秩极小化：理论、算法与应用*

林 宙 辰

北京大学机器感知与智能教育部重点实验室，北京 100871

1. 引言

稀疏表示（Sparse Representation）[Elad2010]和压缩传感（Compressed Sensing）[EK2012]是目前信号处理和机器学习领域的热门课题。所谓稀疏性，指的是有意义的信号在适当选取的一组过完备基（Overcomplete Bases, 或称为字典, Dictionary）下可以只用其中少数几个基来表达。写成数学表达式就是：

如果$W \in \mathbb{R}^{m\times n}$ $(m<n)$是一个合适的字典，y是有意义的信号，x_0是y在W[1]下的最稀疏表示，即：

$$x_0 = \arg\min_x \|x\|_0, \quad s.t. \quad y = Wx,$$

则$\|x_0\|_0 \ll m$，其中$\|x\|_0$表示x中非零元的个数。

上面的描述是基于向量的稀疏性来刻画的。但在实际应用中，我们将面临着各种各样的数据，如图像、视频和基因微阵列（Microarray），它们天然就是矩阵甚至是张量。于是我们就自然面对着一个问题：如何度量矩阵和张量的稀疏性？如果套用向量的稀疏性，把它们强行展开成向量、按照向量来处理，势必破坏数据内在的结构，在很多问题上就会行不通。比如图像或视频压缩，没有一个人会把图像或视频当作向量来压缩，因为这样没有充分利用空间和时间上的相关性。再如 Netflix 挑战[2]（图 1），如果把用户/视频评价矩阵直接按向量处理，必然导致用户未评价的视频都是她/他不喜欢的视频这样不合理的结论。

本文主要讨论矩阵的稀疏性。那么什么才是矩阵的稀疏性度量呢？回想上面的两个例子，大家很容易都能想到要充分利用图像或矩阵的行及列之间的相关性。另外，作为流形学习[LV2007]的基本假定，我们知道真实数据都是存在于高维空间中非常低维的流形上的，而且往往都可以用低维的子空间来近似，比如前几个主分量所张成的线性子空间。行列相关性和低维子空间都共同指向了线性代数的一个基本概念：矩阵的秩。以上的例子都提示我们：**秩是矩阵稀疏性的合理度量**。事实上，秩是非常强的全局约束。一个$m\times n$矩阵如果没有任何约束，它将有mn个自由度；如果它的秩是r，则自由度将下降为$r(m+n-r)$。因此，秩是很好的针对矩阵的正则化子（regularizer）。

* 本文得到国家自然科学基金(61272341)资助。

[1] 对于压缩传感，W前面还会有一个压缩测量矩阵，这里不作赘述。

[2] Netflix 是一家视频租赁公司，拥有很多用户对视频的评价，但这个用户/视频评价矩阵非常稀疏。该公司提供 100 万美元奖金希望能够把预测用户对视频的评价的准确率提高 10%，以便有针对性地推荐，从而提高营收。见 http://www.netflixprize.com/

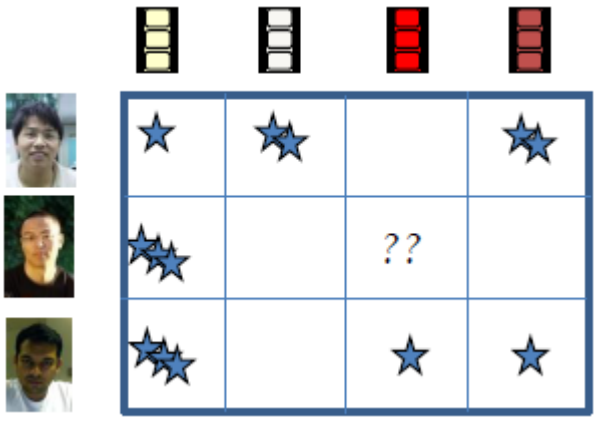


图 1 Netflix 挑战。需要预测用户对其未评价过的视频的喜好程度。

正如其他“新鲜”事物（如稀疏表示）一样，秩其实在统计学中早已被用作矩阵的正则化子，如减秩回归（Reduced-Rank Regression）[Tso1981]；在三维立体视觉里，秩约束更是随处可见[MSKS2004]。但是 E. Candès 等人的工作赋予了秩极小化新的内涵，另外传统的已经获得巨大成功的稀疏表示和压缩传感理论向矩阵推广的需求也是秩极小化焕发新春的内在动力。

本章的主旨是简要介绍秩极小化的理论、算法与应用。第 2 节将介绍秩极小化的几个主要数学模型。第 3 节将介绍一些基本理论结果。第 4 节将介绍求解秩（核范数）极小化的高效算法。第 5 节将介绍秩极小化的一些典型应用。第 6 节将总结本章。

2. 主要数学模型

如前所述，秩很早以前就已被应用于某些领域，如统计学里的减秩回归[Tso1981]和三维立体视觉[MSKS2004]。但是秩极小化重新赢得许多学者的注意却是近几年的事。2008 年，压缩传感发明人之一、斯坦福大学教授 E. Candès 考虑了矩阵填充（Matrix Completion, MC）问题[CR2009]：已知某矩阵 D 在某些位置的值，可否恢复出该矩阵？显然这个问题的答案是不确定的，于是他建议选秩最小的那个解 A：

$$\min_{A} \operatorname{rank}(A), \quad s.t. \quad \pi_{\Omega}(D) = \pi_{\Omega}(A), \tag{1}$$

其中 Ω 是已知值的矩阵元素的位置的集合，π_{Ω} 是保留位置在 Ω 里的矩阵元素的值、其他位置填 0 的投影算子。稍后，E. Candès 又进一步考虑了带噪声的 MC 问题[CP2010]：

$$\min_{A} \operatorname{rank}(A), \quad s.t. \quad \left\|\pi_{\Omega}(D) - \pi_{\Omega}(A)\right\|_F^2 \le \varepsilon. \tag{2}$$

2009 年，Chandrasekaran 等人[CSPW2009]和 Wright 等人[WGRM2009]同时提出了鲁棒主元分析（Robust PCA, RPCA）。他们考虑的是数据中有稀疏大噪声时如何恢复数据的低秩结构：

$$\min_{A,E} \operatorname{rank}(A) + \lambda \left\|E\right\|_0, \quad s.t. \quad D = A + E, \tag{3}$$

其中 $\left\|E\right\|_0$ 表示 E 中非零元的个数。传统的主元分析用 Frobenius 范数来度量噪声，相当于假定噪声是高斯噪声，因此是稠密的（即每个矩阵元素都可能带有噪声）。对于非高斯噪声，如果噪声很强，即使是极少数的噪声，也会使传统的主元分析失败。由于主元分析在应用上的极端重要性，大量学者付出了很多努力在提高主元分析的鲁棒性上，提出了许多号称“鲁棒”的主元分析方法，但是没

有一个方法被理论上严格证明是能够在一定条件下一定能够精确恢复出低秩结构的。

J. Wright 的工作后来得到 E. Candès 的加入，获得了更强的结果，即观测矩阵 D 可以只在部分位置知道值。推广后的模型为[CLMW2011]：

$$\min_{A,E} \operatorname{rank}(A) + \lambda \|E\|_0, \quad s.t. \quad \pi_\Omega(D) = \pi_\Omega(A+E). \tag{4}$$

在他们的论文里，也讨论了带稠密高斯噪声的广义 RPCA 模型[CLMW2011]：

$$\min_{A,E} \operatorname{rank}(A) + \lambda \|E\|_0, \quad s.t. \quad \|\pi_\Omega(D) - \pi_\Omega(A+E)\|_F^2 \le \varepsilon. \tag{5}$$

