Sparse and Low-Dimensional Representation

Lecture 3: Modeling High-dimensional (Visual) Data

Yi Ma

Visual Computing Group
Microsoft Research Asia, Beijing

ECE Department
University of Illinois, Urbana






CONTEXT – *Data increasingly massive, high-dimensional…*

Images

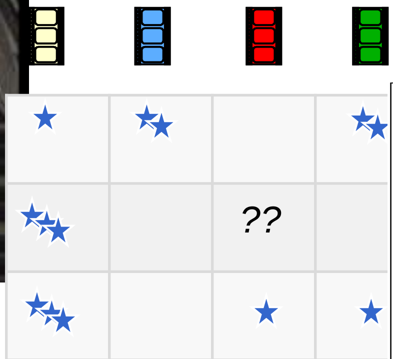
⇓ ➢ ***1M pixels***

Compression
De-noising
Super-resolution
Recognition…

Videos

⇓ ➢ ***1B voxels***

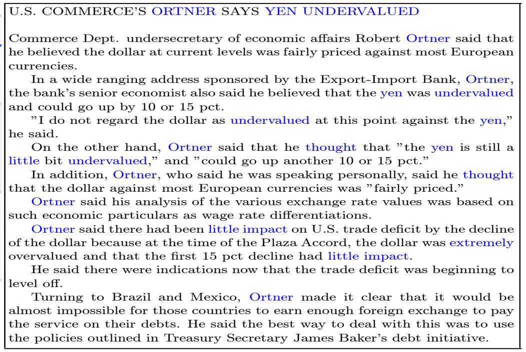
Streaming
Tracking
Stabilization…



User data

⇓ ➢ ***1B users***

Clustering
Classification
Collaborative filtering…



U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

Commerce Dept. undersecretary of economic affairs Robert Ortner said that he believed the dollar at current levels was fairly priced against most European currencies.

In a wide ranging address sponsored by the Export-Import Bank, Ortner, the bank's senior economist also said he believed that the yen was undervalued and could go up by 10 or 15 pct.

"I do not regard the dollar as undervalued at this point against the yen," he said.

On the other hand, Ortner said that he thought that "the yen is still a little bit undervalued," and "could go up another 10 or 15 pct."

In addition, Ortner, who said he was speaking personally, said he thought that the dollar against most European currencies was "fairly priced."

Ortner said his analysis of the various exchange rate values was based on such economic particulars as wage rate differentiations.

Ortner said there had been little impact on U.S. trade deficit by the decline of the dollar because at the time of the Plaza Accord, the dollar was extremely overvalued and that the first 15 pct decline had little impact.

He said there were indications now that the trade deficit was beginning to level off.

Turning to Brazil and Mexico, Ortner made it clear that it would be almost impossible for those countries to earn enough foreign exchange to pay the service on their debts. He said the best way to deal with this was to use the policies outlined in Treasury Secretary James Baker's debt initiative.

Web data

⇓ ➢ ***100B webpages***

Indexing
Ranking
Search…

How to extract low-dim structures from such high-dim data?

Everything old …

*A **long and rich history** of estimating unknown models (or signals) from noisy or erroneous observations:*

R. J. Boscovich. *De calculo probailitatum que respondent diversis valoribus summe errorum post plures observationes … , before 1756*

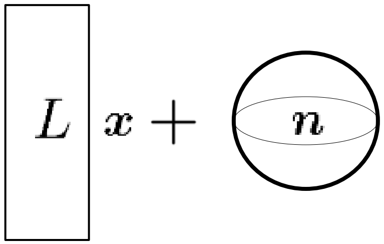
A. Legendre. *Nouvelles methodes pour la determination des orbites des cometes*, 1806

C. Gauss. *Theory of motion of heavenly bodies, 1809*

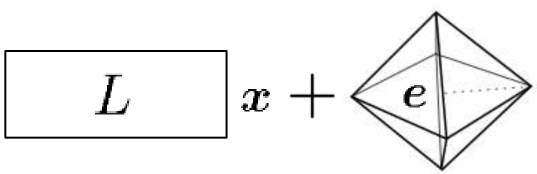
A. Beurling. *Sur les integrales de Fourier absolument convergentes et leur application a une transformation functionelle*, 1938

B. Logan. *Properties of High-Pass Signals*, 1965

⋮



over-determined
+ dense, Gaussian



underdetermined
+ sparse, Laplacian

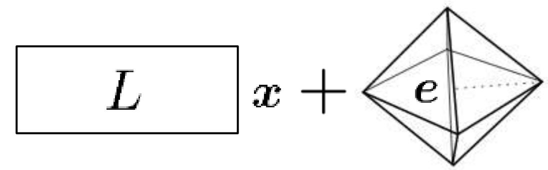
… is new again

*Today, robust estimation of low-dim models in high-dim space is **urgently needed** and increasingly **better understood.***

Theory – high-dimensional geometry & statistics, measure concentration, combinatorics, coding theory…

Algorithms – large scale convex optimization, geometric convergence rate, parallel and distributed computing …

Applications – massive data driven methods, hashing, compressing, denoising, superresolution, MRI, bioinformatics, image classification, recognition …

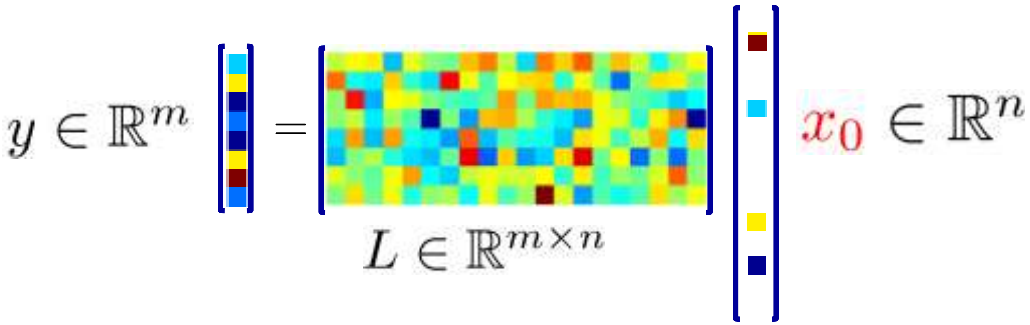


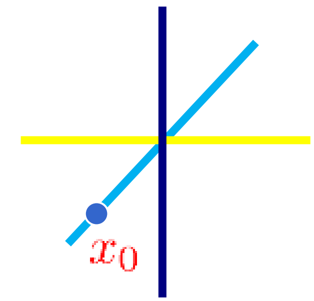
underdetermined
+ sparse, Laplacian

$$\min \|x\|_1 + \|e\|_1$$

CONTEXT – *Sparse models*

Sparse recovery: Given $y = Lx_0$, $L \in \mathbb{R}^{m \times n}$, $m \ll n$, recover x_0.





Impossible in general ($m \ll n$)

Well-posed if x_0 is structured *(sparse),* but still **NP-hard**

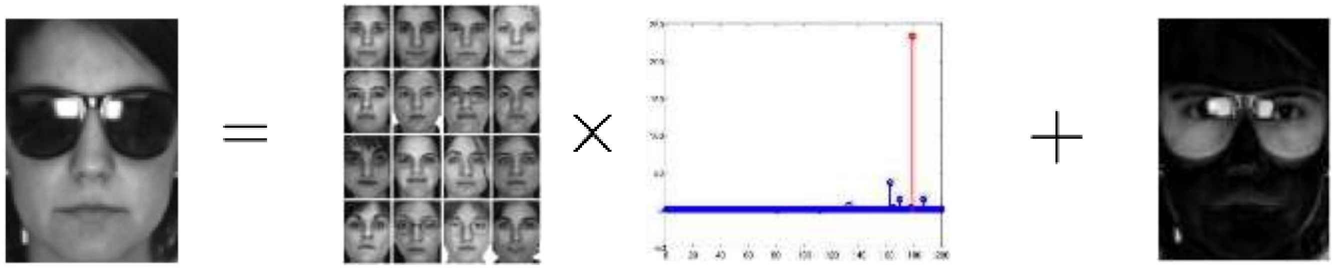
Tractable via convex optimization: $\min \|x\|_1$ s.t. $y = Lx$

… if L is "nice" *(random, incoherent, RIP)*

Hugely active area: *Donoho+Huo '01, Elad+Bruckstein '03, Candès+Tao '04,'05, Tropp '04, '06, Donoho '04, Fuchs '05, Zhao+Yu '06, Meinshausen+Buhlmann '06, Wainwright '09, Donoho+Tanner '09 … and many others*

CONTEXT – *Sparse models*

Robust recovery: Given $y = L{\color{red}x_0} + {\color{green}e_0}$, $L \in \mathbb{R}^{m \times n}$, $m \ll n$, recover ${\color{red}x_0}$ and ${\color{green}e_0}$.



y L ${\color{red}x}$ ${\color{green}e}$

Impossible in general ($m \ll n + m$)

Well-posed if ${\color{red}x_0}$ is *sparse,* errors ${\color{green}e_0}$ not too dense, but still **NP-hard**

Tractable: via convex optimization: $\min \|{\color{red}x}\|_1 + \|{\color{green}e}\|_1 \text{ s.t. } y = L{\color{red}x} + {\color{green}e}$

… if L is "nice" *(cross and bouquet)*

***Hugely active area**: Candès+Tao '05, Wright+Ma '10, Nguyen+Tran '11, Li '11, also Zhang, Yang, Huang'11, etc…*

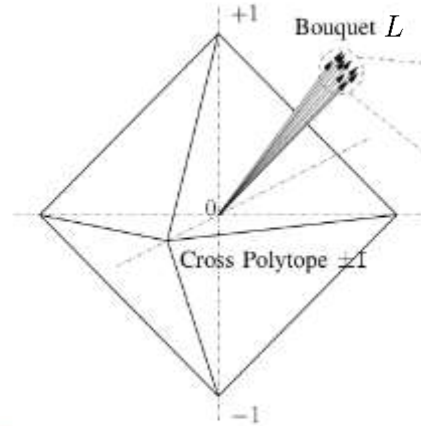
CONTEXT – *Dense Error Correction*

Extended Yale B Database (38 subjects)

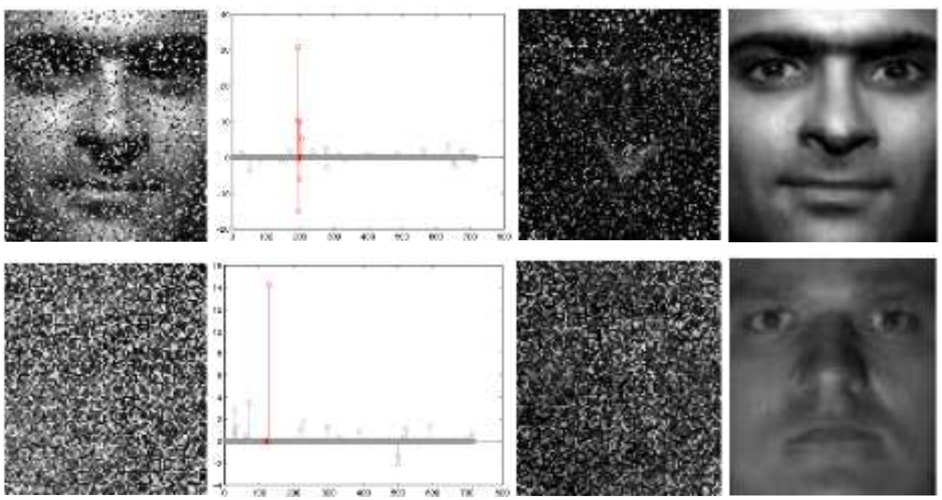
Training: subsets 1 and 2 (717 images)

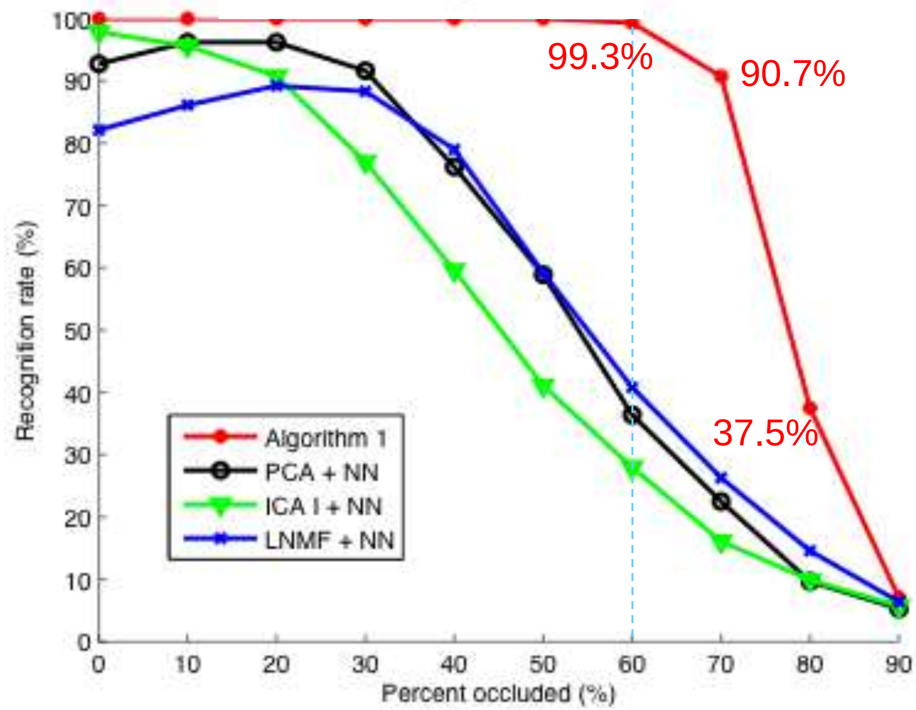
Testing: subset 3 (453 images)

$$y = Lx + e$$



y $\quad$ $\hat{x}$ $\quad$ $\hat{e}$ $\quad$ $\hat{y} = L\hat{x}$





CONTEXT – *Extension to Single Gallery Image Case*

$$y = Lx + Ab + e$$

A: a common dictionary for intraclass variabilities: illumination, expression, and pose.

x, b, e are sparse

FERET Dataset

General training: 1,002 images of 429 people

Gallery training: 1,196 images of 1,196 people

Probe sets:

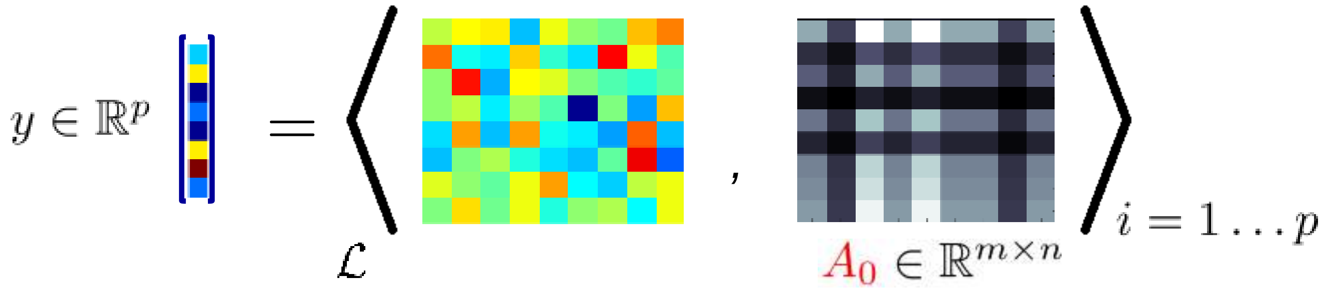
fb (1,195, expression), fc (194, lighting), dup1 (722, different time), dup2 (234, a year)

TABLE 3
Comparative Recognition Rates of SRC and ESRC on the FERET Database Using the FERET'96 Testing Protocol

	Feature	Dsampled Image	Pixel-Rfaces	Pixel	Gabor-Rfaces	Gabor	LBP-Rfaces	LBP
Probe set	Dim	24×24	540	16384	540	10240	540	15104
fb	SRC ESRC	86.4 94.8(+8.4)	82.4 91.5(+9.1)	85.3 92.8(+7.5)	89.5 94.1(+4.6)	92.8 **97.3**(+4.5)	91.5 95.2(+3.7)	96.7 **97.3**(+0.6)
fc	SRC ESRC	69.6 67.5(–2.1)	75.8 78.9(+3.1)	76.3 79.4(+3.1)	96.4 96.9(+0.5)	97.4 **99.0**(+1.6)	72.7 71.1(–1.6)	93.3 95.4(+2.1)
dup1	SRC ESRC	62.7 75.6(+12.9)	60.9 73.1(+12.2)	63.7 77.0(+13.3)	63.0 73.5(+10.5)	72.7 85.0(+12.3)	75.2 81.0(+5.8)	87.7 **93.8**(+6.1)
dup2	SRC ESRC	52.6 62.4(+9.8)	53.0 59.8(+6.8)	55.6 66.2(+10.6)	70.1 72.6(+2.5)	76.5 85.9(+9.4)	69.7 71.4(+1.7)	83.8 **92.3**(+8.5)

CONTEXT – *Low-rank models*

Low-rank sensing: Given $y = \mathcal{L}[A_0]$, $\mathcal{L} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$, recover A_0.



Impossible in general ($p \ll mn$)

Well-posed if A_0 is structured *(low-rank),* but still **NP-hard**

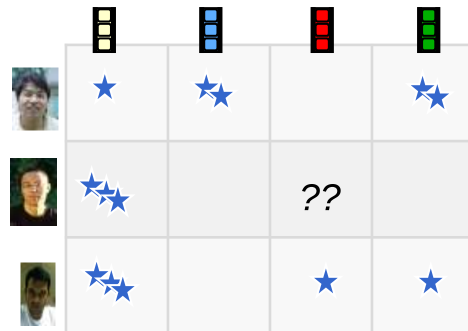
Tractable via convex optimization: $\min \|A\|_* \text{ s.t. } y = \mathcal{L}(A)$

… if $\mathcal{L}$ is "nice" *(random, rank-RIP)*

Hugely active area: *Recht+Fazel+Parillo '07, Candès+Plan '10, Mohan+Fazel '10, Recht+Xu+Hassibi '11, Chandrasekaran+Recht+Parillo+Willsky '11, Negahban+Wainwright '11 …*

CONTEXT – *Low-rank models*

Matrix completion: Given $y = \mathcal{P}_\Omega[A_0]$, $\Omega \subset [m] \times [n]$, recover A_0.



