

指导教师： 杨涛

提交时间：

2015/3/29

The task of
Digital Image Processing

数字图像处理

School of Computer Science

No: 1

姓名：高一凡

学号：2012302467

班号：10011204

格拉斯曼平均可扩展的鲁棒性 PCA

Søren Hauberg
DTU Compute*
Lyngby, Denmark
sohau@dtu.dk

Aasa Feragen
DIKU and MPIs Tübingen*
Denmark and Germany
aasa@diku.dk

Michael J. Black
MPI for Intelligent Systems
Tübingen, Germany
black@tue.mpg.de

摘要

随着大数据收集的快速自发增长，极端值的出现也在日益增多-“大数据”暗示“更多极端值”。与此同时 PCA（主要成分分析）主要用于缩减数据规模，并且我们还有更好的结果，众所周知，一个极端值可以破坏整个结果。不幸的是顶尖的鲁棒性 PCA 也不能衡量大数据。为了解决这个问题，我们来介绍一下格拉斯曼平均（Grassmann Average）简称(GA)，这个表示了降维相当于一个平均子空间的数据跨越。因为平均值可以很好很高效地估算，我们就可以立即得到可扩展性。GA 天生相对于 PCA 有更好的鲁棒性，但是我们发现他们对高斯数据表现一致。我们利用均值可以制造更强大的 Robust Grassmann Average (RGA)作为一种鲁棒性更强的 PCA。鲁棒性可以相对于向量或向量的元素；我们关注于后者并且使用修建均值。由此产生的修剪格拉斯曼平均（TGA）特别适合于计算机视觉，因为它是异常强大的像素。该算法具有最低的计算复杂度和较小的内存需求，使其可以扩展到“大噪声数据”。我们证明 TGA 可以用于背景建模，视频修复，阴影处理。我们通过展示强大的 PCA 在整个 Star Wars IV 上的表现来显示可扩展性。

1.介绍

在许多科学领域，应用领域中，主成分分析（PCA）是其中应用最为广泛的降维，建模，和分析数据的方法。这是一个计算机视觉领域的核心技术。

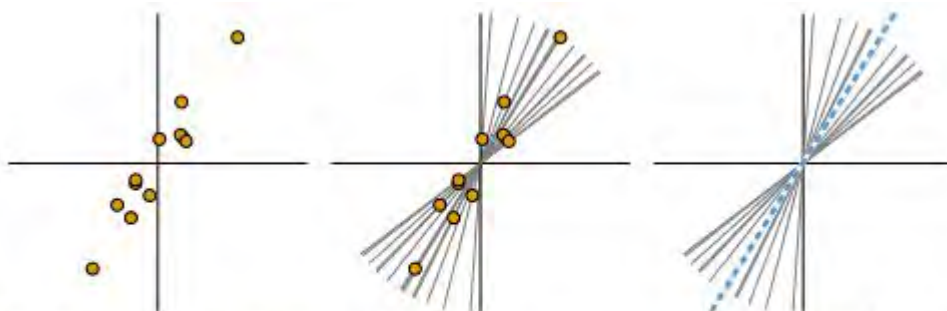


Figure 1. *Left:* A zero-mean dataset represented as a set of points. *Center:* The same data represented as a set of one-dimensional subspaces. *Right:* The blue dotted subspace is the average.

然而现存的升级 PCA 针对大数据级数，还有一个根本的问题没有解决。大数据集往往是自动产生的所以很难由人类来验证。所以“大数据”暗示着“更多极端值”。不过这里有几种方法提升 PCA 对极端值的鲁棒性【6,7,9,11】这些通常的方法没法扩展大数据。特别的是，鲁棒性的 PCA 方法不能扩展打大数据，可以扩展的方法又没有鲁棒性。我们的贡献是一个新的鲁棒性 PCA 公式和可扩展的运算法则，相较于之前方法。

我们的第一个贡献是制定子空间估计作为子空间计算的平均值。给一个平均值为零的 $X1:N$ 包含于 R^d ，我们观察到每个 X 跨越一个一维子空间。数学上，每一个子空间可以被描述