RPCA 模型只能鲁棒地提取一个主子空间，即所有干净数据所张成的线性子空间，它不能对数据进行聚类。于是 2010 年 Liu 等人提出了低秩表示模型（Low-Rank Representation，LRR）[LLY2010][LLYSYM2013]。它受稀疏子空间聚类模型（Sparse Subspace Clustering，SSC）[EV2009]的启发，通过数据的自我表达，要求表达系数矩阵尽可能低秩。在带稀疏噪声的情形，LRR 的数学模型为：

$$\min_{Z,E} \operatorname{rank}(Z) + \lambda \|E\|_0, \quad s.t. \quad D = DZ + E. \tag{6}$$

LRR 的最优表达系数矩阵 Z^* 可以作为样本间的相似性度量，用 $(|Z^*| + |Z^{*T}|)/2$ [3]构建样本间的邻接矩阵（$|Z^*|$ 表示把 Z^* 的元素都取绝对值得到的矩阵），再通过谱聚类就可以把数据聚类成若干线性子空间。

3. 理论分析

上面的低秩模型（1）—（6）都是关于离散函数（秩、非零元个数）的优化问题，一般都是 NP 难的组合优化问题。为了克服计算上的困难，一个通常的做法是把目标函数换成某些凸函数，得到相应的凸规划问题。凸规划问题一般都有多项式时间的解，因此可以认为是初步克服了秩极小化问题所面临的计算上的困难。为了尽可能和原问题接近，该凸函数应和原函数尽可能接近。和某非凸函数“最近”的凸函数应当是该函数的凸包络（Convex Envelope），即不超过该函数的最大凸函数[Rock1970]。Fazel 等人[Fazel2002]证明了秩函数在矩阵谱范数单位球 $B_2 = \{X \mid \|X\|_2 \le 1\}$ 上的凸包络是矩阵的核范数（即矩阵所有奇异值的和）。而向量 0 范数（其实是伪范数，Pseudo Norm）在无穷范数单位球 $B_\infty = \{x \mid \|x\|_\infty \le 1\}$ 上的凸包络是其 1 范数（即所有元素绝对值的和）则是熟知的结论[Dono1995]。做凸包络替换后，就得到如下的凸规划问题：

$$\min_A \|A\|_*, \quad s.t. \quad \pi_\Omega(D) = \pi_\Omega(A), \tag{1'}$$

$$\min_A \|A\|_*, \quad s.t. \quad \|\pi_\Omega(D) - \pi_\Omega(A)\|_F^2 \le \varepsilon, \tag{2'}$$

$$\min_{A,E} \|A\|_* + \lambda \|E\|_1, \quad s.t. \quad D = A + E, \tag{3'}$$

[3] 在后来的 TPAMI2013 论文[LLYSYM2013]中，Liu 等人改用 $|U_{Z^*} U_{Z^*}^T|$ 构建邻接矩阵，其中 U_{Z^*} 是 Z^* 的瘦型奇异值分解的做奇异向量集合。原因在第 3 节解释。

$$\min_{A,E} \|A\|_* + \lambda \|E\|_1, \quad s.t. \quad \pi_\Omega(D) = \pi_\Omega(A+E), \tag{4'}$$

$$\min_{A,E} \|A\|_* + \lambda \|E\|_1, \quad s.t. \quad \|\pi_\Omega(D) - \pi_\Omega(A+E)\|_F^2 \le \varepsilon, \tag{5'}$$

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_1, \quad s.t. \quad D = DZ + E, \tag{6'}$$

其中$\|E\|_1$表示E的所有元素的绝对值的和。于是，自然就面临着一个问题：凸规划（1'）—（6'）是否和原问题（1）—（6）等价？在这个问题上，不同的模型所获得的答案程度不一。

对于模型（4）和（4'）的关系，E. Candès 等人[CLMW2011]证明了当真解(A_0, E_0)符合适当的概率分布且（假定$m \ge n$）

$$\lambda = \frac{1}{\sqrt{m}}, \quad \text{rank}(A_0) = O(n / \ln^2 m), \quad \|E_0\|_0 = O(mn), \quad |\Omega| = O(mn), \tag{7}$$

时，求解（4'）会以压倒性的概率（Overwhelming Probability）恢复出真解(A_0, E_0)，即失败的概率会随着矩阵维度的增长指数式衰减。注意，（7）是个相对来说比较宽泛的条件，它只要求A_0的秩不太低、E_0不需要太稀疏、已知值的个数$|\Omega|$不太少。因此，一般用户可以放心使用（4'），而且不需要限定噪声E_0的大小（Magnitude）。这些特性和传统的 PCA 的差别是非常显著的。Wright 等人[WGMM2012]进一步证明了更一般的多分量情形，即测量数据是多个分量的混合，不管这些分量所用的范数是什么，只要真解在无亚采样时是可精确恢复的，则在压缩测量（Compressive Measurements）下还可以以压倒性概率精确恢复。

对于（5）和（5'）的关系，Zhou 等人[ZLWCM2010]证明了二者的解的差异是随着噪声水平ε连续变化的。

对于（6）和（6'）的关系，Liu 等人[LXY2012]仅证明了当例外（Outlier）样本比例不超过某阈值时，Z_0的行空间和哪些样本是例外值是可以精确恢复的，Z_0的行空间可以从$U_{Z^*}U_{Z^*}^T$得到，其中$U_{Z^*}\Sigma_{Z^*}V_{Z^*}^T$为最优解$Z^*$的瘦型奇异值分解。该分析没有回答$Z_0$和$E_0$本身是否可以精确恢复，但是幸运的是将 LRR 应用于子空间聚类时，我们只需要Z_0的行空间就够了。

Wei 和 Lin[WL2010]进一步分析了 LRR 的数学性质，发现无噪声的 LRR 模型：

$$\min_{Z} \|Z\|_*, \quad s.t. \quad D = DZ, \tag{8}$$

有唯一解，而且该解可以很方便地表达出来：设D的瘦型奇异值分解为$U_D\Sigma_D V_D^T$，则该解为$V_D V_D^T$。矩阵$V_D V_D^T$在立体视觉里被称为形状交互矩阵（Shape Interaction Matrix）[CK1998]。Liu 等人[LLYSYM2013]进一步发现使用一般字典的 LRR 模型：

$$\min_{Z} \|Z\|_*, \quad s.t. \quad D = BZ, \tag{9}$$

也有唯一解，而且该解也可以很方便地表达出来：$Z^* = B^+ D$，其中 B^+ 为 B 的 Moore-Penrose 伪逆。LRR 有闭解的事实说明 LRR 的数学性质要比 SSC 丰富，如何利用其闭解进一步分析 LRR 的性质或设计快速算法是值得深入探讨的课题。

4. 算法

把 NP 难问题（1）—（6）转化为凸规划问题（1’）—（6’）已经实现了降低复杂度的飞跃，因为凸规划问题一般都有多项式时间复杂度的求解算法（比如内点法[BV2004]）。但是对于大规模的数据，一般需要 $O(n\text{poly}\log(n))$ 的复杂度，连平方复杂度都是不可容忍的。为了帮助理解秩极小化模型所面临的计算复杂度的障碍，我们举以下例子说明。对于 RPCA 问题，如果矩阵大小是 $n \times n$，则该问题有 $2n^2$ 个未知数，即使 $n = 1000$，它对应于一个不大的矩阵，未知数个数也达到了一百万。如果使用内点法来求解，比如使用斯坦福大学的 CVX 软件包[GB2009]，则每迭代一次的时间复杂度是 $O(n^6)$，而空间复杂度则是 $O(n^4)$，导致一般内存为 4GB 的 PC 机只能处理 80×80 的矩阵。因此要使得秩极小化模型实用，必须要设计高效的优化算法。