$A_0 \in \mathbb{R}^{m \times n}$

Impossible in general ($|\Omega| \ll mn$)

Well-posed if A_0 is structured *(low-rank),* but still **NP-hard**

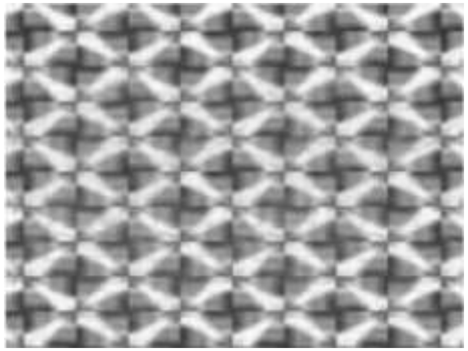
Tractable via convex optimization: $\min \|A\|_* \text{ s.t. } y = \mathcal{P}_Q(A)$

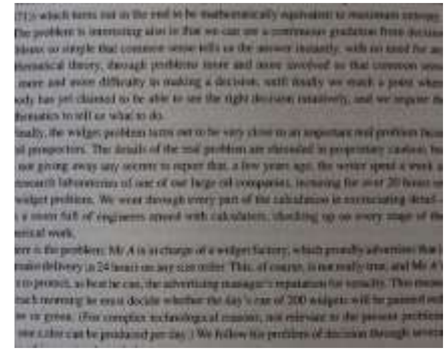
… if Ω is "nice" *(random subset) ...*

… and A_0 interacts "nicely" with $\mathcal{P}_\Omega$ (A_0 *incoherent – not "spiky")*.

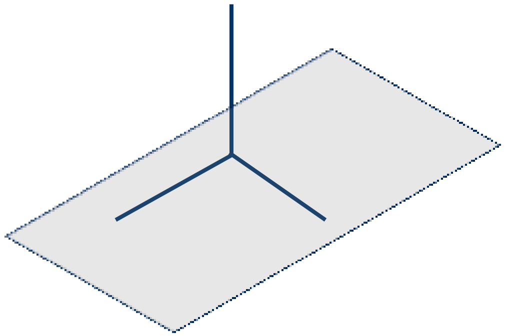
Hugely active area: *Candès+Recht '08, Keshevan+Oh+Montonari '09, Candès+Tao '09, Gross '10, Recht '10, Negahban+Wainwright '10*

CONTEXT – *Low dimensional structures in visual data*






Visual data exhibit ***low-dimensional structures*** due to rich ***local*** regularities, ***global*** symmetries, ***repetitive*** patterns, or ***redundant*** sampling.



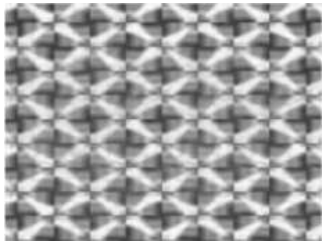
CONTEXT – *PCA: Fitting Data with a Low-dim. Subspace*

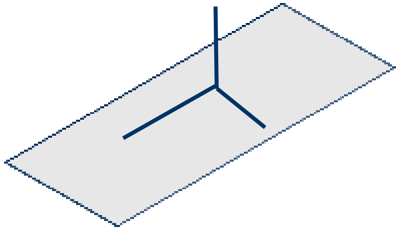
If we view the data (image) as a matrix

$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}$$

then

$$r \doteq \operatorname{rank}(A) \ll m.$$



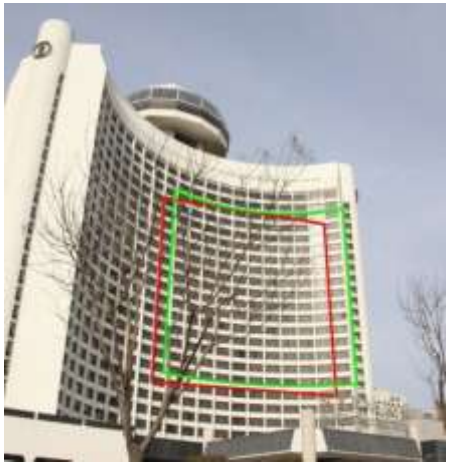


Principal Component Analysis (PCA) via singular value decomposition (SVD):

- Optimal estimate of A under iid Gaussian noise $D = A + Z$
- Efficient and scalable computation
- Fundamental statistical tool, with huge impact in image processing, vision, web search, bioinformatics…

But… **PCA breaks down under even a single corrupted observation.**

CONTEXT – *But life is not so easy…*

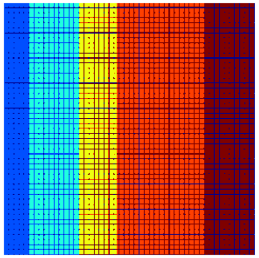


Real application data often contain ***missing observations****,* ***corruptions,*** *or subject to unknown* ***deformation or misalignment.***

Classical methods (e.g., PCA, least square regression) break down…

THIS TALK – *Low-rank + Sparse Models*

The data should be ***low-dimensional (low-rank)****:*

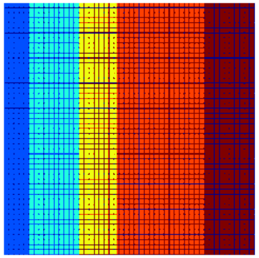


$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \mathrm{rank}(A) \ll m.$$

THIS TALK – *Low-rank + Sparse Models*

*The data should be **low-dimensional**:*

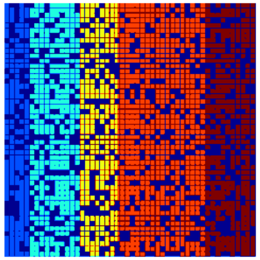
$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \mathrm{rank}(A) \ll m.$$



*… but some of the observations are **grossly corrupted**:*

$$A + E, \qquad |E_{ij}|$$

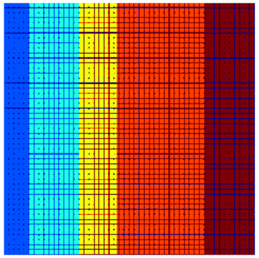
E_{ij} arbitrarily large, but most are zero.



THIS TALK – *Low-rank + Sparse Models*

*The data should be **low-dimensional**:*

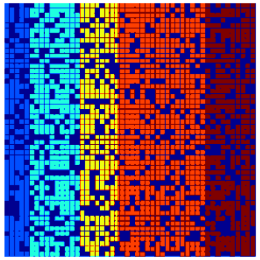
$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \operatorname{rank}(A) \ll m.$$



*… but some of the observations are **grossly corrupted**:*

$$A + E, \qquad |E_{ij}|$$

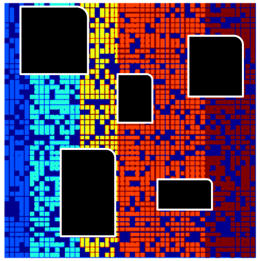
E_{ij} arbitrarily large, but most are zero.



*… and some of them can be **missing** too:*

$$D = \mathcal{P}_\Omega[A + E],$$

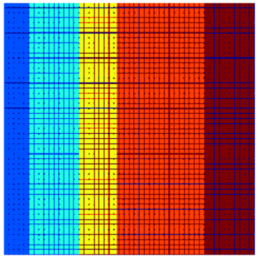
$\Omega \subset [m] \times [n]$ the set of observed entries.



THIS TALK – *Low-rank + Sparse Models*

*The data should be **low-dimensional**:*

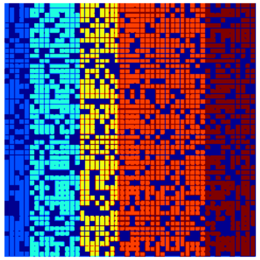
$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \operatorname{rank}(A) \ll m.$$



*… but some of the observations are **grossly corrupted**:*

$$A + E, \qquad |E_{ij}|$$

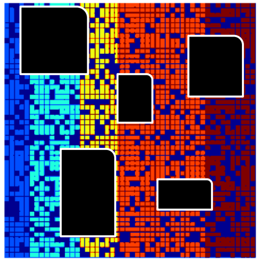
E_{ij} arbitrarily large, but most are zero.



*… and some of them can be **missing** too:*

$$D = \mathcal{P}_\Omega[A + E],$$

$\Omega \subset [m] \times [n]$ the set of observed entries.



… special cases of a more general problem:

$$D = \mathcal{L}_1(A) + \mathcal{L}_2(E) + Z \qquad A, E \text{ either sparse or low-rank}$$

THIS TALK

Given observations $D = \mathcal{P}_Q[A + E + Z]$, *with*

A *low-rank,*

E *sparse,*

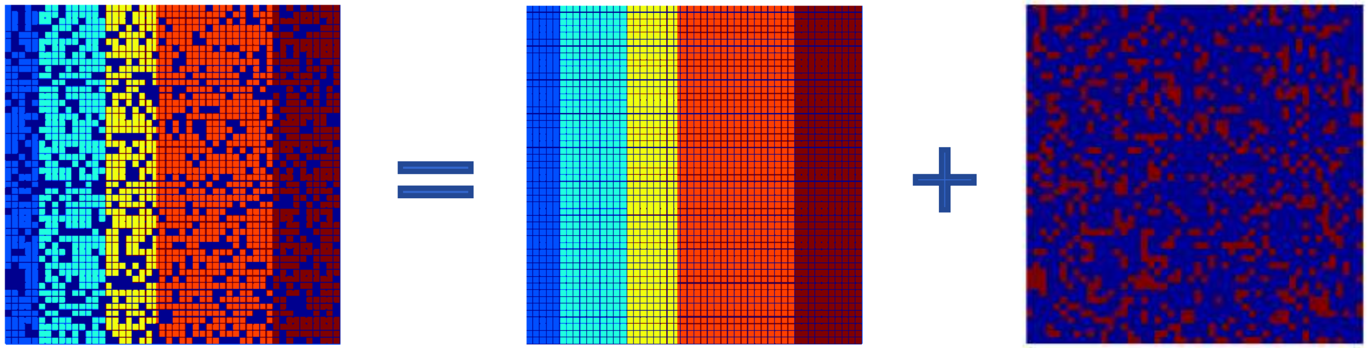
Z *small, dense noise,*

recover a good estimate of A *and* E.

- **Theory and Algorithms**
  - **Provably Correct and Tractable Solution**
  - **Provably Optimal and Efficient Algorithms**
- **Potential Applications**
  - **Visual Data (Reconstruction, Recognition etc.)**
  - **Other Data**
- **Conclusions**

ROBUST PCA – *Problem Formulation*

D - observation $\qquad A_0$ – low-rank $\qquad E_0$ – sparse



Problem: Given $D = A_0 + E_0$, recover A_0 and E_0.

Low-rank component $\qquad$ Sparse component (gross errors)

Numerous approaches in the literature:

- Multivariate trimming [Gnanadesikan and Kettering '72]
- Power Factorization [Wieber'70s]
- Random sampling [Fischler and Bolles '81]
- Alternating minimization [Shum & Ikeuchi'96, Ke and Kanade '03]
- Influence functions [de la Torre and Black '03]

Key question: ***can guarantee correctness with an efficient algorithm?***

ROBUST PCA – *Convex Surrogates for Sparsity and Rank*

Seek the lowest-rank A that agrees with the data up to some sparse error E:

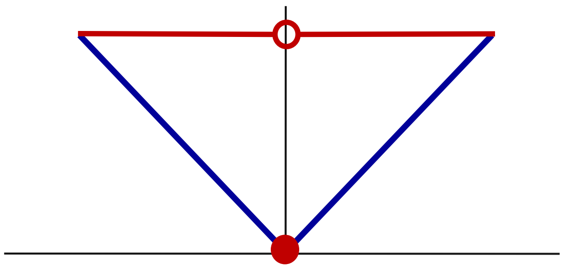
$$\min\ \operatorname{rank}(A) + \gamma\|E\|_0 \quad \text{subj} \quad A + E = D.$$

But INTRACTABLE! Relax with convex surrogates:

$$\|E\|_0 = \#\{E_{ij} \neq 0\} \quad \rightarrow \quad \|E\|_1 = \textstyle\sum_{ij} |E_{ij}|.$$ L_1 norm

$$\operatorname{rank}(A) = \#\{\sigma_i(A) \neq 0\} \quad \rightarrow \quad \|A\|_* = \textstyle\sum_i \sigma_i(A).$$ Nuclear norm

Convex envelope over $B_{2,2} \times B_{1,\infty}$



ROBUST PCA – *By Convex Optimization*

Seek the lowest-rank A that agrees with the data up to some sparse error E:

$$\min \ \operatorname{rank}(A) + \gamma \|E\|_0 \quad \text{subj} \quad A + E = D.$$

But INTRACTABLE! Relax with convex surrogates:

$$\|E\|_0 = \#\{E_{ij} \neq 0\} \quad \rightarrow \quad \|E\|_1 = \textstyle\sum_{ij} |E_{ij}|.$$

L_1 norm

$$\operatorname{rank}(A) = \#\{\sigma_i(A) \neq 0\} \quad \rightarrow \quad \|A\|_* = \textstyle\sum_i \sigma_i(A).$$

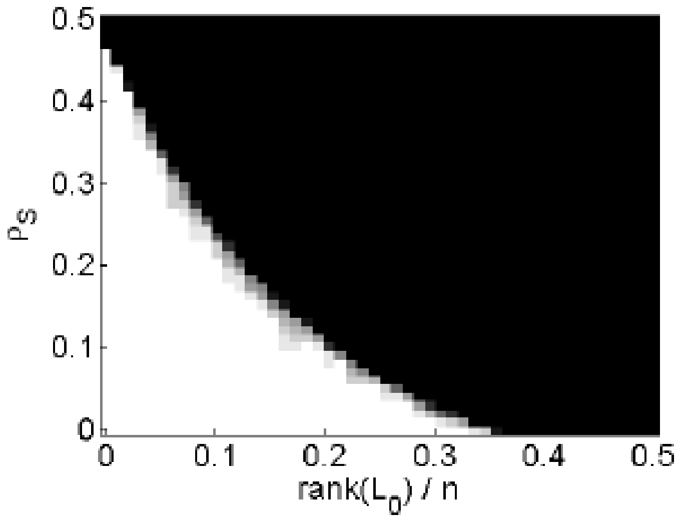
Nuclear norm

$$\boxed{\min \ \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D.}$$

Semidefinite program, solvable in polynomial time

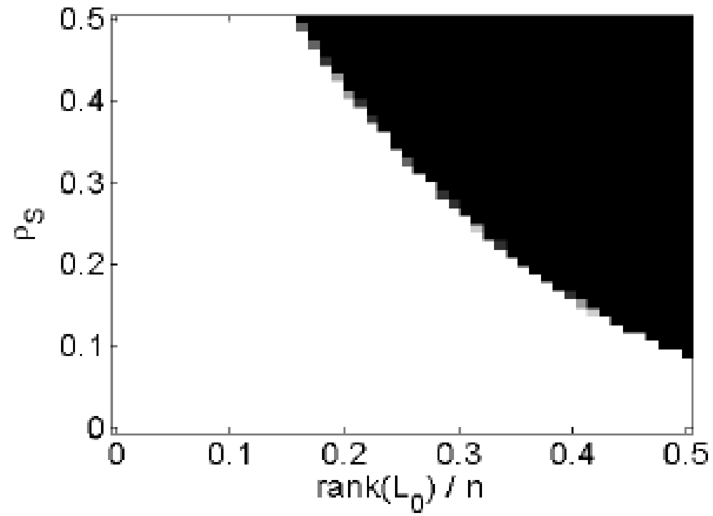
ROBUST PCA – *When the Convex Program Works?*

$$D = A + E$$



Robust PCA, Random Signs

$$D = \mathcal{P}_\Omega[A]$$



Matrix Completion

White regions are instances with perfect recovery.

Correct recovery when A is indeed **low-rank** and E is indeed **sparse**?

MAIN THEORY – *Exact Solution by Convex Optimization*

Theorem 1 (Principal Component Pursuit). *If* $A_0 \in \mathbb{R}^{m \times n}$, $m \geq n$ *has rank*

m

Non-adaptive weight factor

and E_0 *has Bernoulli support with error probability* $\rho \leq \rho_s^*$, *then with very high probability*

$$(A_0, E_0) = \arg\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj} \quad A + E = A_0 + E_0,$$

and the minimizer is unique.

GREAT NEWS: *"Convex optimization recovers almost any matrix of rank* $O\left(\frac{m}{\log^2 n}\right)$ *from errors corrupting* $O(mn)$ *of the observations!"*

Candes, Li, Ma, and Wright, Journal of the ACM, May 2011.

MAIN THEORY – *Corrupted, Incomplete Matrix*

$$D = \mathcal{P}_\Omega[\ A_0 + E_0\ ], \qquad \Omega \sim \mathrm{uni}\binom{[m]\times[n]}{mn}$$

Theorem 2 **(Matrix Completion and Recovery).** *If* $A_0, E_0 \in \mathbb{R}^{m\times n}, m \geq n$, *with*

$$\mathrm{rank}(A_0) \ \leq \ C \frac{n}{\mu \log^2(m)}, \quad \textit{and} \quad \|E_0\|_0 \ \leq \ \rho^\star mn,$$

and we observe only a random subset of size

$$|\Omega| = mn/10$$

entries, then with very high probability, solving the convex program

$$\min \|A\|_* + \tfrac{1}{\sqrt{m}}\|E\|_1 \quad \mathrm{subj} \quad P_\Omega[A+E] = D,$$

uniquely recovers (A_0, E_0).

MAIN THEORY – *With Dense Errors and Noise*

Theorem 3 (Dense Error Correction). *If A_0 has rank $r \leq \rho_r \frac{m}{\mu^2 \log^2(n)}$ and E_0 has random signs and Bernoulli support with error probability $\rho < 1$, then with very high probability*

$$(A_0, E_0) \;=\; \arg\min \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = A_0 + E_0,$$

and the minimizer is unique.

Theorem 4 (Robust PCA with Noise). *Given $D = A_0 + E_0 + Z$ for any $\|Z\|_F \leq \eta$, if A_0 has rank $r \leq \rho_r \frac{m}{\mu^2 \log^2(n)}$ and E_0 has Bernoulli support with error probability $\rho \leq \rho_s^\star$, then with very high probability*

$$(\hat{A}, \hat{E}) \;=\; \arg\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj} \quad \|D - A - E\| \leq \eta,$$

sastisfies $\|(\hat{A}, \hat{E}) - (A_0, E_0)\| \leq C\eta$ for some constant $C > 0$.