成 Grassmann manifold 上的一点。然后我们制定一个平均算子在 Grassmann manifold 上。这表示出跨越整个数据集的平均子空间概念。看图一作为事例。我们看一下 GA 是如何与标准 PCA 关联起来的并且证明他，对于高斯数据，GA 与标准 PCA 的子空间是相一致的。我们表示在形式上和实验中 GA 比 PCA 更鲁棒性，但是我们可以做得更好。

弄明白 PCA 怎样被一个均值估算的形式表示，是我们首要要观察的。具有鲁棒性的均值可以更有效的估算。我们基于这个事实来更准确的说明 RGA，它推广 PCA 到具有鲁棒性的 PCA。这里有一些强大的求平均的方法，任何一个都适合于 RGA 但是有一些更适合。

在许多领域和应用中，极端值存在一个数据向量中然而大部分的数据还是完美的。这是一个计算机视觉领域常见的情况，在一副图中一小部分像素被污染。我们发现 RGA 中的鲁棒性均值可以作用在像素层面上并且我们采用同样鲁棒的修剪均值的方法这样估算效率会很高。修建均值的方法 Trimmed Grassmann Average (TGA)可以解决鼓励的极端值或者在向量（图像）中的或者整个数据中的极端值。

由此产生的算法是简单地，理论上彻底的，估算上高效的。我们提供了一个详细的关于相关算法的鲁棒性和效率的比较。我们关注问题包含图像和图像序列但是方法一般化的。我们与 Candes et al.【7】比较实用他们的数据同时结果相似或更优秀。此外，我们相对于 Candes 和 De la Torre 和 Black [9]在档案影片的修复上有更好的结果。这里我们可以提供很多对比。最后我们展示了如何把它应用到视频方法尺度建模的问题上，这里所有以前的鲁棒性 PCA 都失效了。例如，我们用 TGA 估算星球大战 179415 帧的数据。而且我们是在网上实现的。

2.关联工作

主成分分析发现，投影观测值尽可能接近原始观测子空间。这种方法由一个最小二乘能量测量，这意味着异常值对结果有很大的影响。事实上，我们很容易看到甚至一个单独的极端值都可以任意破坏估计成分。我们说 PCA 有一个打破 0%的点【16】；如果超出 0%的数据是极端值我们不能相信这个结果。

鲁棒性 PCA: PCA 的这个弱点已经激发促进了大量关于强化扩展的工作。大多估计方法把每一个数据点看做一个内层点或一个外层点。这个工作可以通过对每个观测空间用鲁棒性协方差估计权重那么极端值就有较小的权重。其他热门的方法有迭代重新加权方案（iterative reweighting schemes）和最优距离和（optimizing the sum-of-distances）估测的子空间而不是平方距离和。Kwak 优化相同的 L1，和我们做的基本一样，但是他并没有与子空间均值联系，所以，不要结束我们的总像素鲁棒性方法。

在计算机世界领域中，最后欢迎的估计方法基本是来自 De la Torre a 和 Black [9] 和 Candes et al.他们允许像素的异常值（相对于具有整个图像视为一个孤立点）。De la Torre 和 Black 叙述鲁棒性子空间估计时在一个 M -estimators 的框架中并且由研究杰曼-麦克卢尔误差函数优化能量。[12]这种能量是利用梯度下降法与退火方案优化在杰曼-麦克卢尔误差函数的参数下。这就需要一些先验的通过一系列反复更新得到的奇异值分解。然而每一次重复这个体系只有线性的复杂度，基于 SVD 的依赖给出了一个整体立方的复杂度。这使得算法在大数据上不可行。由 Candes 等人的方法。[7]分解数据矩阵 $X \in \mathbb{R}^N \times D$ 为 $L + S$ ，其中 L 是最低 S 是稀疏的，即它包含异常值。在非常广泛的假设下，他们优雅的表演了这可以当做一个 convex program 解决，并提供了一系列的算法，其中所谓的非精确 ALM [31]是最常用的方法。该算法反复计算奇异值分解的数据给它一个立方的复杂性[14]。但是，这种复杂性渲染方法在任何超出小中等规模的数据大小下是不切实际的。这可以在一定程度上通过奇异值分解的并行实现缓解[29]，但这需要大量计算空间的可用性所以并不能从根本上解决复杂性问题。

