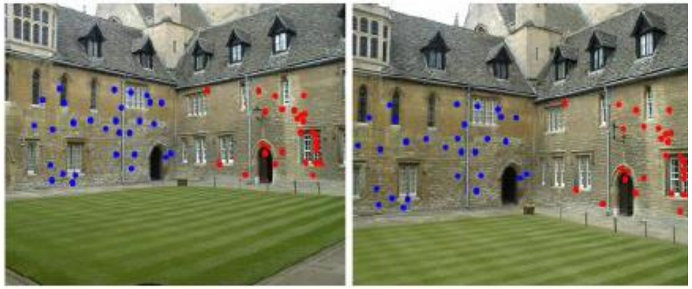
Subspace clustering

- High-dimensional visual data often lie in low-dimensional subspaces
- Subspace clustering is the problem of clustering data into subspaces

Face clustering



Motion segmentation



Planar segmentation

This work addresses sensitivity of subspace clustering to outliers

Data self-representation

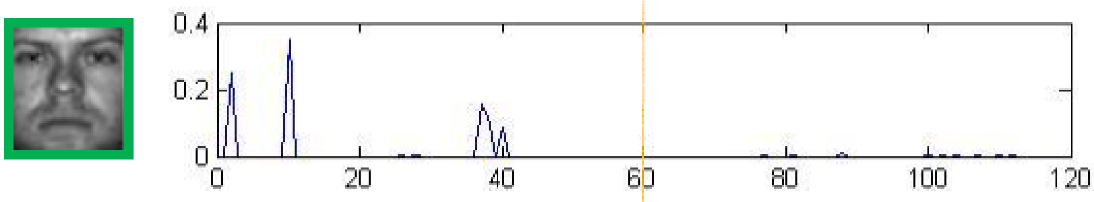
Given data $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$.

S1: Compute self-representation:

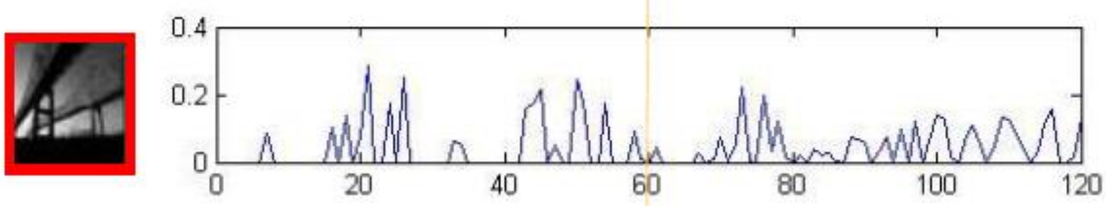
$$\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_1 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, \; c_{jj} = 0$$

X: inliers outliers

$\mathbf{x}_j$



$\mathbf{x}_j$ is an inlier:
$[\mathbf{c}_j]_i \neq 0 \rightarrow \mathbf{x}_i$ is inlier



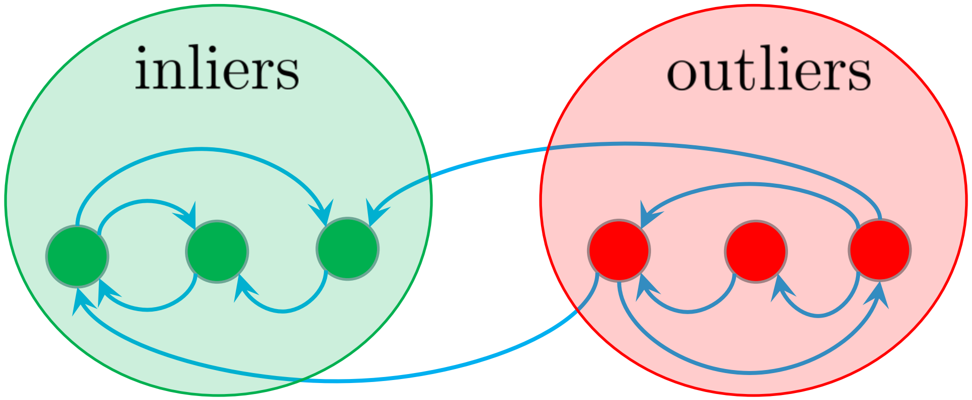
$\mathbf{x}_j$ is an outlier:
$[\mathbf{c}_j]_i \neq 0 \rightarrow \mathbf{x}_i$ can be both inlier and outlier

Random walk

S2: Define a random walk:

$\mathbf{x}_j \rightarrow \mathbf{x}_i$ if $[\mathbf{c}_j]_i \neq 0$

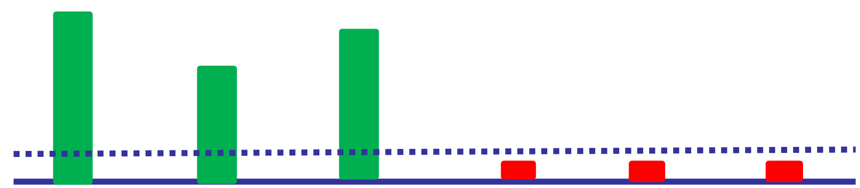
- No transition from inliers to outliers
- Any random walker will end up in the inliers



S3: Compute stationary distribution:

$\bar{\pi}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} \pi^{(0)} P^t$

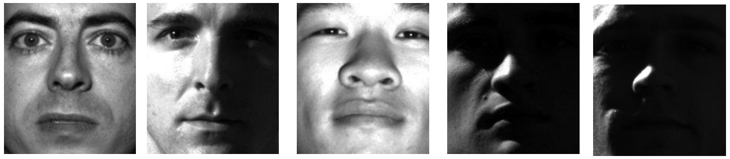
$\bar{\pi}^{(T)}$



Theorem: $[\bar{\pi}]_j^{(\infty)} = 0 \leftrightarrow \mathbf{x}_j$ is an outlier

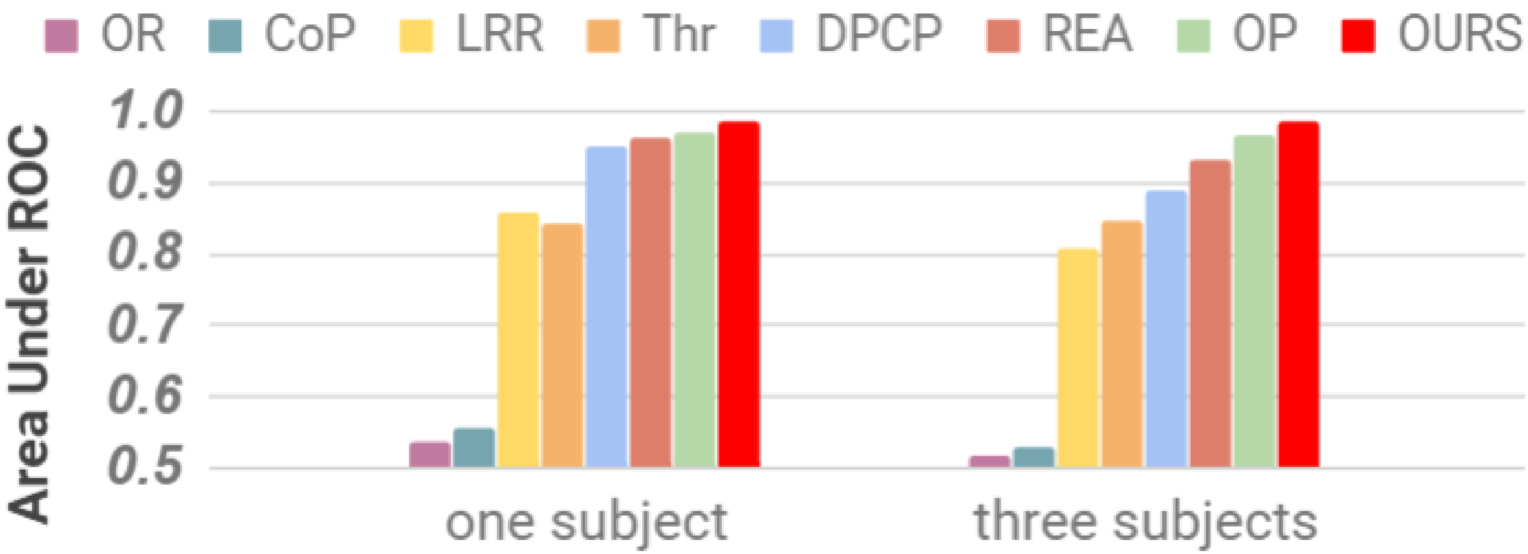
Experiments on Extended Yale B

Datasets: Extended Yale B (shown here), Caltech-256, Coil-100

Inliers:  Outliers: 

Methods:

Results:



- Our method achieves the best outlier detection performance.

Conclusion

- In many computer vision problems, data are unlabeled, composed of multiple groups and corrupted by outliers
- We designed an algorithm to reject outliers by combining self-representation with random walk
- Our method is provably correct

Acknowledgement

Thank you!

Funding: NSF-IIS 1447822

Vision Lab @ Johns Hopkins University