MAIN THEORY – *Compressive Robust PCA*

Theorem 5 (Compressive Principal Component Pursuit). *Let* $A_0 \in \mathbb{R}^{m\times n}$, $m \geq n$ *have rank* $r \leq \rho_r \frac{m}{\mu^2 \log^2(n)}$, *and* E_0 *have a Bernoulli support with error probability* $\rho < \rho^\star$. *Let* $Q^\perp$ *be a random subspac of* $\mathbb{R}^{m\times n}$ *of dimension*

$$\dim(Q) \geq C_Q(\rho mn + mr) \cdot \log^2 m,$$

distributed according to the Haar measure, independent of the support of E_0. *Then with very high probability*

$$(A_0, E_0) = \arg\min \|A\|_* + \frac{1}{\sqrt{m}}\|E\|_1 \quad \text{subj} \quad \mathcal{P}_Q[A+E] = \mathcal{P}_Q[A_0+E_0],$$

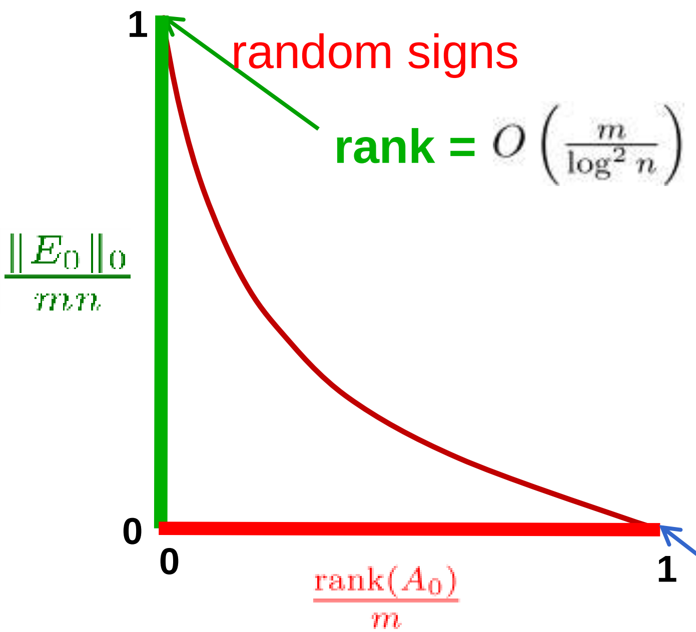
for some numerical constant ρ_r, C_p *and* $\rho^\star$, *and the minimizer is unique.*

A nearly optimal lower bound on minimum # of measurements!

BIG PICTURE – *Landscape of Theoretical Guarantees*

What people have known so far in the past 3-4 years:

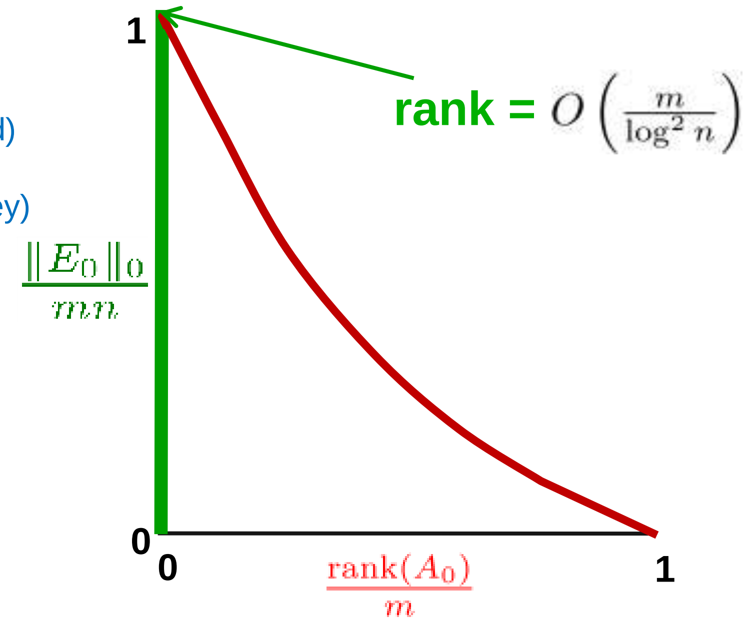
Matrix Recovery (RPCA)



D. Gross
B. Hassibi (Caltech)
J. Tropp (Caltech)
P. Parrilo (MIT)
A. Willsky (MIT)
B. Hastie (Stanford)
E. Candes (Stanford)
A. Montanari (Stanford)
M. Jordan (Berkeley)
M. Wainwright (Berkeley)
B. Yu (Berkeley)
A. Singer (Princeton)
T. Tao (UCLA)
S. Osher (UCLA)
O. Milenkovic (UIUC)
Y. Bresler (UIUC)
Y. Ma (UIUC)
B. Recht (Wisconsin)
M. Fazel (U Wash.)
……

Classical PCA

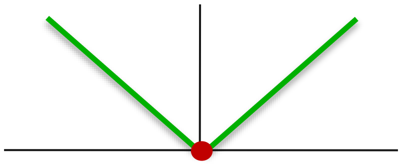
Matrix Completion



ALGORITHMS – *Are scalable solutions possible?*

Seemingly BAD NEWS: Our optimization problem

$$\min_{A,E} \ \|A\|_* + \lambda \|E\|_1 \ \text{ subj } \ A + E = D$$



is high-dimensional and non-smooth.

Convergence rate of solving a generic convex program:

$$\min_x f(x)$$

Second-order Newton method, # of iterations: $O(\log(1/\varepsilon))$, but not scalable!
First-order methods depend strongly on the smoothness of f:

$$\begin{array}{ll} f \text{ smooth, } \nabla f \text{ Lipschitz:} & O(\varepsilon^{-1/2}) \\ f \text{ differentiable:} & O(\varepsilon^{-1}) \\ f \text{ non-smooth:} & O(\varepsilon^{-2}) \end{array}$$

ALGORITHMS – *Why are scalable solutions possible?*

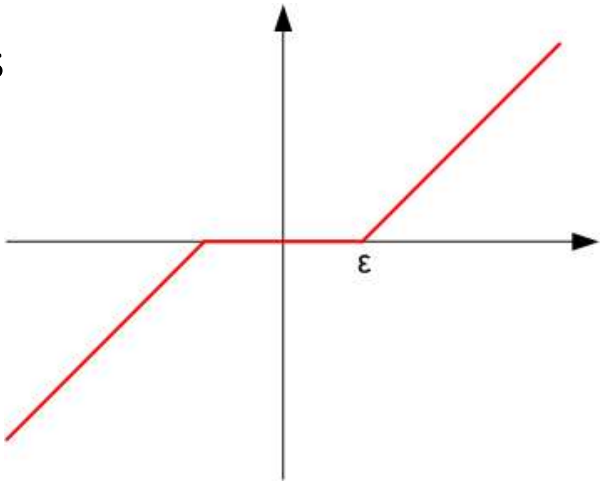
GOOD NEWS: The objective function has special structures

$$\min \ \|A\|_* + \lambda \|E\|_1 \ \text{subj} \ A + E = D$$

KEY OBSERVATION: closed form solutions for the proximal minimizations:

$$\begin{aligned} \mathcal{S}_\varepsilon(Q) &= \operatorname{argmin}_X \varepsilon \|X\|_1 + \frac{1}{2}\|X - Q\|_F^2 \\ \mathcal{D}_\varepsilon(Q) &= \operatorname{argmin}_X \varepsilon \|X\|_* + \frac{1}{2}\|X - Q\|_F^2 \end{aligned}$$

Solutions are given by soft-thresholding the entries and singular values of the matrix, respectively:



GOOD NEWS: Scalable first-order gradient-descent algorithms:

- Iterative Thresholding [Osher, Mao, Dong, Yin '09,Wright et. al.'09, Cai et. al.'09].
- Accelerated Proximal Gradient [Nesterov '83, Beck and Teboulle '09]:
- Augmented Lagrange Multiplier [Hestenes '69, Powell '69]:
- Alternating Direction Method of Multipliers [Gabay and Mercier '76].

A scalable algorithm: alternating direction method (ADM) for ALM:

$$l(A, E, Y) = \|A\|_* + \lambda \|E\|_1 + \langle Y, D - A - E \rangle + \frac{\mu}{2} \|D - A - E\|_F^2$$

$$\text{repeat} \begin{cases} A_{k+1} = \mathcal{D}_{\mu_k^{-1}}(D - E_k + Y_k/\mu_k), & \textit{Shrink singular values} \\ E_{k+1} = \mathcal{S}_{\lambda\mu_k^{-1}}(D - A_{k+1} + Y_k/\mu_k), & \textit{Shrink absolute values} \\ Y_{k+1} = Y_k + \mu_k(D - A_{k+1} - E_{k+1}). & \end{cases}$$

Cost of each iteration is a classical PCA, i.e. a (partial) SVD.

ALGORITHMS – *Evolution of fast algorithms (around 2009)*

For a 1000x1000 matrix of rank 50, with 10% (100,000) entries randomly corrupted: $\min \|A\|_* + \lambda\|E\|_1 \;\; \text{subj} \;\; A + E = D.$

Algorithms	Accuracy	Rank	||E||_0	# iterations	time (sec)
IT	5.99e-006	50	101,268	8,550	119,370.3
DUAL	8.65e-006	50	100,024	822	1,855.4
APG	5.85e-006	50	100,347	134	1,468.9
APG_P	5.91e-006	50	100,347	134	82.7
$EALM_P$	2.07e-007	50	100,014	34	37.5
$IALM_P$	3.83e-007	50	99,996	23	11.8

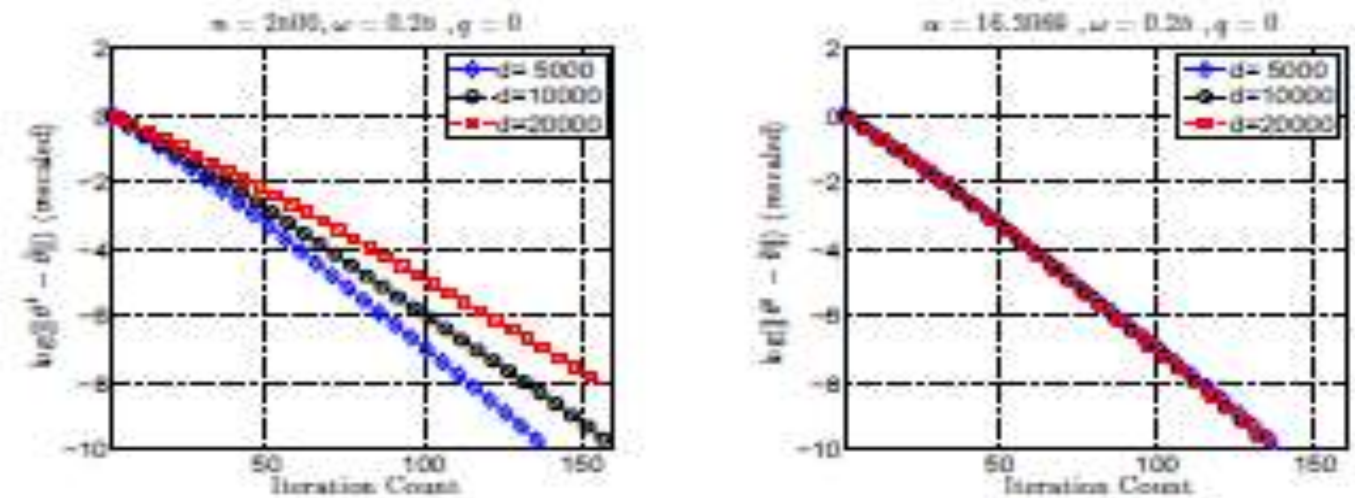
10,000 times speedup!

***Provably Robust PCA** at only a constant factor (≈20) more computation than conventional PCA!*

ALGORITHMS – *Convergence rate with strong convexity*

GREAT NEWS: Geometric convergence for gradient algorithms!

f restricted strong convex:	$O(\log(1/\varepsilon))$	[Agarwal, Negahban, Wainwright, NIPS 2010]
f smooth, ∇f Lipschitz:	$O(\varepsilon^{-1/2})$	
f differentiable:	$O(\varepsilon^{-1})$	
f non-smooth:	$O(\varepsilon^{-2})$	



(a) (b)

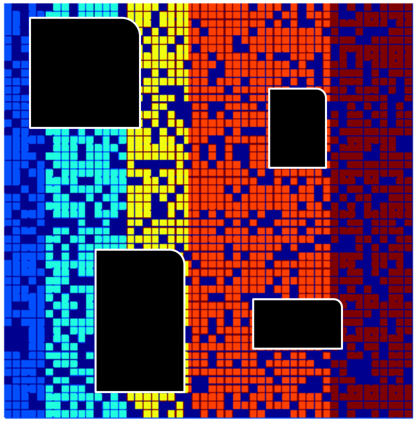
Figure 1. Convergence rates of projected gradient descent in application to Lasso programs (ℓ_1-constrained least-squares). Each panel shows the log optimization error $\log\|\theta^t - \widehat{\theta}\|$ versus the iteration number t. Panel (a) shows three curves, corresponding to dimensions $d \in \{5000, 10000, 20000\}$, sparsity $s = \lceil\sqrt{d}\rceil$, and all with the same sample size $n = 2500$. All cases show geometric convergence, but the rate for larger problems becomes progressively slower. (b) For an appropriately rescaled sample size ($\alpha = \frac{n}{s \log d}$), all three convergence rates should be roughly the same, as predicted by the theory.

APPLICATIONS

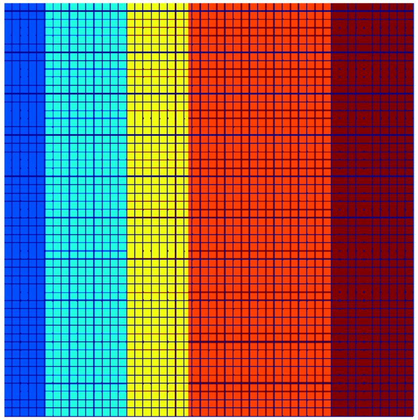
- **Repairing Images and Videos**
  - Image Repairing, Background Extraction, Street Panorama
- **Reconstructing 3D Geometry**
  - Shape from Texture, Featureless 3D Reconstruction
- **Registering Multiple Images**
  - Multiple Image Alignment, Video Stabilization
- **Recognizing Objects**
  - Faces, Texts, etc
- **Other Data and Applications**

Recover low-dimensional structures with a fraction of missing measurements with structured support.

Observed Entries

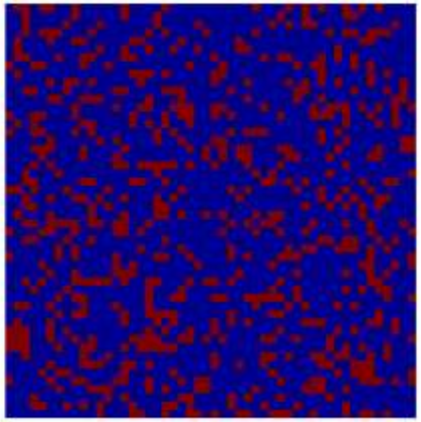


Low-rank Structures



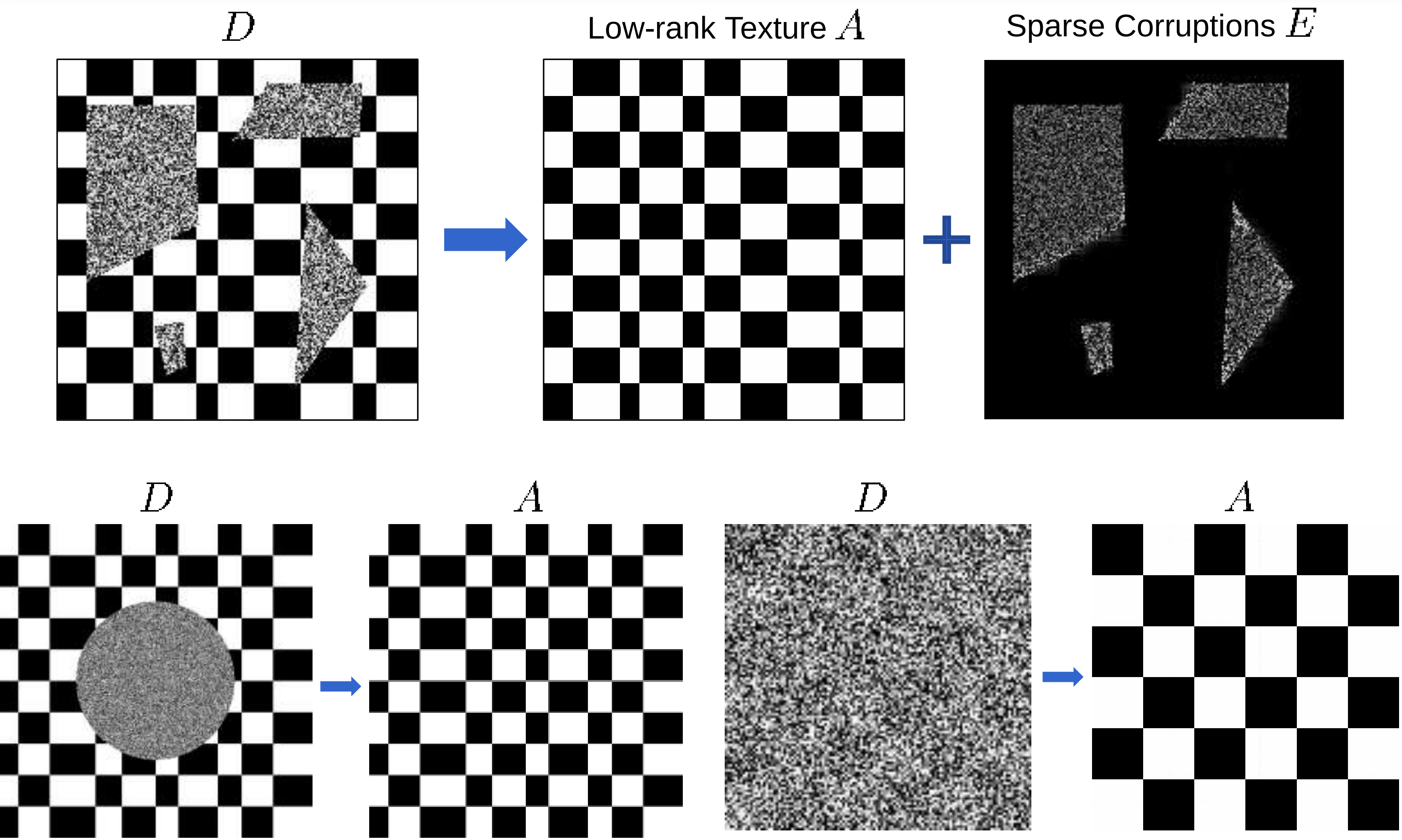
+

Sparse Structures



$$D = \mathcal{P}_\Omega[A + E], \quad \Omega \subset [m] \times [n] \text{ the set of observed entries.}$$

Repairing Images: Highly Robust Repairing of Low-rank Textures!