近似方法: PCA 的不同鲁棒性公式的非线性复杂度引导了多种近似方法的发展。Mackey et al.[25]和 Lui [24] 提出抽样的算法来解决 Candes 的分解问题。这使得算法得以改进，他们只需要解决较小的奇异值分解问题。然而分解方法的理论分析表明他能提供良好的近似值，当分析大量数据要忽略大部分是这种方法总令人失望。在相同的思路中，Mu 提出用一个在优化组合中通过随机预测发现的低级副本代替数据矩阵。这又允许计算收益较小的奇异值分解问题，需要计算部分的数据是没有用的。

可升级的 PCA: 最著名的方法执行 PCA 可以通过对数据的特征值分解协方差矩阵，或通过数据矩阵的奇异值分解[19]。前者只能在低维的情况下，其中的协方差可以很容易地存储在内存中，后者在当数观察或维数太大时是不切实际的。

考虑到 PCA 在数据分析中的重要作用，可升级的成就都是现成的。一个平常的选择就是 EM PCA 公式，这个公式估算每一个主要成分通过重复迭代直至收敛。

$$\tilde{\mathbf{v}} \leftarrow \sum_{n=1}^N (\mathbf{x}_n^T \mathbf{v}_{i-1}) \mathbf{x}_n \quad \text{and} \quad \mathbf{v}_i \leftarrow \frac{\tilde{\mathbf{v}}}{\|\tilde{\mathbf{v}}\|} \quad (1)$$

这里 \mathbf{v}_i 表示主成分迭代中公式的第 i 项。这个也可以衍生成 EM 公式在无先验信息，作为一个梯度下降法，或一个幂迭代。一个公式迭代有线性的复杂度并且越少的内存使用是越有更高的扩展性。理想中我们想要一个鲁棒性的公式有着相同的特点。

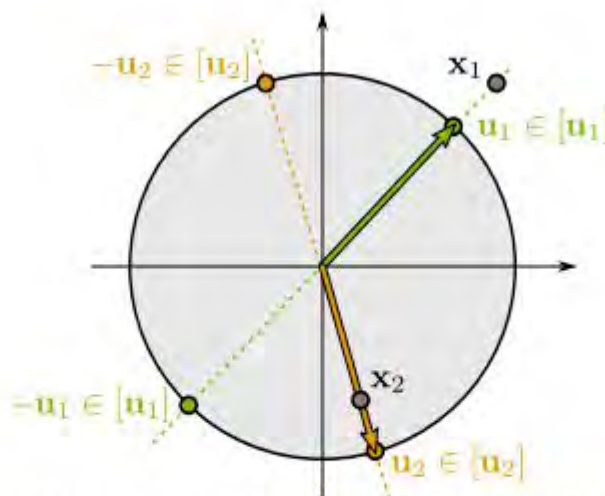


Figure 2. An illustration of the Grassmann manifold $\text{Gr}(1, 2)$ of 1-dimensional subspaces of \mathbb{R}^2 . Any subspace is represented by a point on the unit sphere with the further identification that opposing points on the sphere correspond to the same point on $\text{Gr}(1, D)$.