目前面向大规模计算的优化算法都是一阶算法。代表性的算法包括加速近邻梯度法（Accelerated Proximal Gradient，APG）和交错方向法（Alternating Direction Method，ADM）。在一些有特殊结构的问题上，APG 可以被推广（Generalized APG，GAPG）。为了方便计算，ADM 还有一个线性化的版本（Linearized Alternating Direction Method, LADM）。

4.1 加速近邻梯度法（APG）及其推广（GAPG）

APG 主要面向无约束凸优化。考虑如下无约束凸优化问题

$$\min_{x \in \mathbb{R}^m} f(x), \tag{10}$$

其中 $f(x)$ 是 $C^{1,1}$ 的凸函数，即 $f(x)$ 的导数是 Lipschitz 连续的：

$$\|\nabla f(x) - \nabla f(y)\| \le L(f)\|x - y\|, \quad \forall x, y \in \mathbb{R}^m. \tag{11}$$

称 $L(f)$ 为 f 的一阶 Lipschitz 系数。假设（10）的解集为 Q，$R(f)$ 为原点到 Q 的距离。不失一般性，我们还假定迭代从原点开始。定义如下函数类：

$$S_m(L, R) = \left\{ f \mid f : \mathbb{R}^m \to \mathbb{R}, f \in C^{1,1}, L(f) \le L, R(f) \le R \right\}. \tag{12}$$

则有如下经典结果[Nem1994]：

定理 1：如果使用一阶算法求解问题（10），其中 $f(x)$ 属于函数类 $S_m(L,R)$，则算法复杂度是：

$$O(1)\min\left\{m,\sqrt{\frac{LR^2}{\varepsilon}}\right\}\le complexity(\varepsilon)\le\sqrt{\frac{4LR^2}{\varepsilon}},$$

其中 $complexity(\varepsilon)$ 表示要得到精度为 ε 的解所需的计算量。

定理 1 意味着一阶算法在 $S_m(L,R)$ 上的收敛速度是 $O(k^{-2})$，其中 k 是迭代次数。但是通常的梯度下降法只能达到 $O(k^{-1})$ 的收敛速度[Nem1994]：

定理 2：假定使用固定步长的梯度下降法求解（10）：

$$x_0=0, x_{k+1}=x_k-\gamma\nabla f(x_k),$$

其中 $\gamma\in(0,2/L(f))$，则收敛速度是 $O(k^{-1})$：

$$f(x_k)-f^*\le\frac{R^2(f)}{\gamma(2-\gamma L(f))k},$$

其中 f^* 是目标函数最优值。

定理 1 和 2 所揭示的收敛速度在阶上的差异一直激励着学者们寻找收敛速度为 $O(k^{-2})$ 的一阶算法。1983 年，Y. Nesterov 终于构造出了一个这样的算法[Nest1983]。算法描述如下：

Input: $L(f)$。

Process:

Initialize：$x_0 = y_1 = 0$，$t_1 = 1$，$k = 1$。

Loop：

$$x_k = y_k - \frac{1}{L(f)}\nabla f(y_k)，\quad t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}，\quad y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})；$$

$k \leftarrow k+1$。

Output:

y_k

图 1　Nesterov 方法[Nest1983]

可以证明上面的算法的收敛速度是$O(k^{-2})$[Nem1994]：

定理 3：Nesterov 方法在$S_m(L,R)$上的收敛速度是$O(k^{-2})$：

$$f(y_k) - f^* \le \frac{2L(f)R^2(f)}{k^2}.$$

问题（10）要求目标函数为$C^{1,1}$，这在应用中很受限制，因为涉及稀疏性和低秩性的目标函数往往都是不可导的。于是 Beck 和 Teboulle[BT2009]把 Nesterov 方法做了推广，可以处理如下形式的无约束凸优化问题：

$$\min_{x\in\mathbb{R}^m} f(x) + g(x)，\tag{13}$$

其中$f(x)$和$g(x)$都是凸函数，但是只要求$f(x)$为$C^{1,1}$，$g(x)$可以不可导。其算法称为 APG，描述如下：

Input: $L(f)$。

Process:

Initialize：$x_0 = y_1 = 0$，$t_1 = 1$，$k = 1$。

Loop：

$$x_k = \arg\min_x \left\{ g(x) + \frac{L(f)}{2} \left\| x - \left(y_k - \frac{1}{L(f)} \nabla f(y_k) \right) \right\|^2 \right\},$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1});$$

$k \leftarrow k+1$。

Output:

y_k

图 2　APG 算法[BT2009]

APG 的收敛速度也可以被证明是$O(k^{-2})$的[BT2009]。值得注意的是，APG 假定子问题

$$\min_x \left\{ g(x) + \frac{\alpha}{2} \| x - z_k \|^2 \right\} \tag{14}$$

是容易求解的，这在应用中经常可以满足，因为$g(x)$一般会取作矩阵或向量的范数，此时该子问题甚至有闭解。比如，当$g(x) = \|x\|_1$时，最优解为$\Theta_{\alpha^{-1}}(z_k)$ [Dono1995]，其中

$$\Theta_\varepsilon(c) = \mathrm{sgn}(c)\max(|c| - \varepsilon, 0) \tag{15}$$

为收缩算子（Shrinkage Operator）[Dono1995]；当x和z_k为矩阵X和Z_k、$g(X) = \|X\|_*$时，最优解为

$$U_k \Theta_{\alpha^{-1}}(\Sigma_k) V_k^T, \tag{16}$$

其中$U_k \Sigma_k V_k^T$为Z_k的奇异值分解[CCS2010]。

对于常见的带约束凸优化问题：

$$\min_x f(x), \quad s.t. \quad Ax = b, \tag{17}$$

可以通过罚函数的方式转化为无约束问题：

$$\min_x \frac{\alpha}{2}\|Ax-b\|^2 + f(x), \tag{18}$$

来求近似解。但是要注意的是，如果一开始就让惩罚系数α很大，则收敛会极其地慢。所以一般使用 Continuation 技巧[GLWWCM2009]，即一开始α取较小的值，随着迭代进行再逐渐增大到一个较大的值，这样简单处理后就会显著地加快收敛。另外，APG 需要估计 Lipschitz 系数$L(f)$，如果$L(f)$估计得过于保守（太大），则会影响收敛速度，所以 Beck 和 Teboulle[BT2009]还进一步提出了使用动态估计的$L(f)$，以加速收敛。

Zuo 和 Lin[ZL2011]对条件（11）做了放松以加速收敛。他们指出（11）实际上只是为了保证：

$$f(x) \le f(y) + \langle x-y, \nabla f(y)\rangle + \frac{1}{2}L(f)\|x-y\|^2, \quad \forall x, y \in \mathbb{R}^m, \tag{19}$$

对算法收敛性起关键作用的只是（19）式。因此，Zuo 和 Lin[ZL2011]指出（19）可以放宽为

$$f(x) \le f(y) + \langle x-y, \nabla f(y)\rangle + \frac{1}{2}\|x-y\|_{L_f}^2, \quad \forall x, y \in \mathbb{R}^m, \tag{20}$$

其中L_f为正定矩阵，$\|x\|_{L_f} = \sqrt{x^T L_f x}$。相应地，图 2 中$x_k$的更新算法为：

$$x_k = \arg\min_x \left\{ g(x) + \langle \nabla f(y_k), x-y_k\rangle + \frac{1}{2}\|x-y_k\|_{L_f}^2 \right\}, \tag{21}$$

其余不变，$O(k^{-2})$的收敛速度也得到保持。当$g(x)$可分解时（参见（24）），L_f常常可以选成对角阵，为不同的变量提供不同的一阶 Lipschitz 系数，因此可以加速收敛。Zuo 和 Lin[ZL2011]把 GAPG 应用到图像恢复问题上得到了比 APG 快 2 倍以上的加速比。