Liang, Ren, Zhang, and Ma, *Repairing Sparse Low-Rank Texture*, in ECCV 2012.

Repairing Low-rank Textures

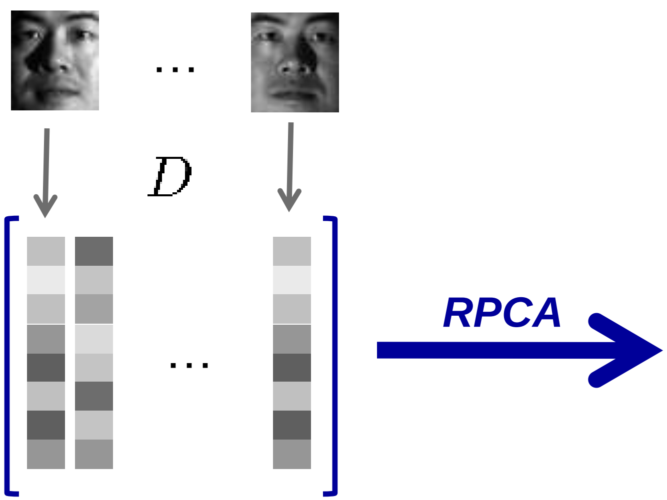
Low-rank Method | Photoshop

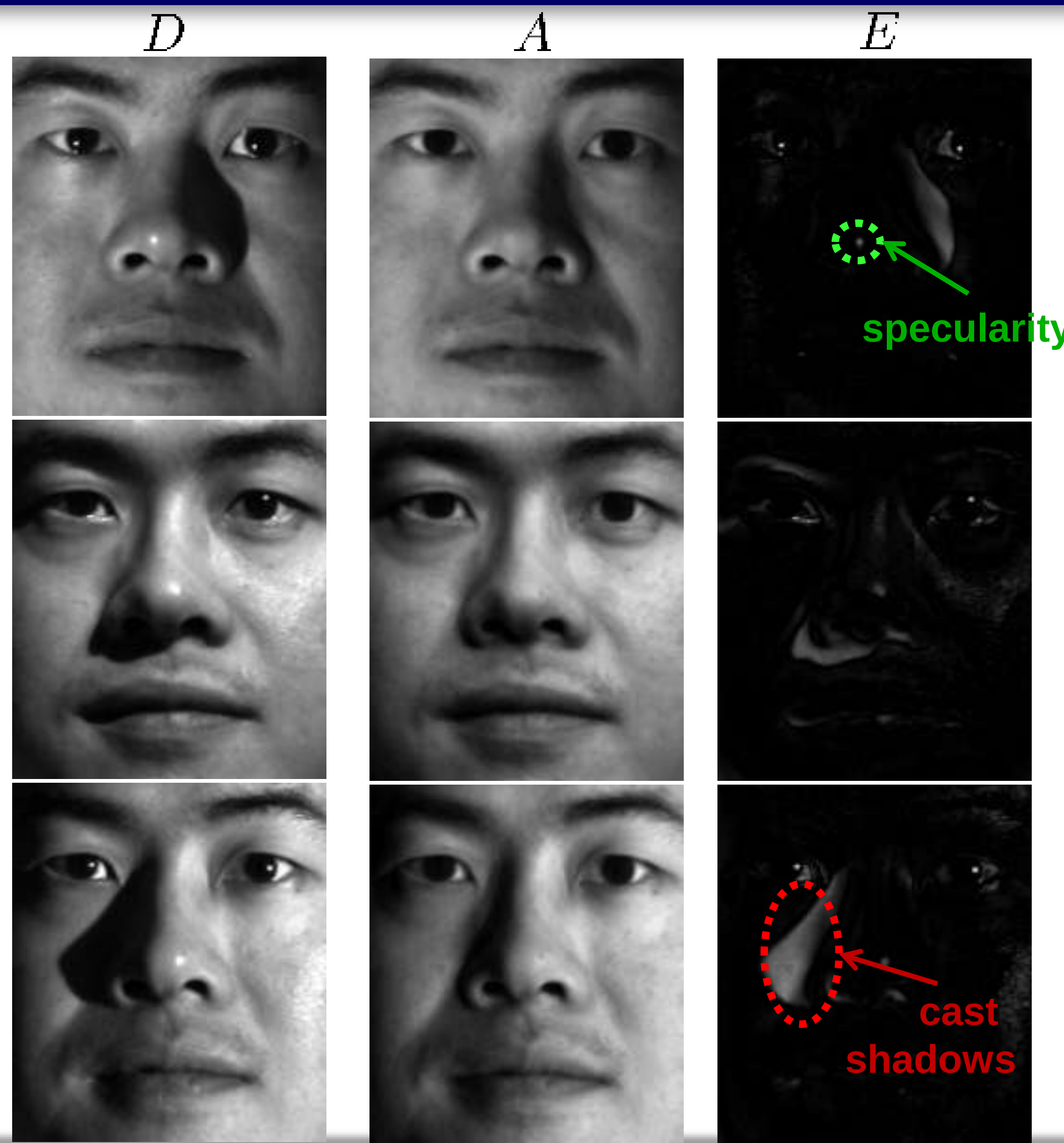
Input

Output

Repairing Multiple Correlated Images

58 images of one person
under varying lighting:

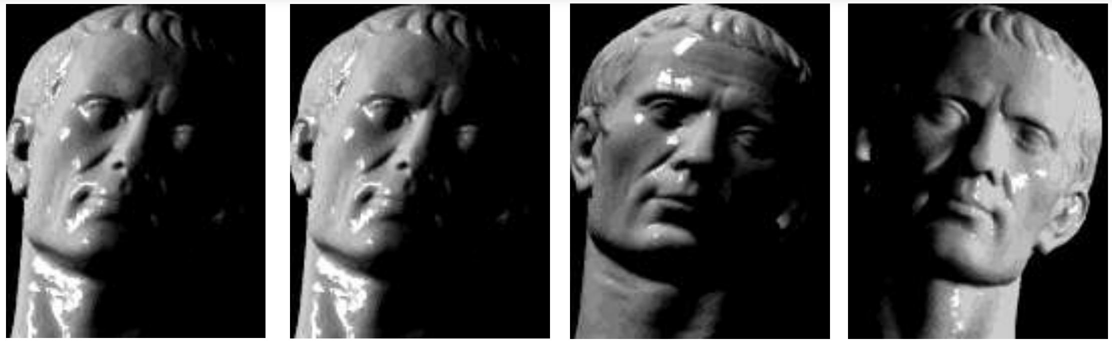




Candes, Li, Ma, and Wright, Journal of the ACM, May 2011.

Repairing Images: robust photometric stereo

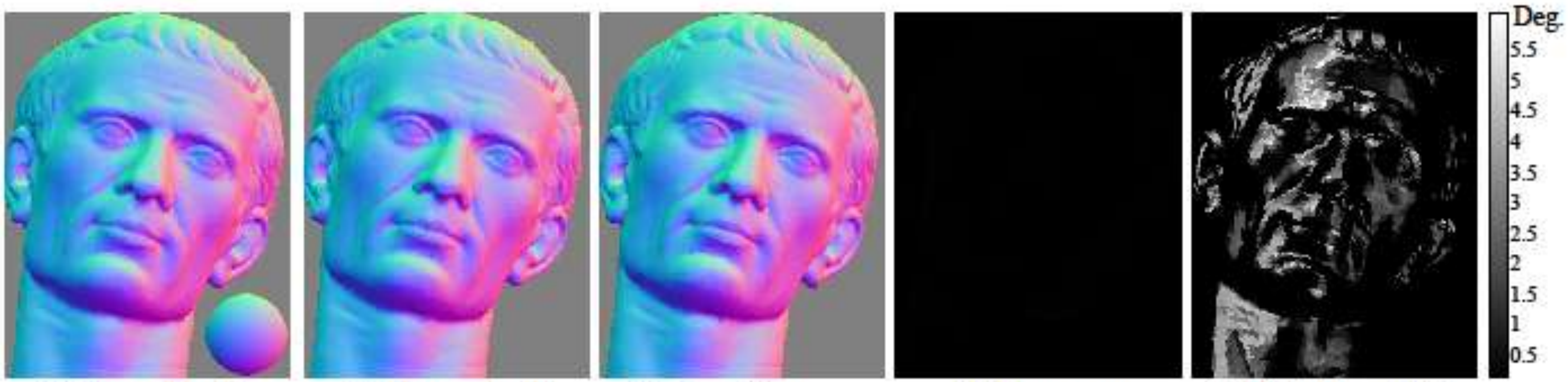
Input images



$$\min \ \|A\|_* + \lambda\|E\|_1 \quad \text{subj} \quad D = \mathcal{P}_\Omega(A+E).$$

$\Omega^c \sim \text{shadow}(20.7\%)$

$E \sim \text{specularities}(13.6\%)$



(a) Ground truth (b) Our method (c) Least Squares (d) Error map (our method) (e) Error map (LS)

Mean error	**0.014º**	0.96º
Max error	**0.20º**	8.0º

Surveillance video

200 frames,
144 x 172 pixels,

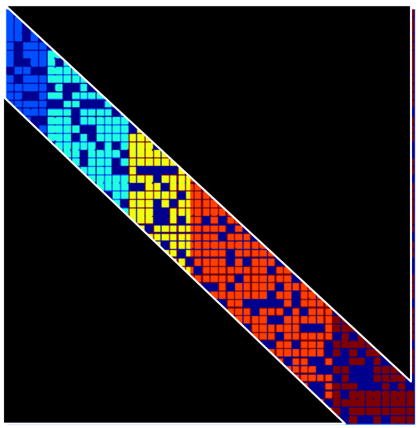
Significant foreground motion

Video D = Low-rank appx. A + Sparse error E

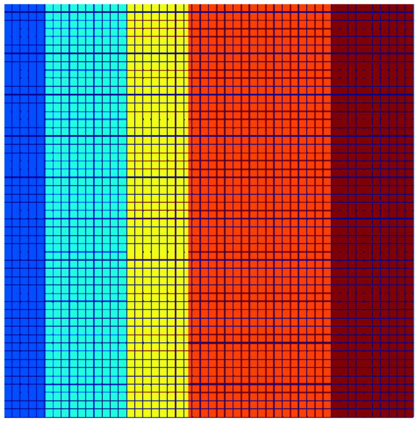
Implications: Highly Compressive Sensing of Structured Information!

Recover low-dimensional structures from diminishing fraction of corrupted measurements.

compressive samples

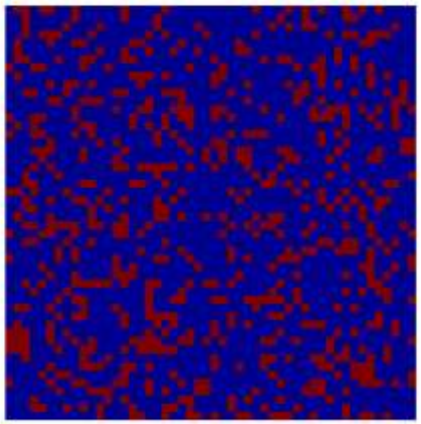


Low-rank Structures

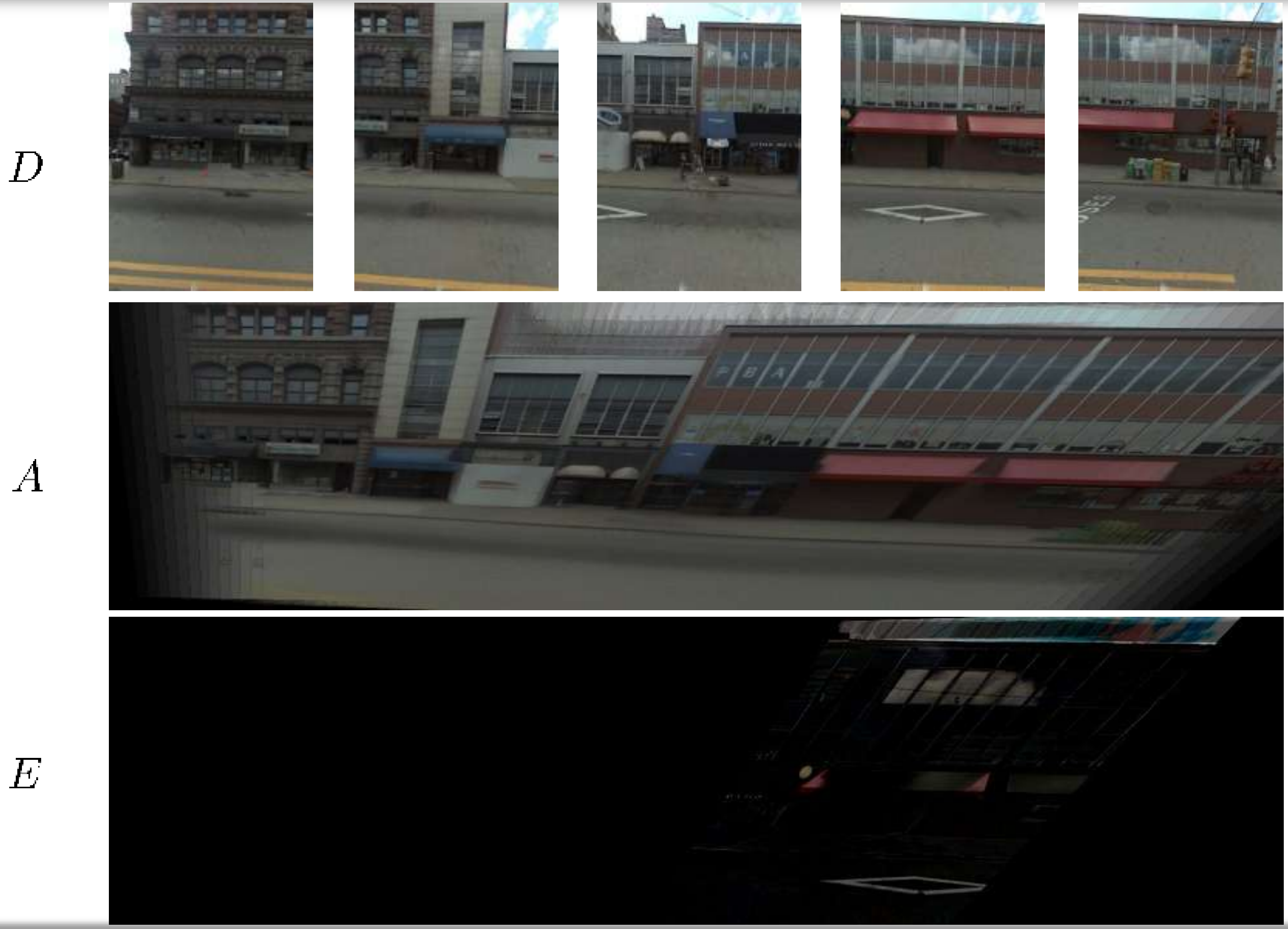


+

Sparse Structures



Repairing Video Frames: *Street Panorama*



Repairing Video Frames: Street Panorama

Low-rank

AutoStitch

Photoshop

Repairing Video Frames: Street Panorama

Low-rank

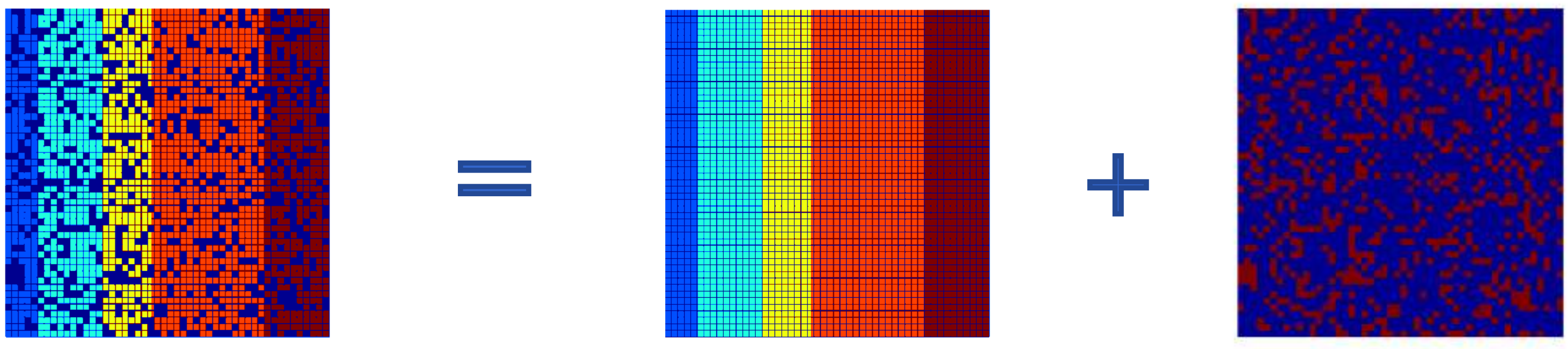
AutoStitch

Photoshop

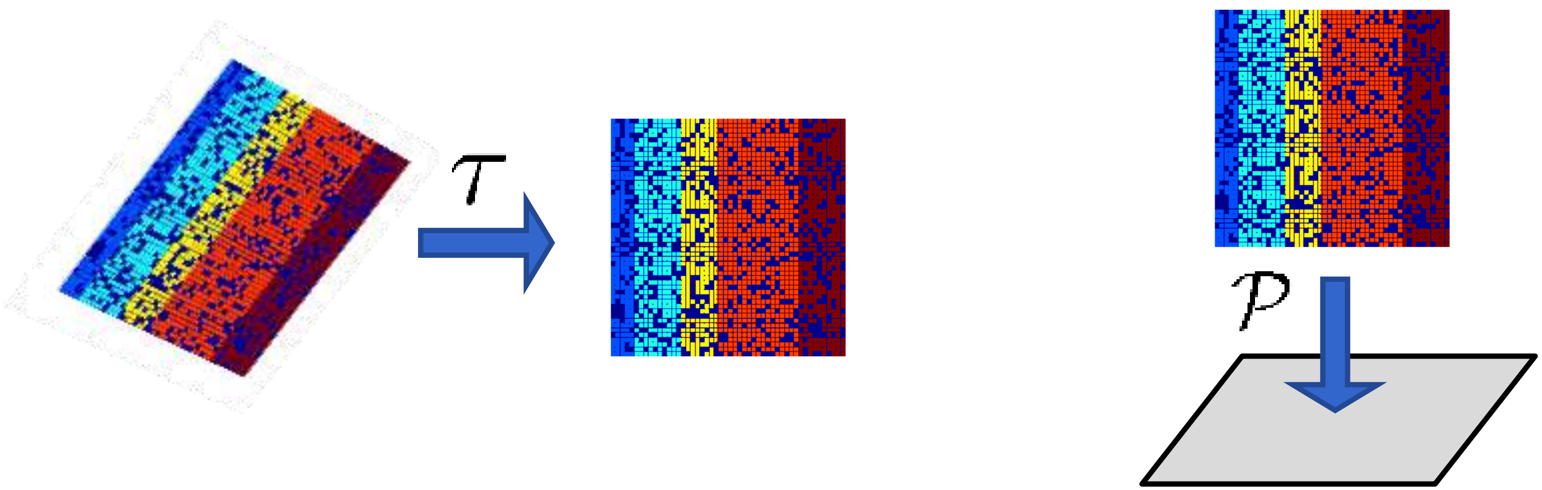
Sensing or Imaging of Low-rank and Sparse Structures

Fundamental Problem: ***How to recover low-rank and sparse structures from corrupted data***

Low-rank Structures　　　Sparse Structures

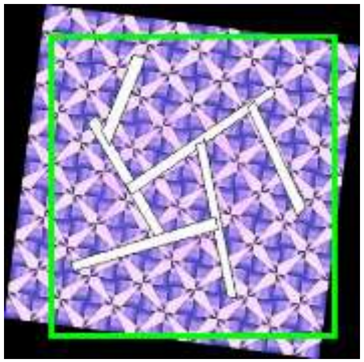


subject to either nonlinear deformation τ or linear compressive sampling $\mathcal{P}$?