3. 格拉斯曼平均

我们的主要贡献是制定降维的子空间的平均估计的数据跨越。非正式地，如果我们寻求一个返回子空间的平均操作，那么这个操作的输入必须是子空间的集合。正式这种想法我们需要“空间的子空间的概念”，“这是用经典的几何结构称为格拉斯曼流形 (Grassmann manifold)

定义 1 格拉斯曼流形 $\text{Gr}(k, d)$ 是所有的 \mathbb{R}^d 的 k 维线性子空间的空间。

假设我们有一平均为零的数据集 $\mathbf{X} 1: N$ ，然后每一个观察的一维子空间跨越 \mathbb{R}^d ，因此，在 Gr 上就是一点 $(1, d)$ 。一维子空间需要一个特别简单的 \mathbb{R}^d 因为这样的任何子空间可以

由一个单位向量表现为 U 或其对映体 $-U$ 。因此 $Gr(1, D)$ 可写作为商 $S^{D-1}/\{\pm 1\}$ 相对于单位圆和组内的相反作用。我们表示点的 $Gr(1, D)$ 为 $[U] = \{u, -u\}$ ，记为 $[U] = [-U]$ 。换句话说，给定一个观察 X_n 我们替代跨越的子空间为 $[U_n] \pm x N / \|X\|$ ，其中符号是未知的。这是在图 2 所示

3.1 平均空间

对于均值为零的数据，我们现在定义对应于第一主成分。加权平均数通常被定义为最小的加权值的平方距离；即其中 $w_1: n$ 的权重和 $dist_{Gr(1, D)}$ 是一个距离 $Gr(1, D)$ 。我们定义这个距离是通过商空间构建 $Gr(1, D)$ 为 [5, 65 页] 这里是一个有待 $dist_{S^{D-1}}$ 定义

$$\begin{aligned} dist_{Gr(1, D)}^2([u_1], [u_2]) \\ = \min\{dist_{S^{D-1}}^2(v_1, v_2) | v_1 \in [u_1], v_2 \in [u_2]\}, \end{aligned} \quad (3)$$

距离单位球面上的 S^{D-1} 。这个距离将被选择为计算平均 Gr 的一个高效的算法 (1, d)，但首先我们证明一个有用的关系在平均与 $Gr(1, d)$ 和在单位球面 S^{D-1} 之间：

引理 1 对于任何加权平均 $[q] \in Gr(1, D)$ 满足公式 2，并且任意选择 $q \in [q] \in S^{D-1}$ ，这里存在 $u_1: N \in [u_1: N] \in S^{D-1}$ 所以 Q 是一个 S^{D-1} 上的加权均值。

$$q = \arg \min_{v \in S^{D-1}} \sum_{n=1}^N w_n dist_{S^{D-1}}^2(u_n, v). \quad (4)$$

证明选取任意 $q \in [q]$ 。注意，公式 3 所以 $dist_{S^{D-1}}(q, U_n) \geq dist_{Gr(1, D)}([q], [U_n])$ ，对于所有的 n 且 $U_n \in [U_n]$ ，因此，在公式 4 中找到的最小值不会比公式 2 中找到的更小。

此外，我们可以找到 $U_n \in [U_n]$ ， $dist_{S^{D-1}}(U_n, q) = dist_{Gr(1, D)}([U_n], [q])$ 。但是 q 是公式 4 中的最小值并且是 S^{D-1} 上的加权平均。引理 1 是有用的因为它减少了对 Gr 的计算问题 (1,

D) 来计算的问题对 S^{D-1} 手段，优化在可能的替代 $\pm u_1: N \in [u_1: N]$ 。

3.2 估算子空间均值

为了估算 $Gr(1, D)$ 上的均值，我们提出一个简单的算法，重复以下步骤直至收敛。

1. 选取替代值 U_1 在 $[u_1: N]$ 之中，选最接近当前 $dist_{S^{D-1}}$ 下的平均估计 q 。

2. 更新 q 作为 U_1 的球形平均。

为了这个公式的实用性，我们需要有能力快速的估计球形平均。带着这样的想法，我们选择

$$\text{dist}_{S^{D-1}}^2(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2} \|\mathbf{u}_1 - \mathbf{u}_2\|^2 = 1 - \mathbf{u}_1^T \mathbf{u}_2. \quad (5)$$