4.2 交错方向法（ADM）及其线性化（LADM）

在有约束情形，APG 只能求得近似解，如果需要较高的数值精度，可以使用交错方向法（ADM）。ADM 来源于增广 Lagrange 乘子法（Augmented Lagrange Multiplier，ALM），所以 ADM 又称为非精

确增广 Lagrange 乘子法[LCM2009]。为简单起见，只考虑上面带等式约束的凸优化问题（17）。首先引入问题（17）的增广 Lagrange 函数：

$$L(x,\lambda,\beta)=f(x)+\langle\lambda,Ax-b\rangle+\frac{\beta}{2}\|Ax-b\|^2, \tag{22}$$

其中 λ 为 Lagrange 乘子，β 为惩罚系数。则有如下迭代：

Process:

Initialize：x_0，λ_0，$k=0$。

Loop：

$x_{k+1}=\arg\min_x L(x,\lambda_k,\beta)$；

$\lambda_{k+1}=\lambda_k+\beta(Ax_{k+1}-b)$；

$k\leftarrow k+1$。

Output:

x_k

图 3　ALM 算法[LCM2009]

ALM 算法假定子问题

$$x_{k+1}=\arg\min_x L(x,\lambda_k,\beta) \tag{23}$$

容易求解，这个假定并不一定成立。但是在应用中 $f(x)$ 经常有结构，比较常见的结构是可分解性，即：

$$f(x)=f_1(y)+f_2(z),\quad x=(y^T,z^T)^T. \tag{24}$$

此时约束 $Ax=b$ 可分解成

$$A_1y+A_2z=b. \tag{25}$$

交错方向法就是利用了目标函数的可分解性。其迭代过程如下：

Process:

Initialize：y_0， z_0， λ_0， $k=0$。

Loop：

$$y_{k+1}=\arg\min_y L(y,z_k,\lambda_k,\beta)=\arg\min_y f_1(y)+\frac{\beta}{2}\left\|A_1y+A_2z_k-b+\lambda_k/\beta\right\|^2;$$

$$z_{k+1}=\arg\min_z L(y_{k+1},z,\lambda_k,\beta)=\arg\min_z f_2(z)+\frac{\beta}{2}\left\|A_2z+A_1y_{k+1}-b+\lambda_k/\beta\right\|^2;$$

$$\lambda_{k+1}=\lambda_k+\beta(A_1y_{k+1}+A_2z_{k+1}-b);$$

$k\leftarrow k+1$。

Output:

(y_k,z_k)

图 4　ADM 算法[LCM2009]

ADM 假定了如下形式的子问题：

$$y_{k+1}=\arg\min_y f_1(y)+\frac{\beta}{2}\left\|A_1y-p_k\right\|^2, \tag{26}$$

$$z_{k+1}=\arg\min_z f_2(z)+\frac{\beta}{2}\left\|A_2z-q_k\right\|^2, \tag{27}$$

是容易求解的，这在 $A_1=I$， $A_2=I$ 时是经常满足的，比如 RPCA 问题[LCM2009]（参见 APG 的介绍）。但是对于更一般的问题，A_1,A_2 可能不是单位阵，导致上面的子问题不易求解。于是 Lin 等人[LLS2011]提出了线性化的 ADM（Linearized ADM, LADM），把（26）的二次项通过做一阶 Taylor 展开并加上一个近邻项（proximal term）来近似，把（26）改成：

$$\begin{aligned}y_{k+1}&=\arg\min_y f_1(y)+\beta\left\langle A_1^T(A_1y-p_k),y-y_k\right\rangle+\frac{\beta\eta_1}{2}\left\|y-y_k\right\|^2\\&=\arg\min_y f_1(y)+\frac{\beta\eta_1}{2}\left\|y-y_k+A_1^T(A_1y-p_k)/\eta_1\right\|^2\end{aligned}. \tag{28}$$

这样就可以很方便地得到 y_{k+1} 了。对（27）也进行类似处理，可得：

$$\begin{aligned} z_{k+1} &= \arg\min_z f_2(z) + \beta\left\langle A_2^T(A_2 z - q_k), z - z_k\right\rangle + \frac{\beta\eta_2}{2}\|z - z_k\|^2 \\ &= \arg\min_z f_2(z) + \frac{\beta\eta_2}{2}\left\|z - z_k + A_2^T(A_2 z - q_k)/\eta_2\right\|^2 \end{aligned}. \quad (29)$$

这样 z_{k+1} 也可以很方便地得到了。如果采用动态调整的惩罚系数 β：

$$\beta_{k+1} = \begin{cases} \rho\beta_k, & \text{if } \beta_k \max\left\{\|y_{k+1} - y_k\|, \|z_{k+1} - z_k\|\right\}/\|b\| \le \varepsilon_2, \\ \beta_k, & \text{otherwise,} \end{cases} \quad (30)$$

则有可能加速收敛，其中 $\rho \ge 1$，$0 < \varepsilon_2 \ll 1$ 为一阈值。β 的初值的选择不要太大，应使得 β_k 在头几次迭代就增长。ρ 的选择要适中，最好能使 β_k 随着迭代稳步增长。迭代在下列条件满足时停止：

$$\begin{cases} \|A_1 y_{k+1} + A_2 z_{k+1} - b\|/\|b\| \le \varepsilon_1, \\ \beta_k \max\left\{\|y_{k+1} - y_k\|, \|z_{k+1} - z_k\|\right\}/\|b\| \le \varepsilon_2. \end{cases} \quad (31)$$

Lin 等人[LLS2011]证明了当序列 $\{\beta_k\}$ 不减且有上界时，$\{(y_k, z_k, \lambda_k)\}$ 收敛于问题

$$\min_{y,z} f_1(y) + f_2(z), \quad s.t. \quad A_1 y + A_2 z = b, \quad (32)$$

的 KKT 点，即 $\{(y_k, z_k)\}$ 收敛到问题（32）的最优解。Lin 等人[LLS2011]把 LADM 应用到 LRR 的求解上，结合奇异值分解的求解技巧，得到了第一个复杂度为 $O(rmn)$ 的算法，其中 $m \times n$ 为 D 的维度，r 为最优 Z 的秩。

4.3 奇异值分解的计算

在秩极小化问题里，优化迭代时子问题之一就是如下形式的问题：

$$A_{k+1} = \arg\min_A \varepsilon_k \|A\|_* + \frac{1}{2}\|A - W_k\|_F^2. \quad (33)$$

正如前面（见（16））所指出的，它的解为：

$$A_{k+1} = \mathrm{T}_{\varepsilon_k}\left(W_k\right) = U_k \Theta_{\varepsilon_k}\left(\Sigma_k\right) V_k^T, \tag{34}$$

其中$U_k \Sigma_k V_k^T$为W_k的奇异值分解，Θ_ε为软阈值算子，其图像如下所示：

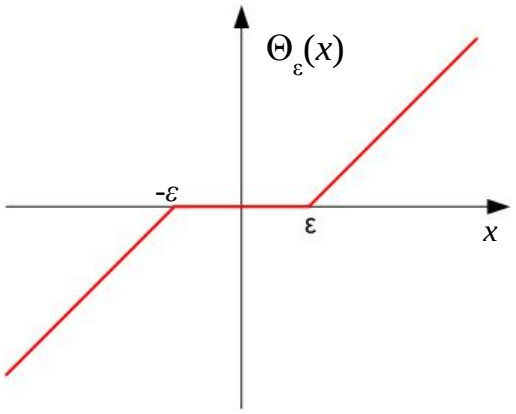


图 5 软阈值算子的图像

因此求解秩极小化问题往往离不开奇异值分解，而维度为$m \times n$的矩阵的奇异值分解的复杂度为$O(mn\min(m,n))$，所以一般来说求解秩极小化问题计算量都很大，在面临大矩阵时问题尤为突出。所幸从（34）可以看出小于ε_k的奇异值及其奇异向量无需求出，因为这些奇异值将被收缩为 0，从而对A_{k+1}没有贡献。于是就可以只计算大于ε_k的奇异值及其奇异向量，这可以使用 PROPACK[Lar1998]来实现，计算量相应地下降到$O(rmn)$，其中r为最优Z的秩。值得一提的是，PROPACK 只能提供指定个数的最大奇异值及其奇异向量，所以调用 PROPACK 时要动态地预测r的值[LCM2009]。当解不是低秩的时，比如在图像处理和计算机视觉里已经有广泛应用的 Transform Invariant Low-Rank Textures (TILT)[ZGLM2012]（见（35）及 5.3 节），可以使用增量奇异值分解[RL2013]进行加速。