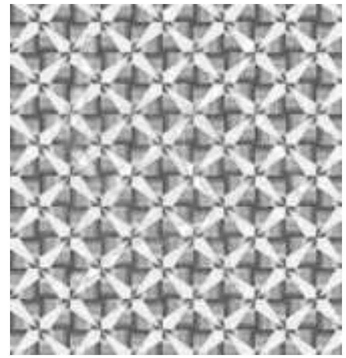
Reconstructing 3D Geometry and Structures

D – deformed observation



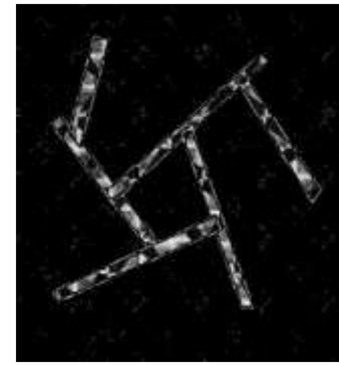
$\circ\, \tau \; =$

A – low-rank structures



$+$

E – sparse errors



Problem: Given $D \circ \tau = A_0 + E_0$, recover τ, A_0 and E_0 simultaneously.

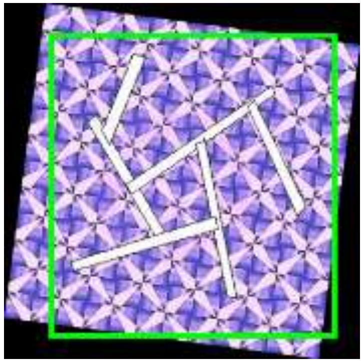
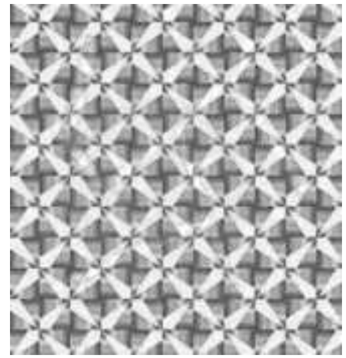
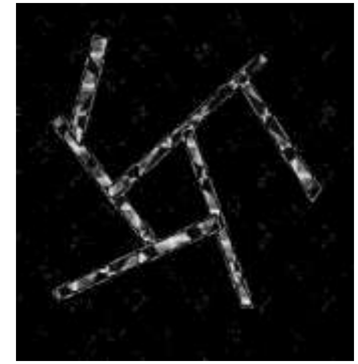
Low-rank component
(regular patterns…)

Sparse component
(occlusion, corruption, foreground…)

Parametric deformations
(affine, projective, radial distortion, 3D shape…)

Transform Invariant Low-rank Textures (TILT)

D – deformed observation | A – low-rank structures | E – sparse errors

 $\circ\, \tau \; =$  $+$ 

Objective: *Transformed Principal Component Pursuit:*

$$\min \ \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \tau$$

Solution: *Iteratively solving the linearized convex program:*

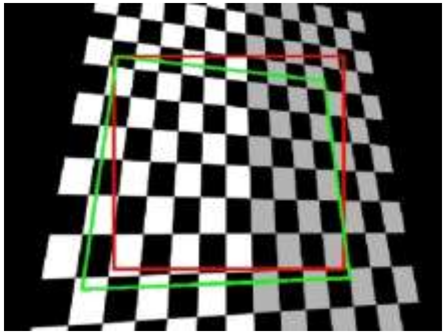
$$\min \ \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \tau_k + J \cdot \Delta\tau$$

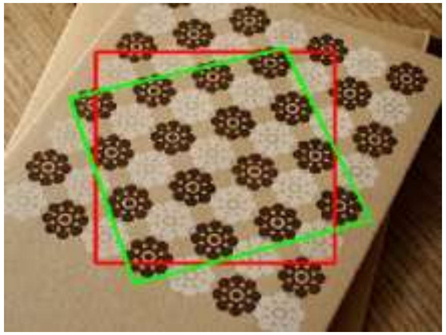
Or reduced version: $\quad \text{subj} \quad \mathcal{P}_Q[A + E] = \mathcal{P}_Q[D \circ \tau_k], \ \mathcal{P}_Q[J] = 0$

Zhang, Liang, Ganesh, Ma, ACCV'10, IJCV'12

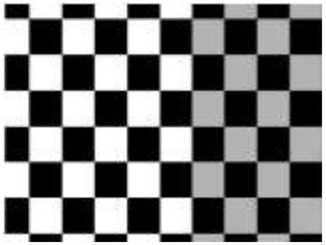
TILT: *Shape from texture*

Input (red window D)



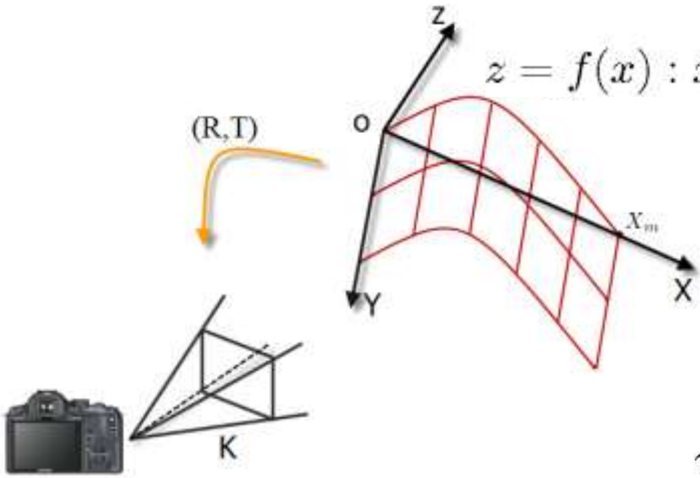


Output (rectified green window A)



Zhang, Liang, Ganesh, Ma, ACCV’10, IJCV’12

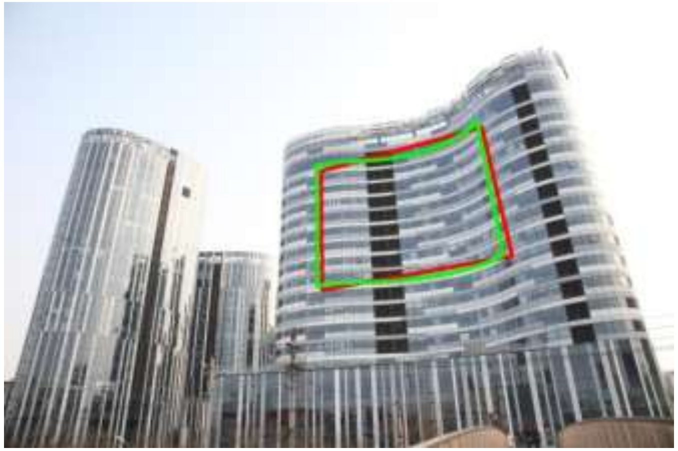
TILT: *Shape and geometry from textures*



$$z = f(x) : x(x - x_m) \sum_{i=1}^{d} a_i x^i$$

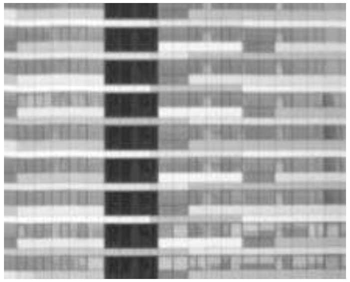
$$D \circ \tau = A + E$$

$$\tau = (K, R, T, \{a_i\})$$

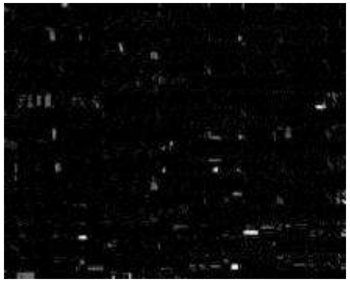


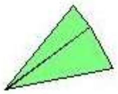
D

A



E

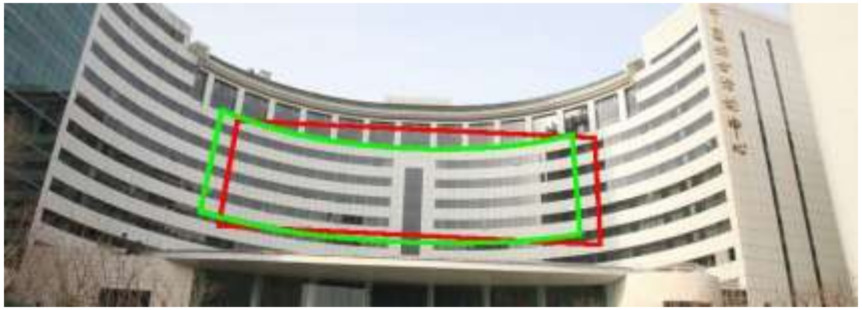




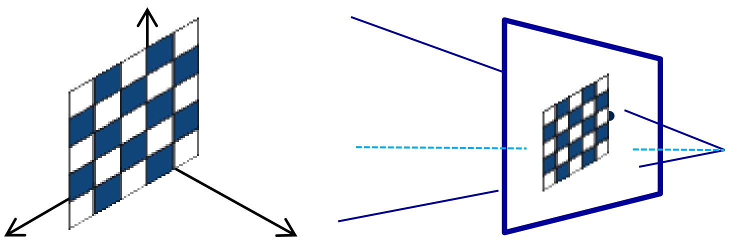
TILT: *Shape and geometry from textures*

360º panorama

TILT: *Virtual reality*



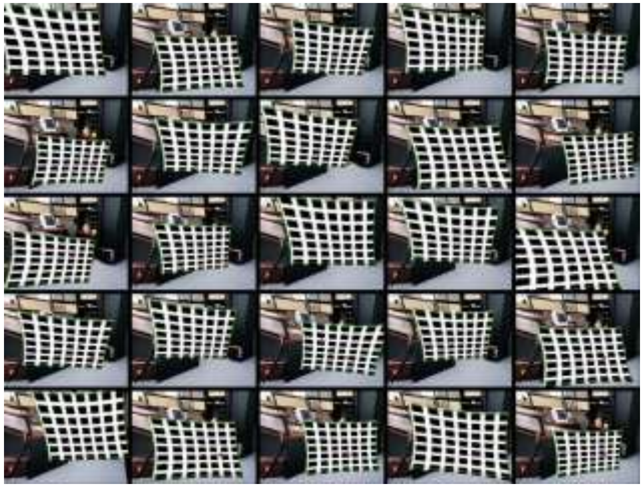
TILT: *Camera Calibration with Radial Distortion*

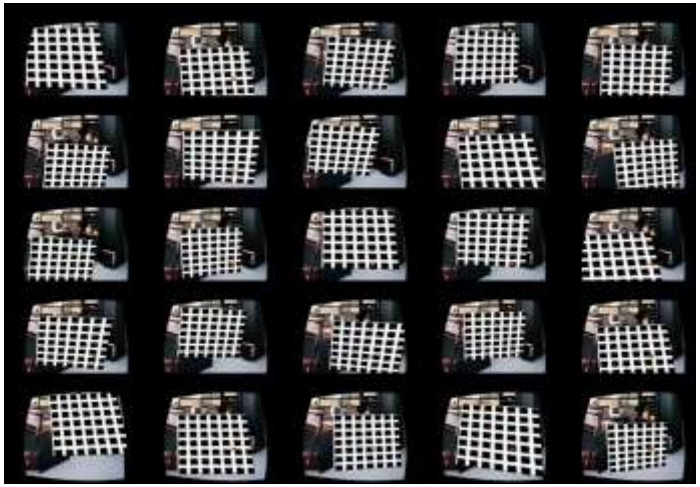


$$r = \sqrt{x_0^2 + y_0^2}, f(r) = 1 + kc(1)r^2 + kc(2)r^4 + kc(5)r^6$$

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f(r)x_0 + 2kc(3)x_0y_0 + kc(4)(r^2 + 2x_0^2) \\ f(r)y_0 + 2kc(4)x_0y_0 + kc(3)(r^2 + 2y_0^2) \end{pmatrix}$$

$$K = \begin{bmatrix} f_x & \theta & o_x \\ 0 & f_y & o_y \\ 0 & 0 & 1 \end{bmatrix}$$





TILT: *Camera Calibration with Radial Distortion*

$$\min \sum_{i=1}^{N} \|A_i\|_* + \lambda \|E_i\|_1 \quad \text{subj} \quad A_i + E_i = D \circ (\tau_0, \tau_i)$$

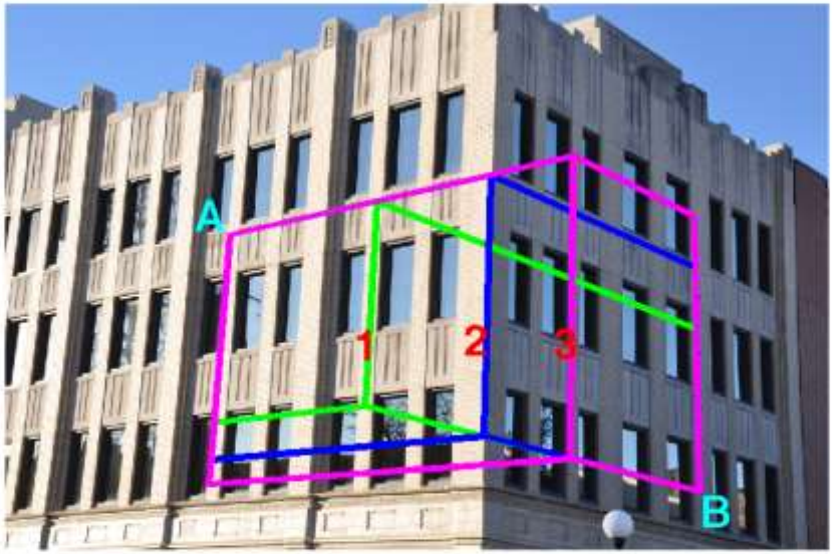
$$\tau_0 = (K, K_c), \quad \tau_i = (R_i, T_i).$$