再这样的距离之下， $\mathbf{u}_{1:N}$ 的加权平均可以用近似下面的格式表示

$$\mathbf{q} = \frac{\boldsymbol{\mu}(w_{1:N}, \mathbf{u}_{1:N})}{\|\boldsymbol{\mu}(w_{1:N}, \mathbf{u}_{1:N})\|}, \quad (6)$$

这里

$$\boldsymbol{\mu}(w_{1:N}, \mathbf{u}_{1:N}) = \left(\sum_{n=1}^N w_n \right)^{-1} \sum_{n=1}^N w_n \mathbf{u}_n \quad (7)$$

表示欧氏加权平均。我们需要规范数据 \mathbf{x}_n ：N 前查看每个单位长度观察作为一种空间自然是要挑选权重 $w_n = \|\mathbf{x}_n\|$ 。这种选择，我们得到以下迭代计算平均空间方案：

Algorithm: Grassmann Average (GA)

$$w_n \leftarrow \text{sign}(\mathbf{u}_n^T \mathbf{q}_{i-1}) \|\mathbf{x}_n\|, \quad \mathbf{q}_i \leftarrow \frac{\boldsymbol{\mu}(w_{1:N}, \mathbf{u}_{1:N})}{\|\boldsymbol{\mu}(w_{1:N}, \mathbf{u}_{1:N})\|}, \quad (8)$$

这里 $w_n = \|\mathbf{x}_n\|$ 和 i 表示迭代的次数。我们用一个随机单位向量来初始化这个算式。

在前面 3.2.1 到 3.2.3 我们讨论了如何扩展估计多成分的方法，显示公式 8 与普通 PCA 的高斯数据齐是第一主成分。

3.2.1 多成分

上述算法只计算主要成分的数据。我们计算的附加组件在一个贪婪的通过添加组件的进一步约束时应该是正交的。以下 EM PCA [28]，我们实现这从数据中去除估计的成分计算从剩余的下一个组件。正式，如果 $\mathbf{X} \in \mathbb{R}^{N \times D}$ 表示的数据矩阵，则 $\mathbf{X} \mathbf{X}^T \mathbf{Q} \mathbf{Q}^T$ 是更新后的数据矩阵，组成 \mathbf{Q} 除去。然后下一个组件可以计算 \mathbf{X} 以上。这种方法是最优的高斯数据，但我们没有其他分布结果。这个贪婪的方法是有问题的 EM PCA 作为一个估计成分不准确影响下列组件。这不是遗传算法实例它返回一个精确的结果。

3.2.2 能量最小化的观点

被公式 8 替代的 GA 算法衍生一个估算跨越整个数据的子空间平均值计算，但是这个可以证明能量被优化了。从公式 3 和 5 我们看出

$$\text{dist}_{\text{Gr}(1,D)}^2([\mathbf{u}_1], [\mathbf{u}_2]) = 1 - |\mathbf{u}_1^T \mathbf{u}_2|, \quad (9)$$

并且公式 2 中的均值减少

$$[\mathbf{q}] = \arg \max_{\mathbf{v} \in S^{D-1}} \sum_{n=1}^N |\mathbf{x}_n^T \mathbf{v}|. \quad (10)$$

这是一个在投影追踪家族中的能量并且也被 Kwak 一直研究。Kwak 证明了 GA 公式 8 遗传算法有局部最优解。

$$\mathbf{q}_{\text{PCA}} = \arg \max_{\mathbf{v} \in S^{D-1}} \sum_{n=1}^N (\mathbf{x}_n^T \mathbf{v})^2. \quad (11)$$

我们看到，GA 可以比预期有更多的弹性对于异常值相比普通 PCA 平方已被取代为绝对值。然而，由于 GA 依赖于一个普通平均 (7) 我们看到，它仍然对极端值敏感异常。在 Sec3.3 我们将用强大的平均值改善这种现象。

3.2.3 与 PCA 的关系 (高斯数据)