5. 应用

秩极小化在信号处理和机器学习等领域里已经获得了广泛的应用，NIPS2011 上曾出现了大量的讨论低秩模型的论文。由于专业所限，本节只简要介绍笔者及其合作者在图像处理和计算机视觉领域里找到的典型应用。

5.1 背景建模[CLMW2011]

背景建模的最简单情形是从固定摄相机拍摄的视频中分离背景和前景。此时很容易想到背景是基本不变的，所以如果把背景的每一帧作为矩阵的一列，则该矩阵低秩。同时由于前景是移动的物体，占据像素比例较低，所以前景对应于视频中的稀疏“噪声”部分。由此得到做背景建模的 RPCA

模型（3）或（3’），其中 D 的每一列是视频的每一帧拉直后得到的向量， A 的每一列对应于背景的每一帧拉直后得到的向量， E 的每一列对应于前景的每一帧拉直后得到的向量。部分结果如下：

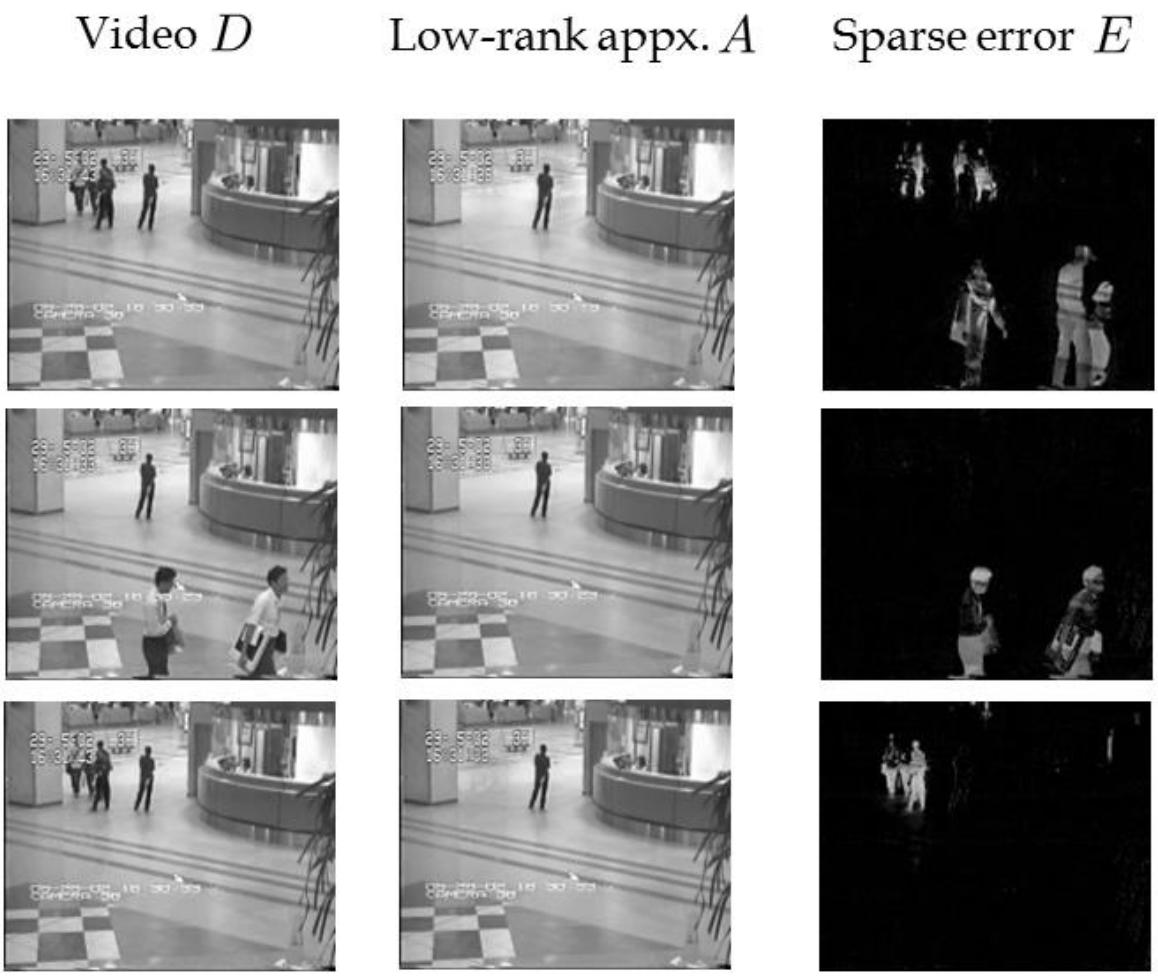


图 6 背景建摸。第一列是监控视频，第二列是背景视频，第三列是前景视频（ E 值取绝对值）。

5.2 图像批量对齐（RASL）[PGWXM2012]

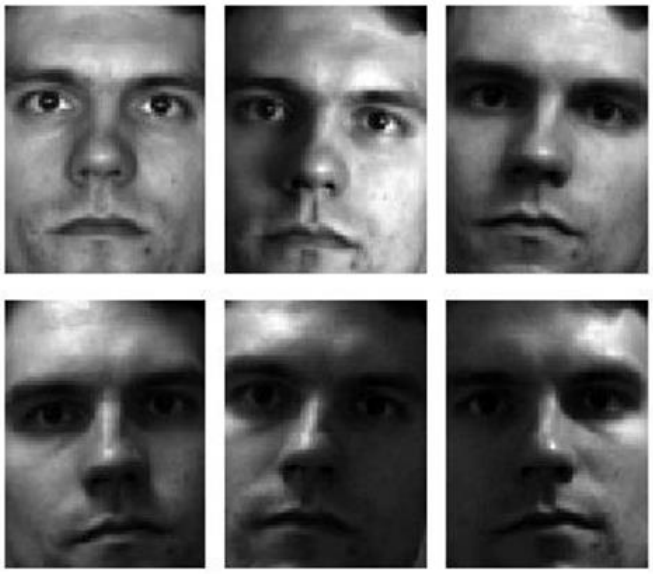
背景建模需要假定背景已经对齐，这样才能得到低秩的背景视频。在没有对齐的情况下，可以考虑把每一帧/每一幅图像作适当的几何变形使它们对齐。此时数学模型为：

$$\min_{\tau,A,E} \|A\|_* + \lambda \|E\|_1, \quad s.t. \quad D \circ \tau = A + E, \tag{35}$$