Previous approach

Low-rank method

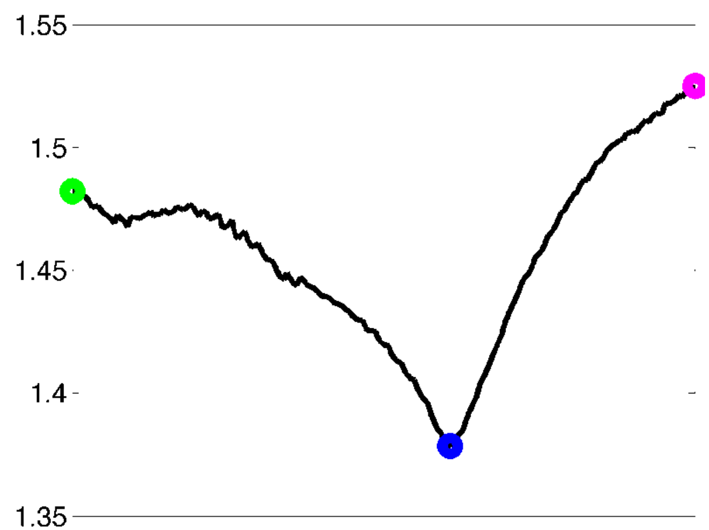
Zhang, Matsushita, and Ma, in CVPR 2011

TILT: *Holistic 3D Reconstruction of Urban Scenes*



$$\min \|A\|_* + \|E\|_1 \quad \text{s.t.}$$

$$A + E = [D_1 \circ \tau_1, D_2 \circ \tau_2]$$

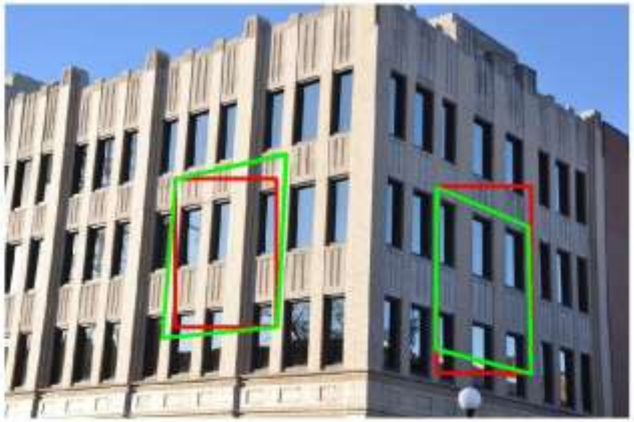


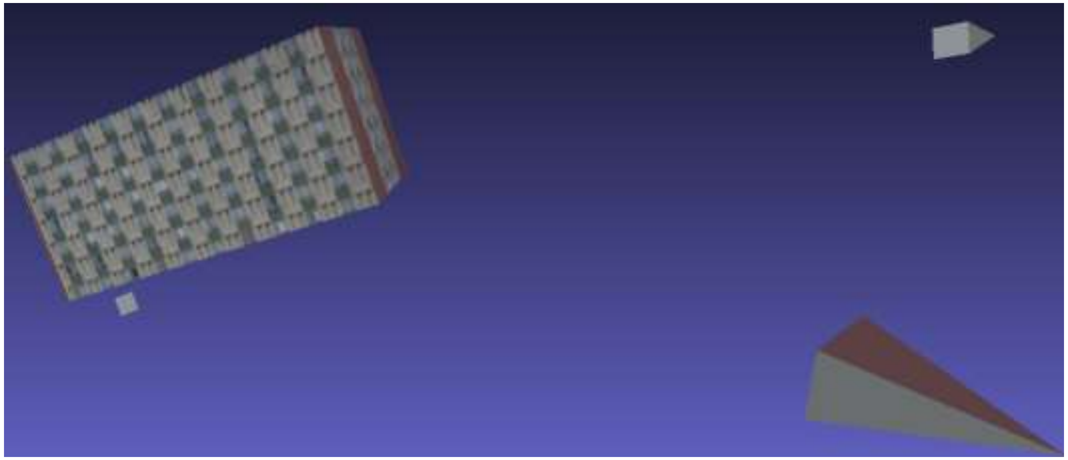
Mobahi, Zhou, and Ma, in ICCV 2011

TILT: *Holistic 3D Reconstruction of Urban Scenes*

From one input image

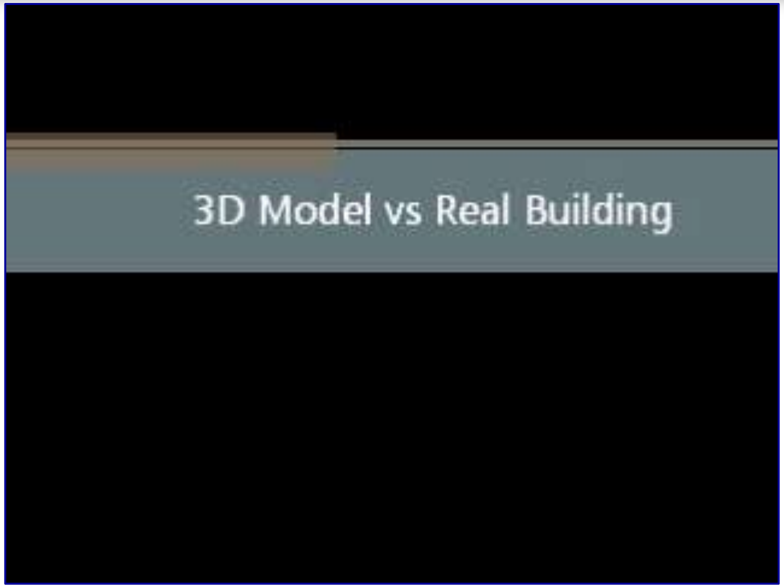
From four input images

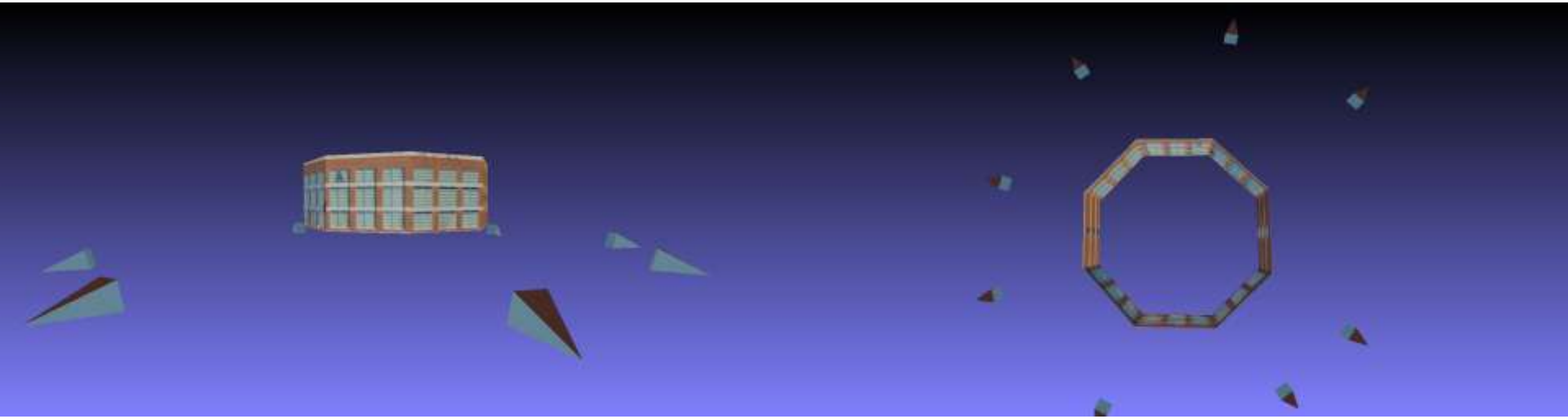




TILT: *Holistic 3D Reconstruction of Urban Scenes*

From eight input images





Virtual reality in urban scenes

Repairing Distorted Low-rank Textures

Low-rank Method

Photoshop

Input

Output

Repair Distorted Low-rank Textures

Liang, Ren, Zhang, and Ma, *Repairing Sparse Low-Rank Texture*, in ECCV 2012.

Registering Multiple Images: *Robust Alignment*

D – corrupted & misaligned observation

A – aligned low-rank signals

E – sparse errors

Problem: Given $D \circ \tau = A_0 + E_0$, recover τ , A_0 and E_0.

Parametric deformations (rigid, affine, projective…)

Low-rank component

Sparse component

Solution: Robust Alignment via Low-rank and Sparse (**RASL**) Decomposition

Iteratively solving the linearized convex program:

$$\min \ \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \tau_k + J\Delta\tau$$
$$(\text{or} \quad Q(A+E) = QD \circ \tau_k, \ QJ = 0)$$

RASL: *Aligning Face Images from the Internet*

*48 images collected from internet

Peng, Ganesh, Wright, Ma, CVPR'10, TPAMI'11

RASL: *Faces Detected*

Input: faces detected by a face detector (D)

Average

RASL: *Faces Aligned*

Output: aligned faces ($D \circ \tau$)

Average

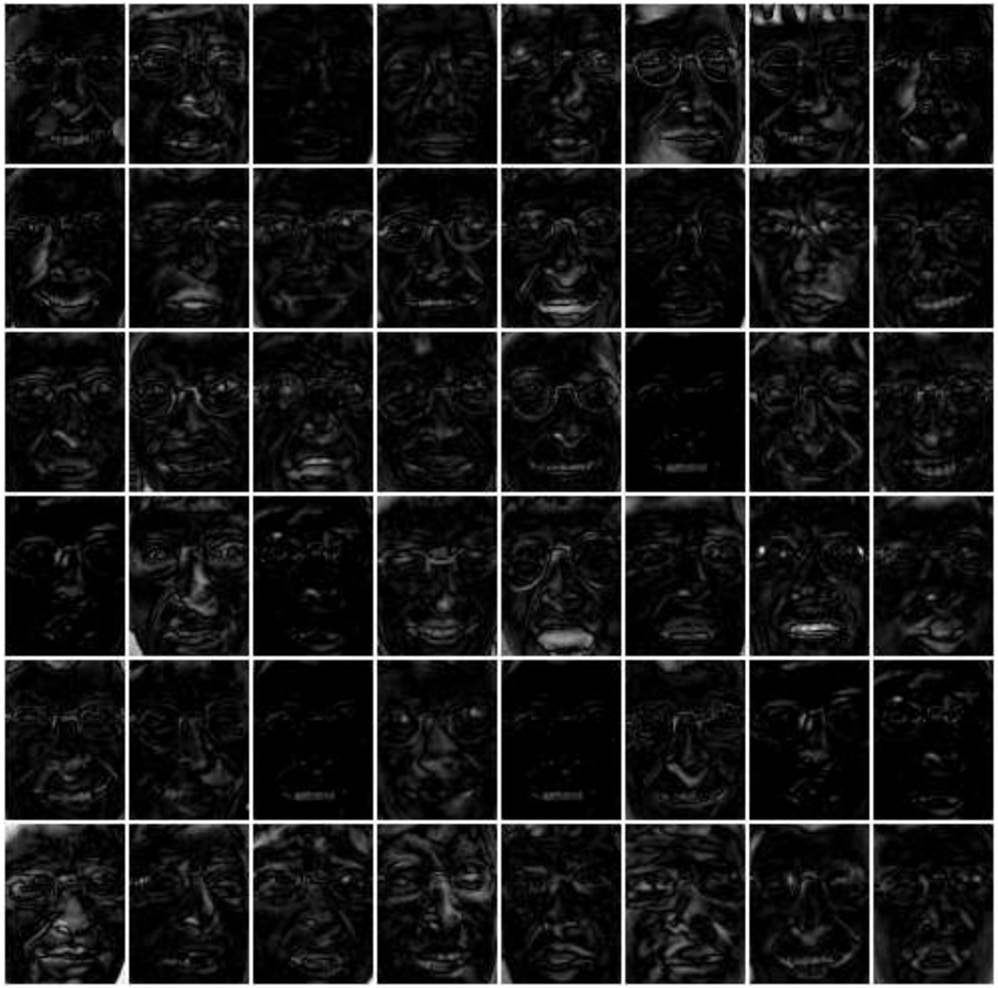
RASL: *Faces Repaired and Cleaned*

Output: clean low-rank faces (A)

Average

RASL: *Sparse Errors of the Face Images*

Output: sparse error images (E)



RASL: *Video Stabilization and Enhancement*

Original video (D) Aligned video ($D \circ \tau$) Low-rank part (A) Sparse part (E)

Peng, Ganesh, Wright, Ma, CVPR'10, TPAMI'11

RASL: *Aligning Handwritten Digits*

D

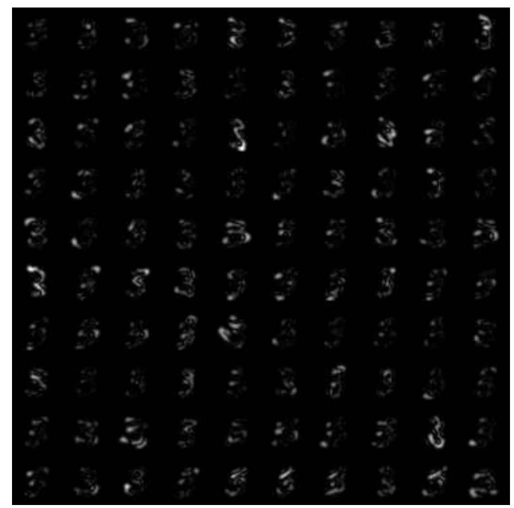
Learned-Miller PAMI'06

Vedaldi CVPR'08

$D \circ \tau$

A

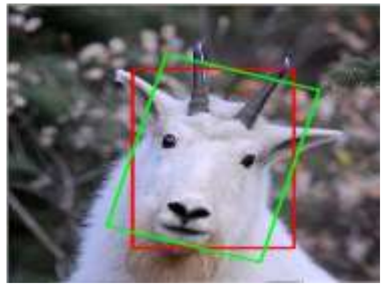
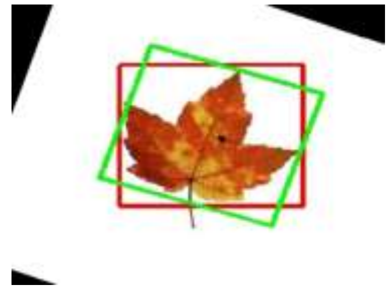
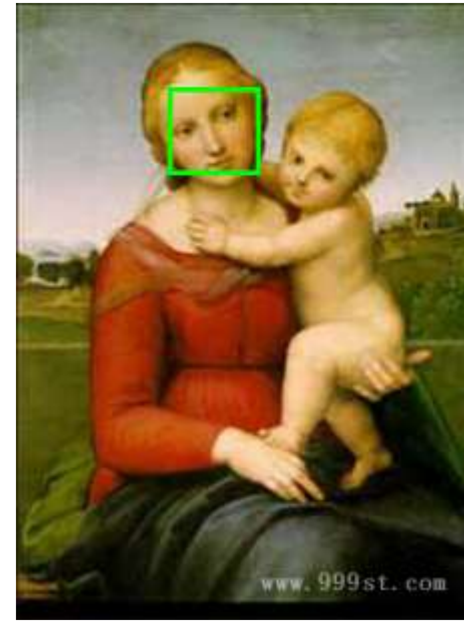
E



Peng, Ganesh, Wright, Ma, CVPR'10, TPAMI'11

Object Recognition: *Rectifying Pose of Objects*

Input (red window D)

Output (rectified green window A)

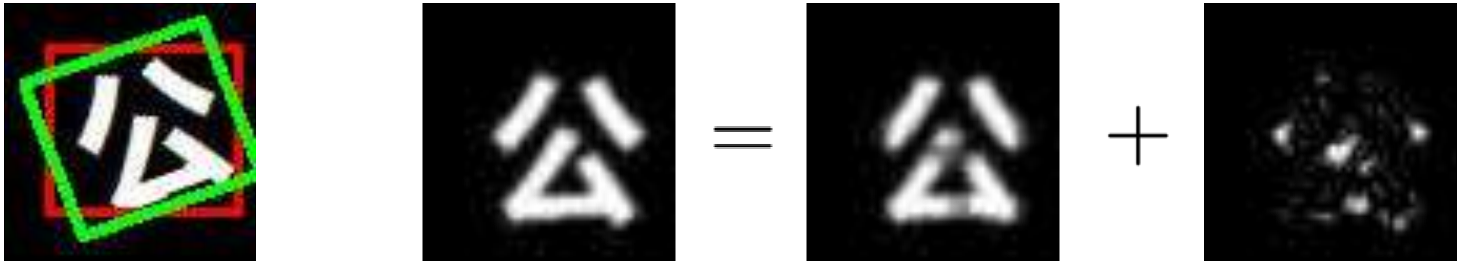



Object Recognition: *Regularity of Texts at All Scales!*

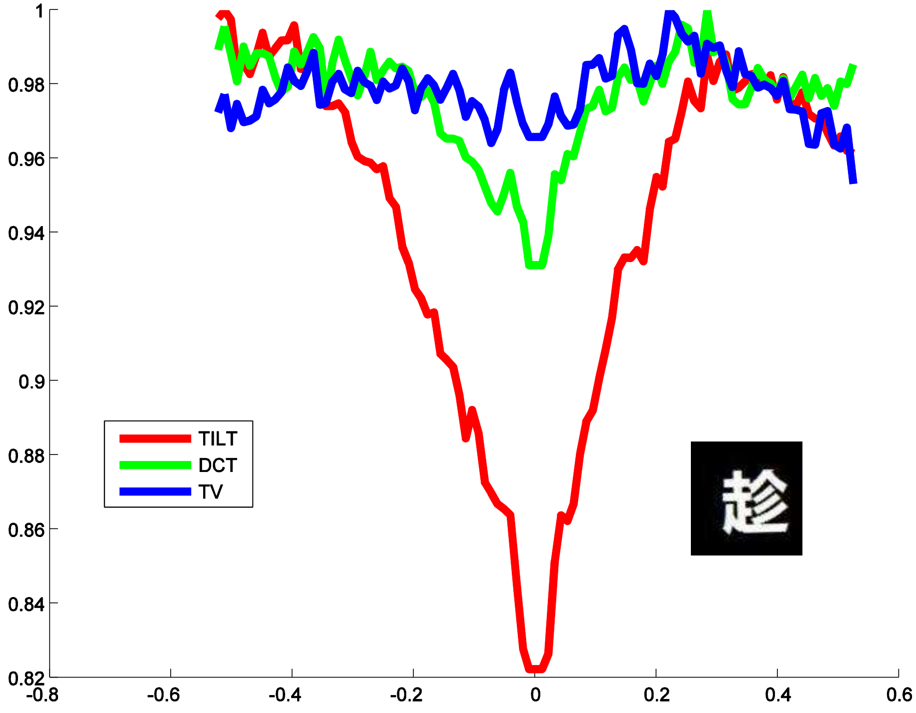
Input (red window D)

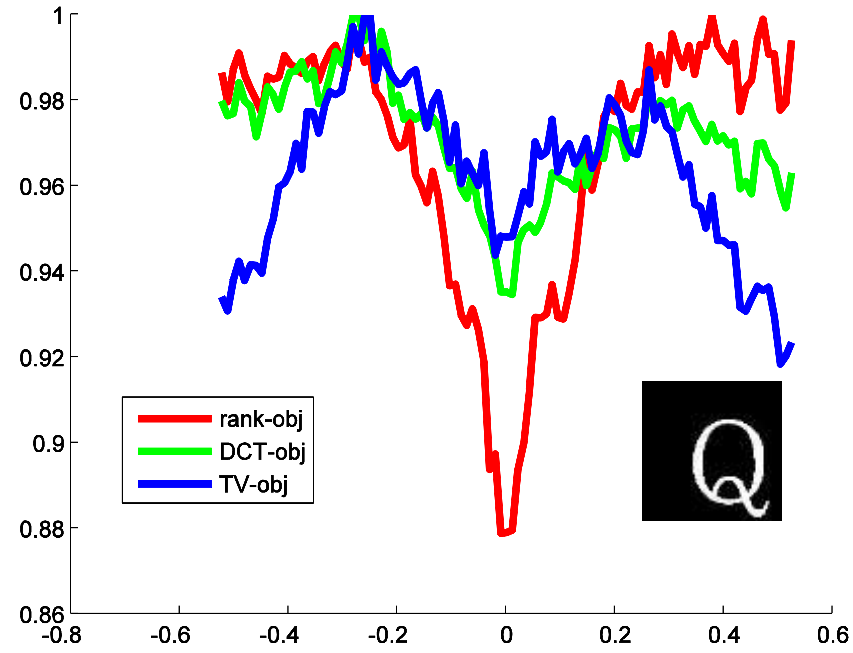
Output (rectified green window A)