为了战士 GS 是如何与普通 PCA 关联的，我们现在展示他们他们的观测数据服从正态分布。让 $\mathbf{x}_{1:N} \sim N(0, \Sigma)$ 从正态分布简化而来。预期值公式 10 是

$$\arg \max_{\mathbf{v} \in S^{D-1}} \mathbb{E} \left(\sum_{n=1}^N |\mathbf{v}^T \mathbf{x}_n| \right) = \arg \max_{\mathbf{v} \in S^{D-1}} \sum_{n=1}^N \mathbb{E}(|\mathbf{v}^T \mathbf{x}_n|). \quad (12)$$

定理 2 RD 子空间的预期 (12) 价值的 X1 GA: n 符合预期第一主成分。

证明 由于 \mathbf{x}_n 为正态分布的采样， $\mathbf{v}^T \mathbf{x}_n$ 遵循单变量正态分布的预测 $N(0, 2 \sigma^2)$ [4, 2.3]。因此， $|\mathbf{v}^T \mathbf{x}_n|$ 按照预期值 $\sigma \sqrt{2}$ 半正态分布比例。 σ 标准偏 V 是最大的时候，V 是主要的特征向量 Σ ，因此算法的期望值与预期的第一主成分的定用普通的 PCA。

下面介绍扩展到多个主成分，我们实验验证的结果。

3.3 鲁棒性 Grassmann Averages

GA 算法的核心计算部分 (8) 是球形平均 (6)，其中，反过来，仅仅计算归一化欧氏平均 (7)。由此，我们看到，即使是一个单一的孤立点可以任意破坏结果。

这个问题的一个直接的解决方案是取代平均一个强大的平均，给出以下方案：

Algorithm: Robust Grassmann Average (RGA)

$$w_n \leftarrow \text{sign}(\mathbf{u}_n^T \mathbf{q}_{i-1}) \|\mathbf{x}_n\|, \quad \mathbf{q}_i \leftarrow \frac{\boldsymbol{\mu}_{\text{rob}}(w_{1:N}, \mathbf{u}_{1:N})}{\|\boldsymbol{\mu}_{\text{rob}}(w_{1:N}, \mathbf{u}_{1:N})\|}, \quad (13)$$

这里 $\mathbf{u}_n = \mathbf{x}_n / \|\mathbf{x}_n\|$ 和 $\boldsymbol{\mu}_{\text{rob}}$ 表示任何鲁棒性均值。

我们把这种方法称为鲁棒格拉斯曼平均 (RGA)。在这里， $\boldsymbol{\mu}_{\text{rob}}$ 可以被选择为具有鲁棒性能该应用程序相关的。在计算机视觉应用中，我们经常处理图像数据的单个像素被损坏。因此，在计算平均图像鲁棒性是对每一个像素的基础上做普通，例，为一系列的一维鲁棒平均。一个标准的方法创建一个强大的一维平均修剪平均 [16]。这消除了最大和最小的观察计算平均之前。如果我们删除的数据 P% 从两边，我们得到一个打破 P% 点。对于 P = 50% 得到的最大鲁棒中值 [16]。

建立一个空间估计是强大的像素的异常值，我们选择 $\boldsymbol{\mu}_{\text{ROB}}$ 为像素的均值。我们把产生的

估计修剪的格拉斯曼平均 (TGA) 和让 $TGA(P, K)$ 表示估计发现 K 组平均修剪 $P\%$ 的最小和最大的数据元素。注意: P 和 K 是唯一的参数估计。每一次迭代的 $TGA(P, K)$ 的计算复杂度为 $O(KND)$ 作为修剪平均值可以计算使用选择算法在线性时间 [8, 9] §。我们有, 因此, 设计一个健壮的子空间估计方法, 可以处理像素的异常具有 GA 相同的计算复杂度 (8) 和 EM PCA (1)

4.结果

在这一节中我们提供了不同的任务结果, 其中我们与 Candes 等人的方法比较。[7] 和 Torre 和黑 [9]。为了比较的目的, 本强大的格拉斯曼平均也在 MATLAB 中实现, 所有的实验是在一个英特尔至强 w3680 进行 24 GB 的内存。当使用 TGA 我们减去像素的中值获得鲁棒估计的“零”数据集, 我们装饰在 50% 时计算子空间。对于 [7, 9] 利用公布的代码默认参数。这是进一步讨论 [15]。