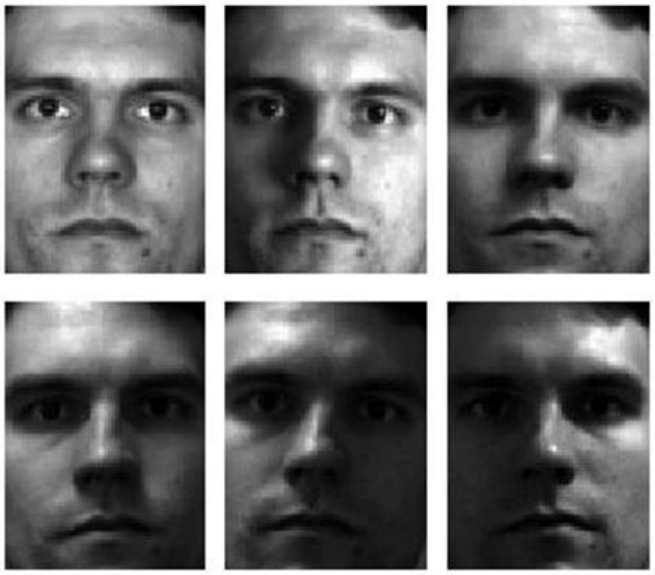
其中 $D \circ \tau$ 是把每一帧/每一幅图像作适当的几何变形 τ 的形式写法（每一帧/每一幅图像的几何变形不同）。此时，（35）是一个非凸优化问题。为了有效求解，Peng 等人[PGWXM2012]提出了把 τ 局部线性化的迭代算法，即做如下循环直至 $\Delta\tau_k$ 足够小：

$$\begin{cases} \min\limits_{\Delta\tau_k, A, E} \|A\|_* + \lambda \|E\|_1, \quad s.t. \quad D \circ \tau_k + J\Delta\tau_k = A + E, \\ \tau_{k+1} \leftarrow \tau_k + \Delta\tau_k, \\ k \leftarrow k+1, \end{cases} \tag{36}$$

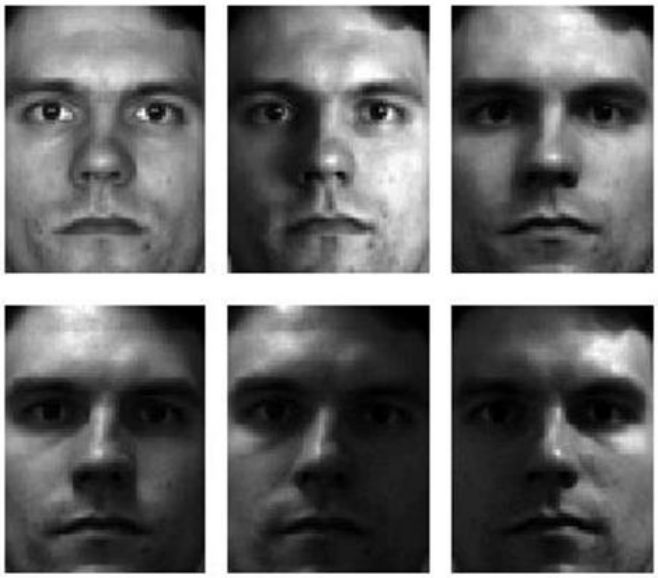
其中 J 是 $D \circ \tau$ 对 τ 的参数的 Jacobi 矩阵。在反射变换模型下，人脸图像的部分对齐结果如下：



人脸的初始姿态



人脸对齐的中间结果



人脸对齐的最终结果

图 7 人脸对齐的迭代过程

5.3 变换不变低秩纹理（TILT）[ZGLM2012]

变换不变低秩纹理（Transform Invariant Low-rank Textures, TILT）的数学模型形式上和 RASL 的（35）完全一样，其求解过程也完全一样，但是这里 D 是图像的一个长方形图像块。TILT 的思想是通过几何变换 τ 把 D 所代表的图像区域校正成正则的区域，如具有横平竖直、对称等特性，这些特性可以通过低秩性来进行刻画。以下是射影变换下图像校正的例子：

图 8 利用 TILT 模型进行图像校正的例子。第一行表示原始的图像块（长方形框）及相应校正变换（四边形框，校正变换就是四边形框相对于长方形框的几何变换），第二行是校正后的图像块。

TILT 原则上对任何参数变换都适用。Zhang 等人[ZLM2011]还考虑了基于广义柱体变换的 TILT，可用于人造建筑物表面纹理的提取。一些例子如下：

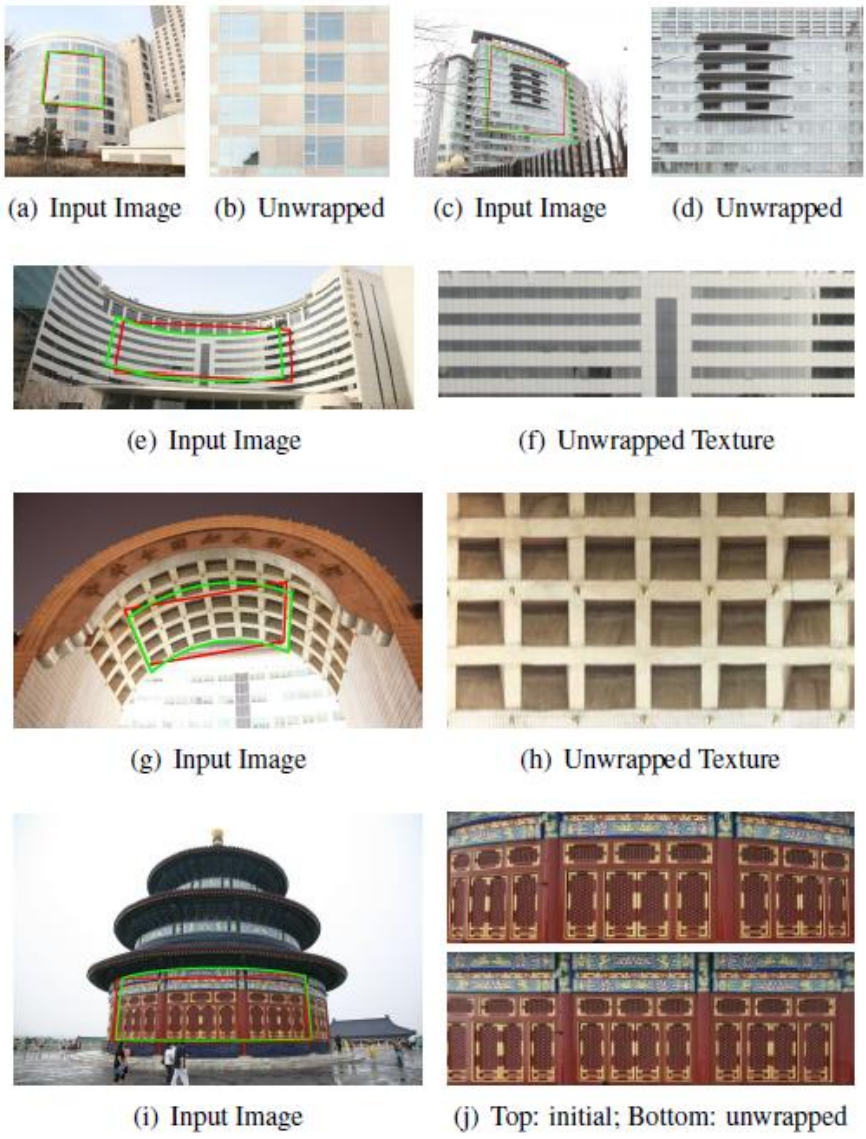


图 9 基于广义柱体变换的 TILT 的建筑物纹理展开

TILT 还被广泛应用于建筑物几何建摸[ZGLM2012]、相机自动标定和镜头畸变自动校正[ZMM2011]。由于其应用上的重要性，Ren 和 Lin[RL2013]特别研究了它的快速算法，把求解速度提高了 5 倍以上。

5.4 运动分割[LLY2010, LLYSYM2013]

LRR 被认为是做刚体运动分割最好的算法之一[AEH2013]。所谓刚体运动分割，就是把视频里做刚体运动的物体上的特征点进行聚类，使得每一类对应于一个独立运动的物体，这样就可以得到物体运动的轨迹。部分例子如下：