Recognition: *Character/Text Rectification*



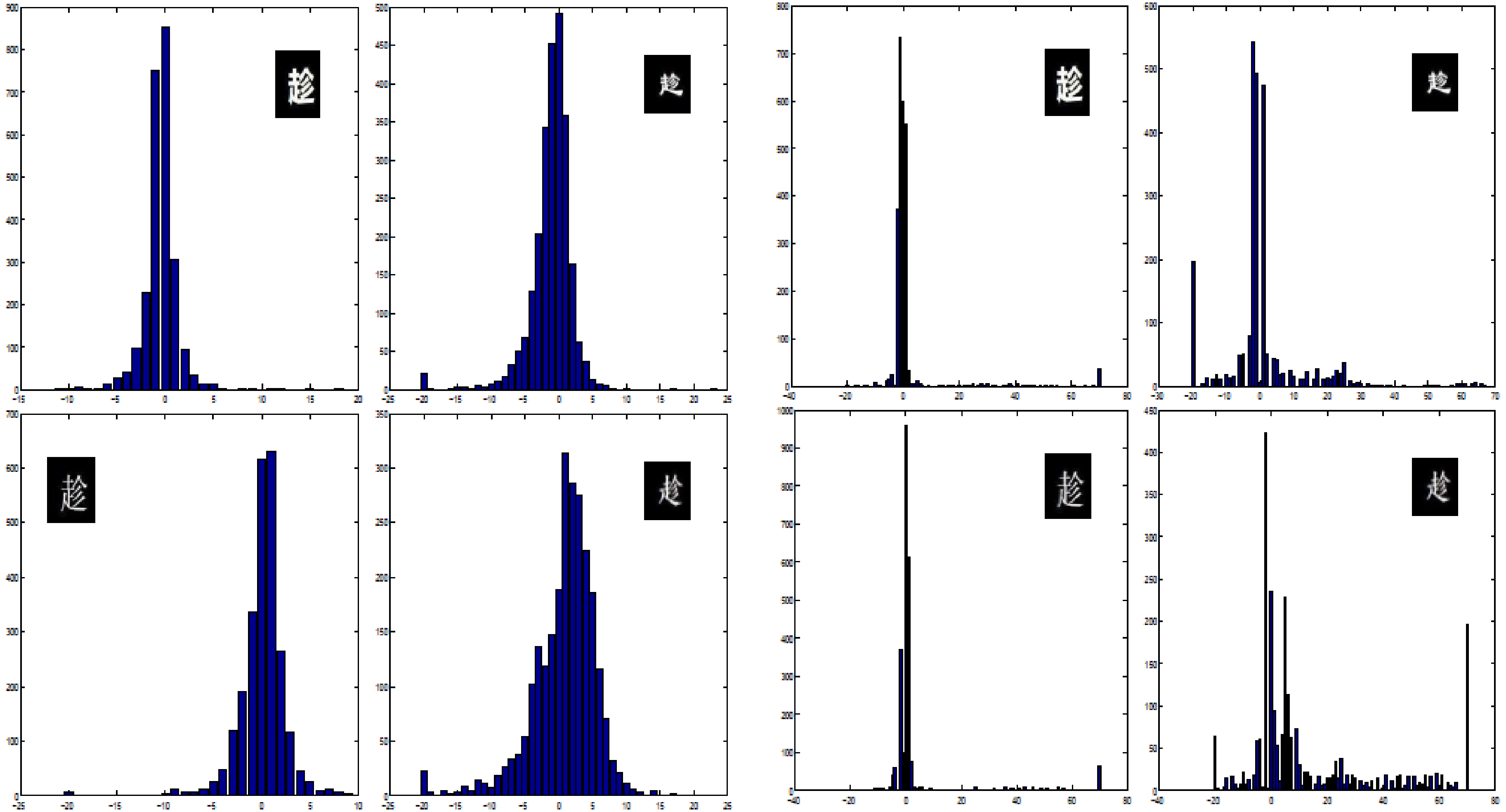
D $D \circ \tau$ A E



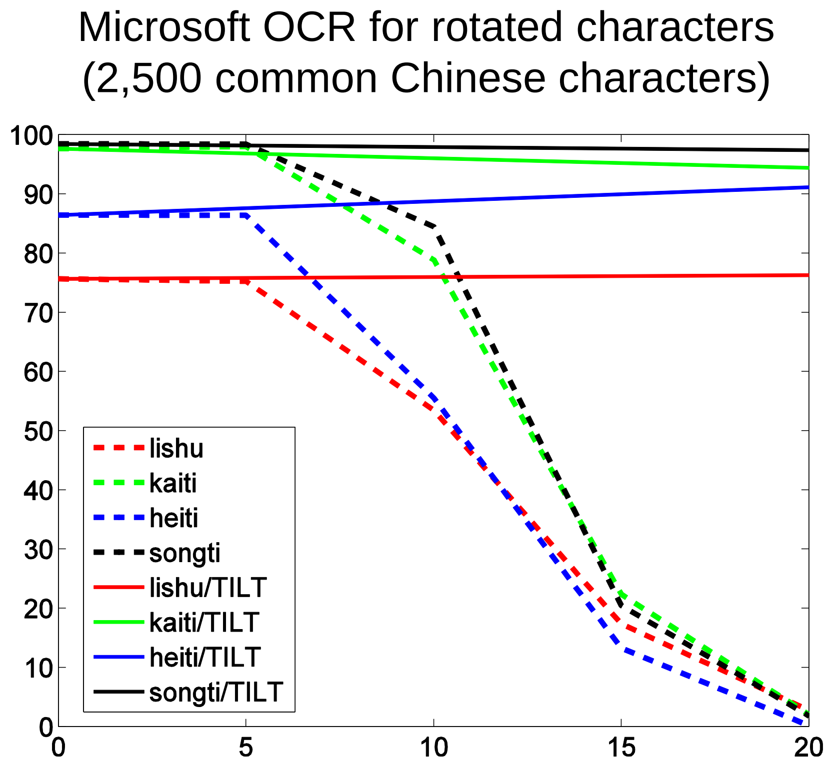


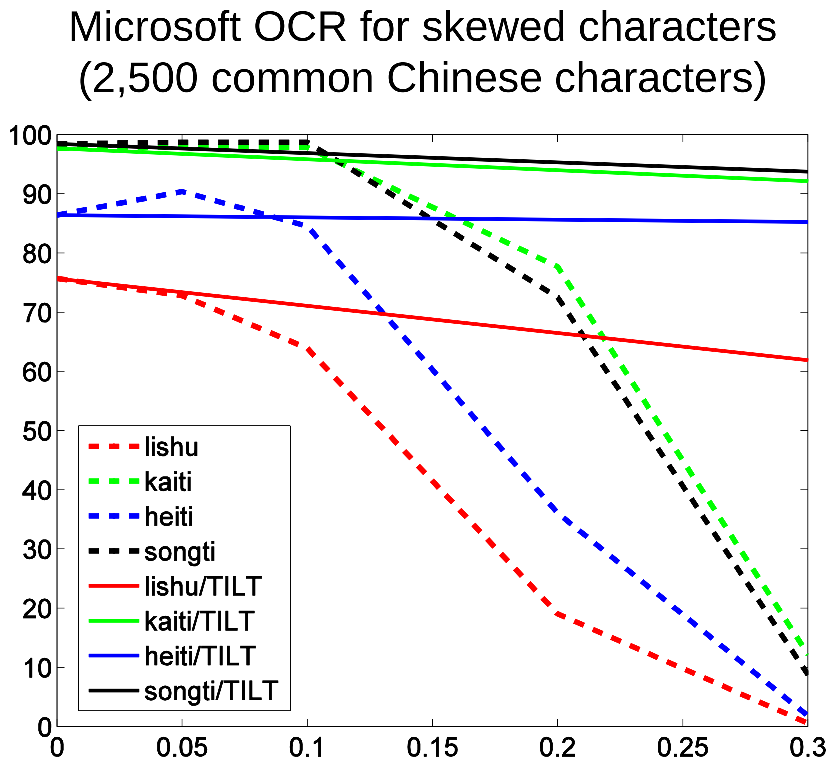
Recognition: *Character/Text Rectification*

TILT versus Hough Transform



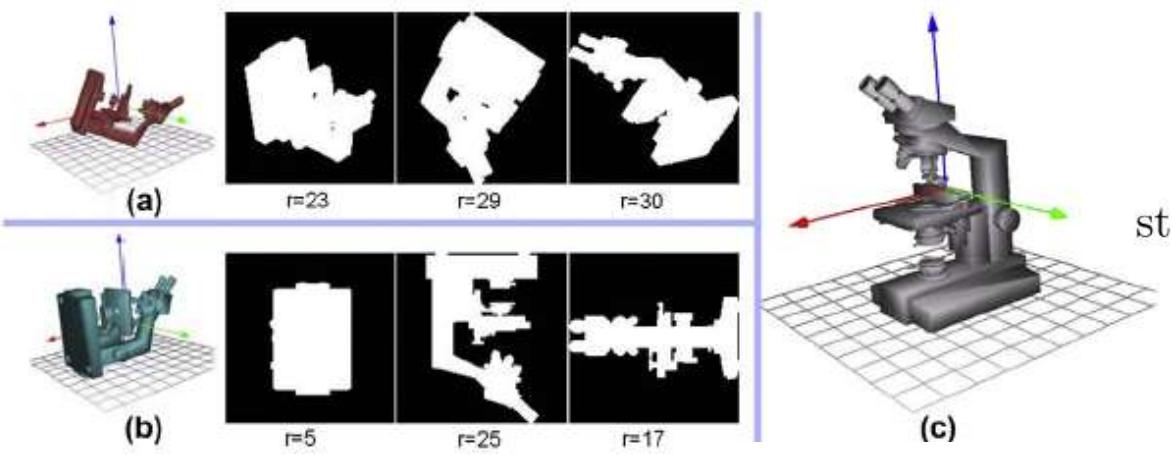
Recognition: Character Rectification and Recognition





Recognition: *Upright orientation of man-made objects*

TILT for 3D: Unsupervised upright orientation of man-made 3D objects



$$\min \sum_{i=1}^{3} \|A_i\|_* + \lambda \|E_i\|_1$$

$$\text{st } D \circ \tau = [A_1, A_2, A_3] + [E_1, E_2, E_3].$$

Fig. 10. More models which have been successfully tested through our algorithm.

Take-home Messages for Visual Data Analysis:

1. (Transformed) **low-rank and sparse** structures are central to visual data modeling, processing, and analyzing;

2. Such structures can now be extracted **correctly, robustly, and efficiently**, from raw image pixels (or high-dim features);

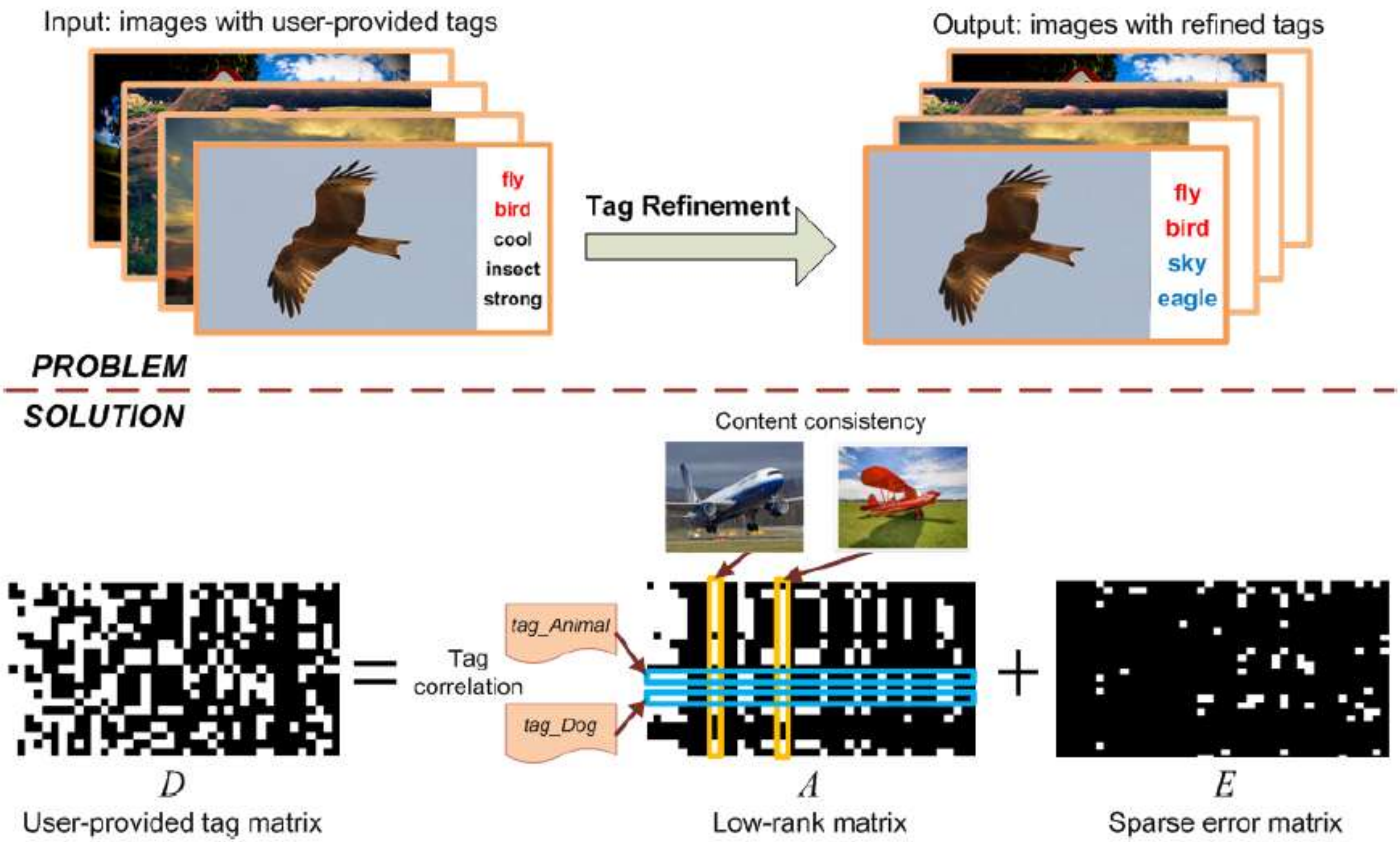
3. These new algorithms **unleash tremendous local or global information** from single or multiple images, emulating or surpassing human capability;

4. These algorithms start to exert significant impact on **image/video processing, 3D reconstruction, and object recognition**.

… …

But try not to abuse or misuse them…

Other Data/Applications: Web Image/Tag Refinement



Other Data/Applications: Web Document Corpus Analysis

Latent Semantic Indexing: the classical solution (PCA)

Documents

Words

CHRYSLER SETS STOCK SPLIT, HIGHER DIVIDEND

Chrysler Corp said its board declared a three-for-two stock split in the form of a 50 pct stock dividend and raised the quarterly dividend by seven pct.

The company said the dividend was raised to 37.5 cts a share from 35 cts on a pre-split basis, equal to a 25 ct dividend on a post-split basis.

Chrysler said the stock dividend is payable April 13 to holders of record March 23 while the cash dividend is payable April 15 to holders of record March 23. It said cash will be paid in lieu of fractional shares.

With the split, Chrysler said 13.2 mln shares remain to be purchased in its stock repurchase program that began in late 1984. That program now has a target of 56.3 mln shares with the latest stock split.

Chrysler said in a statement the actions "reflect not only our outstanding performance over the past few years but also our optimism about the company's future."

D

d_{ij} word frequency (or TF/IDF)

$$D = A + Z$$
$$= U_1 \Sigma_1 V_1^T + \underline{U_2 \Sigma_2 V_2^T}$$

Dense, difficult to interpret

a better model/solution?

$$D = A + \underline{E}$$

Low-rank "background" topic model

Informative, discriminative "keywords"

Other Data/Applications: Sparse Keywords Extracted

Reuters-21578 dataset: 1,000 longest documents; 3,000 most frequent words

CHRYSLER SETS STOCK SPLIT, HIGHER DIVIDEND

Chrysler Corp said its board declared a three-for-two stock split in the form of a 50 pct stock dividend and raised the quarterly dividend by seven pct.

The company said the dividend was raised to 37.5 cts a share from 35 cts on a pre-split basis, equal to a 25 ct dividend on a post-split basis.

Chrysler said the stock dividend is payable April 13 to holders of record March 23 while the cash dividend is payable April 15 to holders of record March 23. It said cash will be paid in lieu of fractional shares.

With the split, Chrysler said 13.2 mln shares remain to be purchased in its stock repurchase program that began in late 1984. That program now has a target of 56.3 mln shares with the latest stock split.

Chrysler said in a statement the actions "reflect not only our outstanding performance over the past few years but also our optimism about the company's future."