在章节 4.1 中, 我们提供的鲁棒 PCA 的应用档案影片数据的恢复, 而在秒。4.2 和 4.3 我们重复实验, 从 [7] 在以前使用的数据以及新的数据。我们定量比较不同方法在 SEC 的鲁棒性。

4.4、如何

不同的方法在大数据规模秒。4.5。额外的实验, 视频, 和细节在 [15]。

4.1 视频恢复

档案影片通常包含划痕, 门毛, 接头, 灰尘, 眼泪, 亮度闪烁和其他形式的退化。

恢复膜通常包括人工识别和修复这种损伤。在不同的时间和空间, 我们把它们统称为像素级的异常。假设一个静态的相机, 我们执行对帧的鲁棒 PCA 使用强大的子空间重构的电影。我们执行利用正交投影重建, 但强大的替代的存在 [3]; 看到有值得更详细的讨论的问题。作为一个例子, 我们考虑一个场景从 1922 经典电影鬼。我们还原框架使用 TGA 和 [7, 9]。TGA 和 [9] 我们计算 80 部分组成, 其中对应于场景中的帧的数量。有代表性的框架, 如图 3 所示,

恢复电影在 [15]。TGA 不加权异常值去除或而大多保留内点。de la Torre 和 Black [9] 平滑的框架和大幅减少的视觉质量视频。这是在补充视频非常明显。

请注意, 这不是一个完全通用的算法的电影恢复但相机稳定的修改和处理快速移动的区域, TGA 可形成在电影修复系统的核心部件, 参见 [18]。

量化算法的性能, 我们产生人为损坏的视频以检测异常的能量, 对他们创造一个异常面具, 把这些面具并随机产生新的孤立点的面具, 然后应用他们从最近的好莱坞电影画面干净。以测量不同重建的准确性, 我们测量之间的平均绝对偏差和重建图像在像素噪声的地方添加。这是在表 1 中, TGA 显示一致比其他算法更好的因为他们往往 oversmooth 结果。在这里, 我们使用一个像素的中值; 即 50% 边。一块这样的微调参数的影响在 [15]。的完整性, 表 1 还显示了用普通的 PCA 算法达到的精度 (8)。

4.2 背景建模

我们重复由 Candes 等人做的实验。[7] 在不同光照条件下的背景建模。静态摄像机记录现场行走的人, 而照明条件的变化在整个序列。鲁棒 PCA 应该能够捕获光照差, 背景可以减去。第一序列, 在机场 [23] 记录, 使用 Candes 等人与 DE 算法的比较 Torre 和 Black [9]。由于缺乏空间, 我们只比较与 [7], 这是报告给最好的结果。图 4 (左) 显示选择帧从机场序列以及不精确的帧的重构 ALM 和 TGA 50% 边和 5 个组成部分。看到还附带视频。在一般情况下, TGA 给明显减少鬼影 [7]。

我们重复这个实验的另一个视频序列从 caviar2 数据集。结果, 主要表现在图 4 (右), 确认从机场序列结果热重法 (50%, 5) 产生明显较少的重影比 [7]



Figure 3. Representative frames from the 1922 film *Nosferatu* as well as their restoration using TGA and the algorithms from [7, 9]. TGA removes many outliers and generally improves the visual quality of the film, while Inexact ALM oversmooths the results and De la Torre and Black oversmooth and introduce artifacts. This is also seen in the absolute difference between the data and the reconstruction, which is also shown (inverted and multiplied by 2 for contrast purposes). We also refer the reader to the supplementary video [15].