(a) Checkerboard (b) Traffic (c) Articulated

图 10 运动分割的例子

5.5 图像分割[CLWHY2011]

图像分割是特殊的聚类问题。首先把图像过分割（Over Segment）成超像素（Super-pixel），然后在超像素上提取适当的特征，通过改进的 LRR 模型综合多种特征（基本上每一种特征对应于一个 LRR 模型），求出整体表示矩阵 Z^*，然后对用 $(|Z^*|+|(Z^*)^T|)/2$ 表出的相似性矩阵进行正则化割（Normalized Cut），得出超像素的聚类关系，每一类就对应于一个图像区域。部分例子如下：

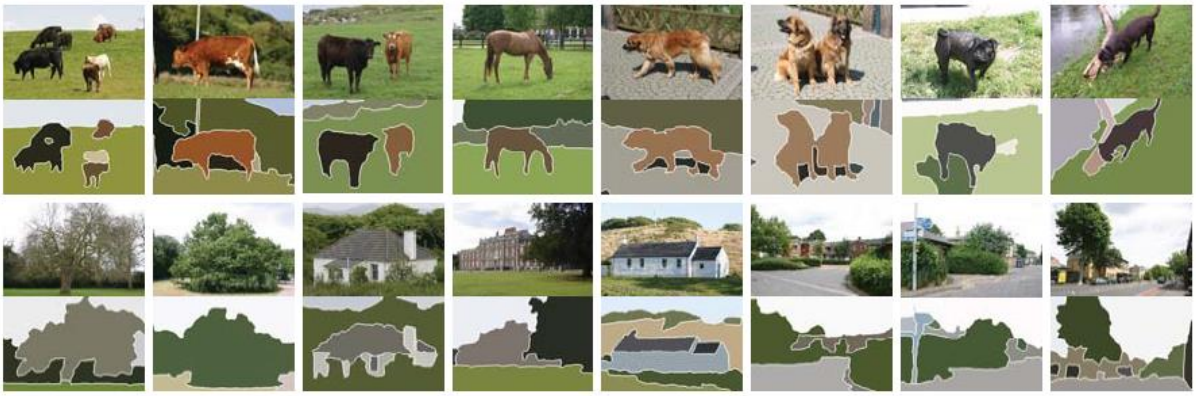


图 11 图像分割的例子

5.6 图像显著区域检测[LLYY2012]

运动分割和图像分割都是利用了 LRR 的表示矩阵 Z，而图像显著区域检测则是利用 LRR 里的稀疏“噪声” E。图像显著区域一般就是图像中“与众不同”的区域，因此如果用其他区域进行“预测”则会产生较大的误差。因此，如果把图像分解成小块，在其上提取适当特征，则图像显著区域对应于 LRR 里的稀疏“噪声” E 较大的部分。部分例子如下：

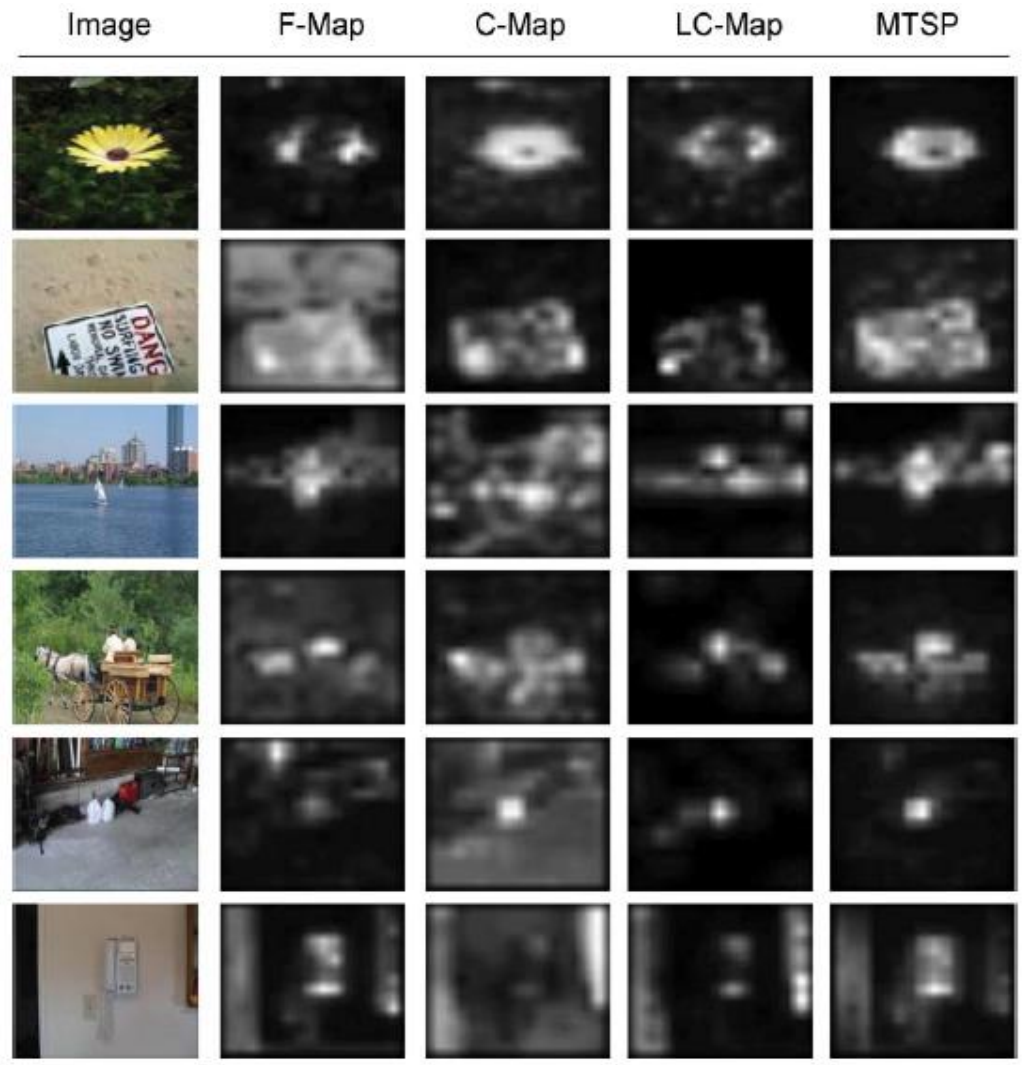


图 12 图像显著区域检测的例子。第一列是输入图像，第二至五列是不同方法的检测结果，其中最后一列是基于 LRR 的检测方法。

秩极小化的其他应用，如基于光度学的立体视觉[WGSMWM2010]、图像标签的改进[ZYM2010]、视觉域适应（Visual Domain Adaption）[JLLC2012]、鲁棒视觉跟踪[ZGLA2012]、三维人脸特征提取[MR2012]、CT 重建[GCSZ2011]、图像半监督分类[ZGLMZY2012]、图像集的协同分割（co-segmentation）[MSXC2012]、甚至音频分析[PK2012]等，限于篇幅就不再一一介绍了。

6. 结束语

秩极小化是近年来信号处理、机器学习等领域的研究热点之一，短短几年从理论、算法到应用各方面都得到了快速的发展，本文只是做了非常粗浅的介绍。许多具体问题，如果结合问题的特性适当地引入低秩性约束，很多情况下都能得到更好的结果。有些问题的数据本身可能没有低秩性，此时可以引入适当变换增强其低秩性（如 RASL/TILT 对 RPCA 的改进）。有些学者没有检查数据是否具有低秩性或对数据进行适当的预处理就声称使用低秩约束效果不好，这是不太严谨的。除了寻找更多的应用外，还需要研究求解秩极小化问题的近线性复杂度算法（即 $O(n\text{polylog}(n))$，其中

$n \times n$为矩阵的大小）以支持大规模数据处理，以及矩阵往张量的推广。

参 考 文 献

[AEH2013] Adler A, Elad M, Hel-Or Y. Probabilistic subspace clustering via sparse representations. IEEE Signal Processing Letters, 2013, 20(1):63 – 66.