Other Data/Applications: Protein-Gene Correlation

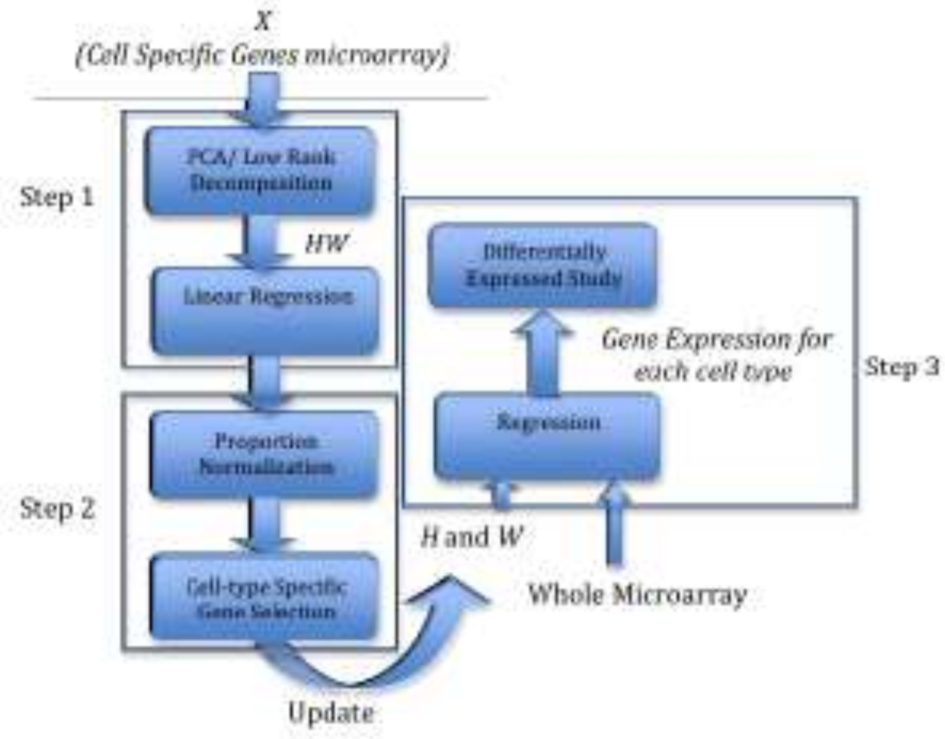


Fig. 1. The diagram of the workflow of the method presented in this paper

Microarray data

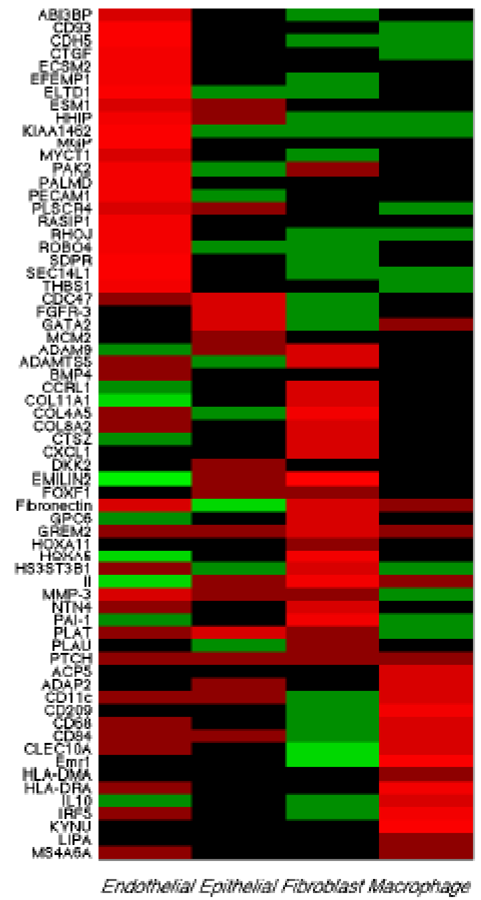
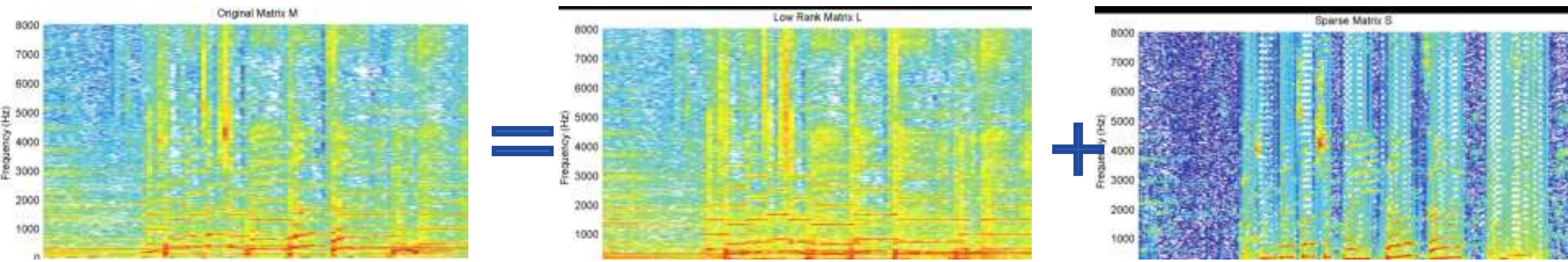
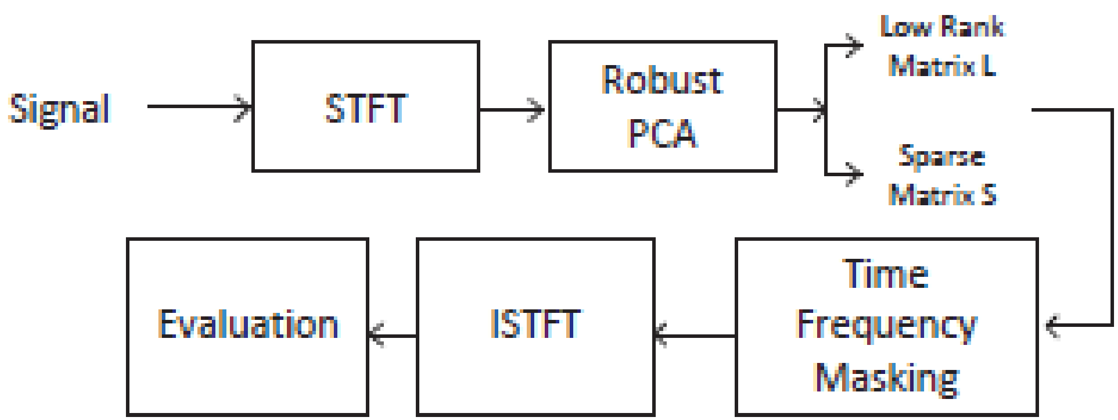


Fig. 6. HeatMap of estimated gene signatures for the sorted cell specific genes after adjustments based on fold changes. RPCA is used in the first step. It is clear that this matrix is close to a block diagonal structure.

Other Data/Applications: Lyrics and Music Separation

Songs (STFT) Low-rank (music) Sparse (voices)





Po-Sen Huang, Scott Chen, Paris Smaragdis, Mark Hasegawa-Johnson, ICASSP 2012.

Other Data/Applications: Internet Traffic Anomalies

Network Traffic = Normal Traffic + Sparse Anomalies + Noise

$$D = L + RS + N$$

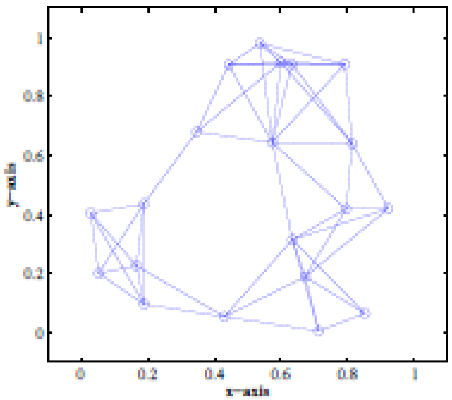
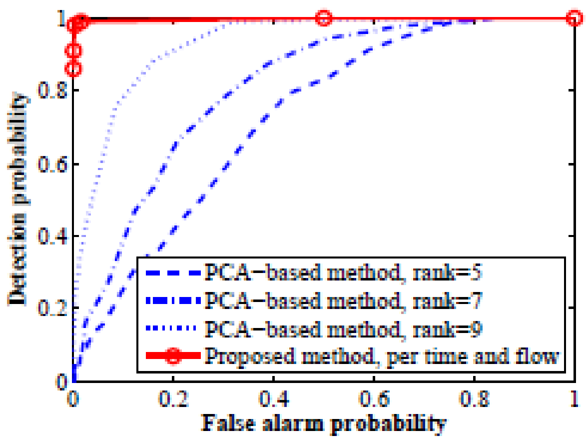
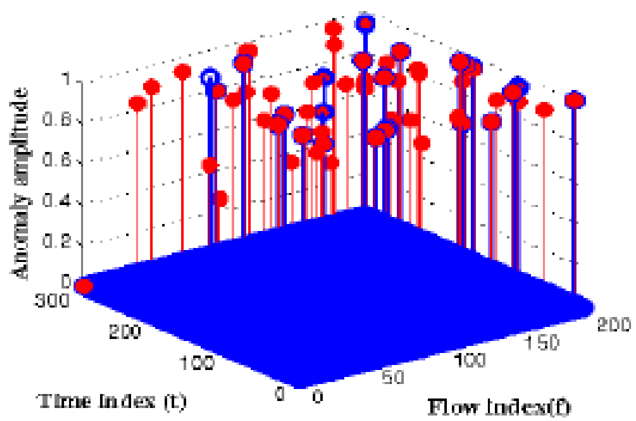


Fig. 2. Network topology graph.





Other Data/Applications: Robust Filtering and System ID

GPS on a Car:

$$\begin{cases} \dot{x} &= Ax + Bu, \quad A \in \Re^{r\times r} \\ y &= Cx + z + \color{red}{e} \end{cases}$$

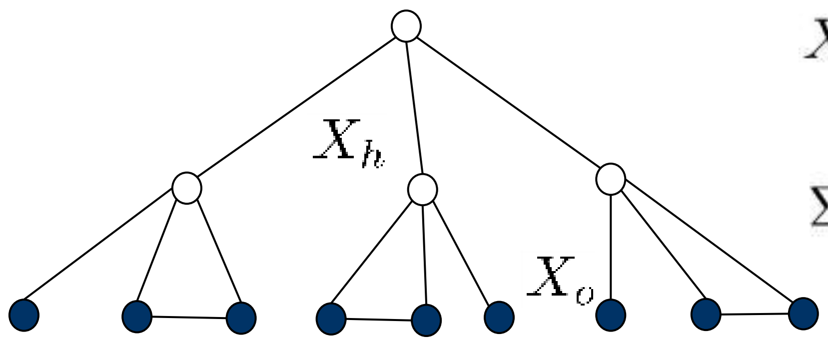
gross sparse errors
(due to buildings, trees…)

Robust Kalman Filter: $\hat{x}_{t+1} = Ax_t + K(y_t - C\hat{x}_t)$

Robust System ID:

$$\underbrace{\begin{bmatrix} y_n & y_{n-1} & y_{n-2} & \cdots & y_0 \\ y_{n-1} & y_{n-2} & \cdots & \ddots & y_{-1} \\ y_{n-2} & \cdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & y_{-n+2} \\ y_0 & y_{-1} & \cdots & y_{-n+2} & y_{-n+1} \end{bmatrix}}_{\text{Hankel matrix}} = \mathcal{O}_{n\times r} X_{r\times n} + S$$

Other Data/Applications: Learning Graphical Models



$$X = (X_o, X_h) \sim \mathcal{N}(0, \Sigma)$$

$$\Sigma = \begin{bmatrix} \Sigma_o & \Sigma_{oh} \\ \Sigma_{ho} & \Sigma_h \end{bmatrix} \Rightarrow \Sigma^{-1} = \begin{bmatrix} J_o & J_{oh} \\ J_{ho} & J_h \end{bmatrix}$$

$X_i, \; X_j$ cond. indep. given other variables $\Leftrightarrow$ $\left(\Sigma^{-1}\right)_{ij} = 0$

Separation Principle:

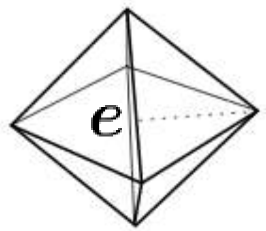
$$\begin{array}{ccccc} \Sigma_o^{-1} & = & J_o & - & J_{oh} J_h^{-1} J_{ho} \\ \text{observed} & = & \text{sparse} & + & \text{low-rank} \end{array}$$

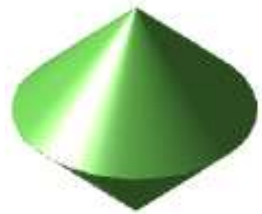
- sparse pattern → conditional (in)dependence
- rank of second component → number of hidden variables

CONCLUSIONS – *A Unified Theory for Sparsity and Low-Rank*

	Sparse Vector	***Low-Rank Matrix***
Low-dimensionality of	individual signal	correlated signals
Measure	L_0 norm $\|\|x\|\|_0$	$\mathrm{rank}(X)$
Convex Surrogate	L_1 norm $\|\|x\|\|_1$	Nuclear norm $\|\|X\|\|_*$
Compressed Sensing	$y = Ax$	$Y = A(X)$
Error Correction	$y = Ax + e$	$Y = A(X) + E$
Domain Transform	$y \circ \tau = Ax + e$	$Y \circ \tau = A(X) + E$
Mixed Structures	$Y = A(X) + B(E) + Z$	

Broader Family of Low-Dimensional Structures

$\boxed{\quad L \quad}\, x +$ 



A norm $\|\cdot\|$ is said to be **decomposable** at $\mathbf{X}$ if there exists a subspace T and a matrix $\mathbf{S}$ such that

$$\partial\|\cdot\|(\mathbf{X}) = \{\Lambda \mid \mathcal{P}_T(\Lambda) = \mathbf{S}, \|P_{T^\perp}(\Lambda)\|^* \leq 1\},$$

where $\|\cdot\|^*$ is the dual norm of $\|\cdot\|$, and $\mathcal{P}_{T^\perp}$ is nonexpansive w.r.t. $\|\cdot\|^*$.

Theorem [Candes, Recht'11] Any low-complexity signal $\mathbf{X}^0$ can be exactly recovered from high compressive measurements via convex optimization:

$$||\mathbf{X}||_\diamond \quad \text{subject to} \quad \mathcal{P}_Q(\mathbf{X}) = \mathcal{P}_Q(\mathbf{X}^0),$$

for a decomposable norm $\|\cdot\|_\diamond$.

Jia, Chan, and Ma, *Robust Face Recognition via Structured Sparsity*, ECCV'12

Compressive Sensing and Separation of Low-dim Structures

Suppose $(\mathbf{X}_1^0, \ldots, \mathbf{X}_k^0) = \arg\min \sum_{i=1}^k \lambda_i \|\mathbf{X}_i\|_{(i)}$ subj $\sum_{i=1}^k \mathbf{X}_i = \sum_{i=1}^k \mathbf{X}_i^0$, for decomposable norms $\|\cdot\|_{(i)}$ that majorize the Frobenius norm.

Theorem 6 (Compressive Sensing of Mixed Low-Comp. Structures). *Let $Q^\perp$ be a random subspac of $\mathbb{R}^{m \times n}$ of dimension*

$$\dim(Q) \geq O(\log^2 m) \times \text{intrinsic degrees of freedomof } (\mathbf{X}_1, \ldots, \mathbf{X}_k),$$

distributed according to the Haar measure, independent of $\mathbf{X}_i$. Then with very high probability

$$(\mathbf{X}_1^0, \ldots, \mathbf{X}_k^0) = \arg\min \sum_{i=1}^k \lambda_i \|\mathbf{X}_i\|_{(i)} \quad \text{subj} \quad \mathcal{P}_Q\Big[\sum_{i=1}^k \mathbf{X}_i\Big] = \mathcal{P}_Q\Big[\sum_{i=1}^k \mathbf{X}_i^0\Big],$$

and the minimizer is unique.

A Unified THEORY – *A Suite of Powerful Regularizers*

For compressive robust recovery of a family of low-dimensional structures:

- [Bach ’10] – relaxations from submodular functions
- [Negahban+Yu+Wainwright ’10] – geometric analysis of recovery
- [Becker+Candès+Grant ’10] – algorithmic templates
- [Xu+Caramanis+Sanghavi ‘11] column sparse errors $L_{2,1}$ norm
- [Recht+Parillo+Chandrasekaran+Wilsky ’11] – compressive sensing of various structures
- [Candes+Recht ’11] – compressive sensing of decomposable structures

$$X^0 = \arg\min \|X\|_\diamond \quad \text{s.t.} \quad \mathcal{P}_Q(X) = \mathcal{P}_Q(X^0)$$

- [McCoy+Tropp’11] – separation of low-dim decomposable structures

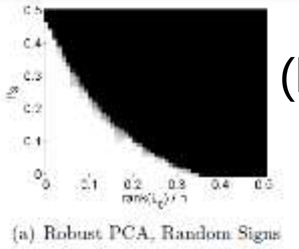
$$(X_1^0, X_2^0) = \arg\min \|X_1\|_{(1)} + \lambda \|X_2\|_{(2)} \;\; \text{s.t.} \;\; X_1 + X_2 = X_1^0 + X_2^0$$

- [Wright+Ganesh+Min+Ma, ISIT’12] – separation of superposition of decomposable structures

$$(X_1^0, \ldots, X_k^0) = \arg\min \sum \lambda_i \|X_i\|_{(i)} \;\; \text{s.t.} \;\; \mathcal{P}_Q(\textstyle\sum_i X_i) = \mathcal{P}_Q(\textstyle\sum_i X_i^0)$$

Take home message: ***Let the data and application tell you the structure…***

A Perfect Storm in the Cloud…

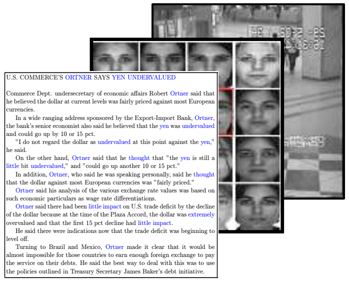


Mathematical Theory
(high-dimensional statistics, convex geometry
measure concentration, combinatorics…)

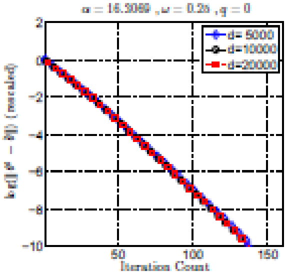
Massive High-dim Data
(images, videos,
texts, audios,
speeches, stocks,
user preferences…)

Cloud Computing
(parallel, distributed,
networked)

New Applications & Services
(data processing,
analysis, compression,
knowledge discovery,
search, recognition…)



Computational Methods
(convex optimization, first-order algorithms
random sampling, approximate solutions…



REFERENCES + ACKNOWLEDGEMENT

Core References:

- *Robust Principal Component Analysis*? Candes, Li, Ma, Wright, Journal of the ACM, 2011.
- *TILT: Transform Invariant Low-rank Textures,* Zhang, Liang, Ganesh, and Ma, IJCV 2012.
- *Compressive Principal Component Pursuit*, Wright, Ganesh, Min, and Ma, ISIT 2012.

More references, codes, and applications on the website:

http://perception.csl.illinois.edu/matrix-rank/home.html

Colleagues:

- Prof. Emmanuel Candes (Stanford)
- Prof. John Wright (Columbia)
- Prof. Zhouchen Lin (Peking University)
- Dr. Yasuyuki Matsushita (MSRA)
- Dr. Arvind Ganesh (IBM Research, India)
- Prof. Shuicheng Yan (NUS, Singapore)
- Prof. Lei Zhang (Hongkong Polytech Univ.)

Students:

- Zhengdong Zhang (MSRA, MIT)
- Xiaodong Li (Stanford)
- Xiao Liang (MSRA, Tsinghua University)
- Xin Zhang (MSRA, Tsinghua University)
- Kerui Min (UIUC), Zhihan Zhou (UIUC)
- Hossein Mobahi (UIUC), Guangcan Liu (UIUC)
- Kui Jia (ADSC, Singapore),
- Tsung-Han Chan (ADSC, Singapore)

THANK YOU!

Questions, please?