[BT2009] Beck, A, Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2009, 2(1):183-202.

[BV2004] Boyd S, Vandenberghe L. Convex Optimization, Cambridge University Press, 2004.

[CCS2010] Cai J-F, Candès E J, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 2010, 20(4): 1956-1982.

[CK1998] Costeira J, Kanade T. A multibody factorization method for independently moving objects. International Journal of Computer Vision, 1998, 29(3):159-179.

[CLMW2011] Candès E J, Li X, Ma Y, Wright J: Robust principal component analysis? Journal of the ACM, 2011, 58(1): 1-37.

[CLWHY2011] Cheng B, Liu G, Wang J, Huang Z, Yan S. Multi-task low-rank affinity pursuit for image segmentation. In: Proceedings of International Conference on Computer Vision (ICCV), 2011, 2439-2446.

[CP2010] Candès E J, Plan Y. Matrix completion with noise. Proceedings of the IEEE, 2010, 98(6): 925-936.

[CR2009] Candès E J, Recht B. Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 2009, 9(6): 717-772.

[CSPW2009] Chandrasekaran V, Sanghavi S, Parrilo P, Willsky A. Sparse and low-rank matrix decompositions. In: Allerton'09 Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing, 2009, 962-967.

[Dono1995] Donoho, D. De-noising by soft-thresholding. IEEE Transaction on Information Theory, 1995, 41(3), 613–627.

[EK2012] Eldar Y C, Kutyniok G (editors). Compressed Sensing: Theory and Applications, Cambridge University Press, 2012.

[Elad2010] Elad M. Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, 2010.

[EV2009] Elhamifar E, Vidal R. Sparse subspace clustering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, vol. 2, 2790-2797.

[Fazel2002] Fazel M. Matrix Rank Minimization with Applications, PhD thesis, 2002.

[GB2009] Grant M, Boyd S. CVX: Matlab software for disciplined convex programming (web page and software). http://stanford.edu/~boyd/cvx, June 2009.

[GCSZ2011] Gao H, Cai J-F, Shen Z, Zhao H. Robust principal component analysis-based four-dimensional computed tomography. Physics in Medicine and Biology, 2011, 56:3181–3198.

[GLWWCM2009] Ganesh A, Lin Z, Wright J, Wu L, Chen M, Ma Y. Fast algorithms for recovering a corrupted low-rank matrix. In: Proceedings of 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2009, 213-216.

[JLLC2012] Jhuo I-H, Liu D, Lee D T, Chang S-F. Robust visual domain adaptation with low-rank reconstruction. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, 2168-2175.

[Lar1998] Larsen R M. Lanczos bidiagonalization with partial reorthogonalization. Department of Computer Science, Aarhus University, Technical Report, DAIMI PB-357, 1998. Code available at http://soi.stanford.edu/~rmunk/PROPACK/

[LCM2009] Lin Z, Chen M, Ma Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrix. University of Illinois at Urbana Champaign, Technical Report UILU-ENG-09-2215, October 2009 (arXiv: 1009.5055).

[LLS2011] Lin Z, Liu R, Su Z. Linearized alternating direction method with adaptive penalty for low-rank representation. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), 2011, 612-620.

[LLY2010] Liu G, Lin Z, Yu Y. Robust subspace segmentation by low-rank representation. In: Proceedings of International Conference on Machine Learning (ICML), 2010, 663-670

[LLYSYM2013] Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y. Robust recovery of subspace structures by low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184.

[LLYY2012] Lang C, Liu G, Yu J, Yan S. Saliency detection by multitask sparsity pursuit. IEEE Transactions on Image Processing, 2012, 21(3): 1327-1338.

[LV2007] Lee J A, Verleysen M. Nonlinear Dimensionality Reduction, Springer, 2007.

[LXY2012] Liu G, Xu H, Yan S. Exact subspace segmentation and outlier detection by low-rank representation. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTAT), 2012, 703-711.

[MR2012] Ming Y, Ruan Q. Robust sparse bounding sphere for 3D face recognition. Image and Vision Computing, 2012, 30:524-534.

[MSKS2004] Ma Y, Soatto S, Kosecka J, Sastry S S. An Invitation to 3-D Vision: From Images to Geometric Models, Springer, 2004.

[MSXC2012] Mukherjee L, Singh V, Xu J, Collins M D. Analyzing the subspace structure of related images: Concurrent segmentation of image sets. In: Proceedings of European Conference on Computer Vision (ECCV), 2012, 128-142.

[Nem1994] Nemirovski A. Efficient Methods in Convex Programming, Lecture Notes, 1994.

[Nest1983] Nesterov, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 1983, 27(2):372-376.

[PGWXM2012] Peng Y, Ganesh A, Wright J, Xu W, Ma Y. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2233-2246.

[PK2012] Panagakis Y, Kotropoulos C. Automatic music tagging by low-rank representation. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, 497-500.

[RL2013] Ren X, Lin Z. Linearized alternating direction method with adaptive penalty and warm starts for fast solving transform invariant low-rank textures. Accepted by International Journal of Computer Vision, 2013.

[Rock1970] Rockafellar R. Convex Analysis, Princeton University Press, 1970.

[Tso1981] Tso M K-S. Reduced-rank regression and canonical analysis. Journal of the Royal Statistical Society, Series B (Methodological), 1981, 43(2): 183-189.

[WGMM2012] Wright J, Ganesh A, Min K, Ma Y. Compressive principal component pursuit. In: International Symposium on Information Theory (ISIT), 2012, 1276-1280.

[WGRM2009] Wright J, Ganesh A, Rao S, Ma Y. Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), 2009, 2080-2088.

[WGSMWM2010] Wu L, Ganesh A, Shi B, Matsushita Y, Wang Y, Ma Y. Robust photometric stereo via low-rank matrix completion and recovery. In: Proceedings of Asian Conference on Computer Vision (ACCV), 2010, vol. 3, 703-717.

[WL2010] Wei S, Lin Z. Analysis and improvement of low rank representation for subspace segmentation.

arXiv:1107.1561, 2010.

[ZGLA2012] Zhang T, Ghanem B, Liu S, Ahuja N. Low-rank sparse learning for robust visual tracking. In: Proceedings of European Conference on Computer Vision (ECCV), 2012, 470-484.

[ZGLM2012] Zhang Z, Ganesh A, Liang X, Ma Y. TILT: Transform invariant low-rank textures. International Journal of Computer Vision, 2012, 99(1): 1-24.

[ZGLMZY2012] Zhuang L, Gao H, Lin Z, Ma Y, Zhang X, Yu N. Non-negative low rank and sparse graph for semi-supervised learning. In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, 2328-2335.

[ZL2011] Zuo W, Lin Z. A generalized accelerated proximal gradient approach for total-variation-based image restoration. IEEE Transactions on Image Processing, 2011, 20(10):2748-2759.

[ZLM2011] Zhang Z, Liang X, Ma Y. Unwrapping low-rank textures on generalized cylindrical surfaces. In: Proceedings of International Conference on Computer Vision (ICCV), 2011, 1347-1354.

[ZLWCM2010] Zhou Z, Li X, Wright J, Candès E J, Ma Y. Stable principal component pursuit. In: International Symposium on Information Theory (ISIT), 2010, 1518 -1522.

[ZMM2011] Zhang Z, Matsushita Y, Ma Y. Camera calibration with lens distortion from low-rank textures. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011: 2321-2328.

[ZYM2010] Zhu G, Yan S, Ma Y. Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceedings of ACM Multimedia, 2010, 461-470.