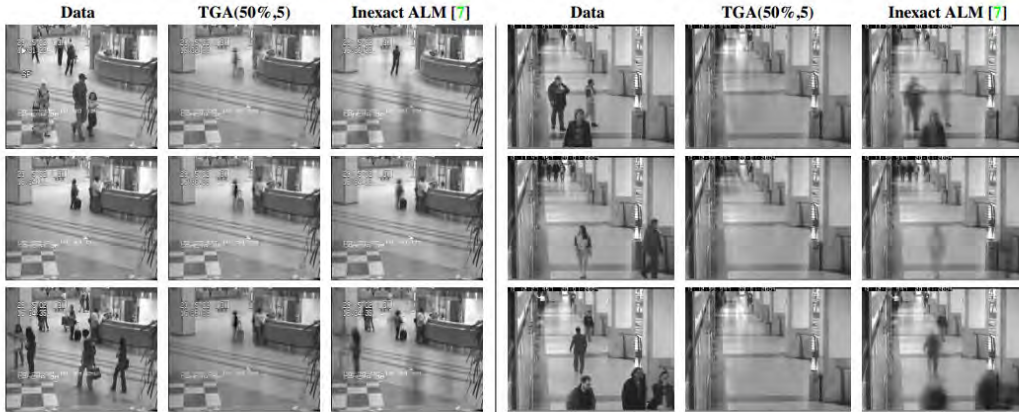


Figure 4. Reconstruction of the background from video sequences with changing illumination; three representative frames. *Left*: the airport sequence (3584 frames), which was also used in [7]. *Right*: a sequence from the CAVIAR dataset (1500 frames).

| | Groundhog Day (85 frames) | Pulp Fiction (85 frames) |
|---------------------------|------------------------------|-----------------------------|
| TGA(50%,80) | 0.0157 | 0.0404 |
| Inexact ALM [7] | 0.0168 | 0.0443 |
| De la Torre and Black [9] | 0.0349 | 0.0599 |
| GA (80 comp) | 0.3551 | 0.3773 |
| PCA (80 comp) | 0.3593 | 0.3789 |

Table 1. The mean absolute reconstruction error of the different algorithms on recent movies with added noise estimated from *Nosferatu*. TGA is consistently better as it does not oversmooth.

4.3 人脸数据的阴影去除

我们也重复阴影去除实验由 Candes 等人。[7]。扩展 Yale 人脸数据库 B [13] 包含不同的光照条件下的人脸图像，和，因此，不同的阴影。假设面凸朗伯体，他们应该躺在附近一九维线性空间[2]。相对于这个模型，阴影可以被认为异常值，我们应该能够消除他们的鲁棒 PCA。我们，因此，估计无阴影的图像使用 TGA (50%, 9)。

我们研究了两人在 Candes 等人。[7]。图 5 显示了原始数据，重建不同的算法以及图像和重建之间的绝对差。虽然没有“地面真理”，TGA 和非精确 ALM [7] 做几乎同样在去除阴影，虽然 TGA 似乎保持更多的阴影和镜面反射而变化非精确 ALM 产生的结果似乎更哑光。

4.4 矢量级鲁棒性

到目前为止，我们考虑了异常值出现像素的例子。为了完整性，我们也考虑的情况下在整个

观测向量被损坏。我们采取两 425 架从当代好莱坞电影的剪辑（在 SEC 定量实验相同的场景。4.1）我们再添加越来越多的帧从第二层和衡量影响估计的成分。我们使用标准测量误差方差，表示 [11]，这是比所估计的方差之间的捕获量主成分，通过真实拍摄组件。在这里，我们只考虑这一部分强调的方法之间的差异；为 [7] 我们首先剔除，然后降低到一维使用 PCA 子空间。图 6（左）的结果显示。两 [7, 9] 有这个问题的困难，但这并不奇怪，因为他们是专为像素的异常。两 TGA 和 GA 处理数据更优雅。

总之，遗传算法具有很好的鲁棒性当有 vectorlevel 异常和是一个很好的基础的替代方法。当每一个向量有一些数据异常，遗传表现主成分分析（表 1），并提供了显著的鲁棒性分析。

4.5 可扩展性

格拉斯曼平均的一个重要特点是，算法的尺度优雅的大型数据集。显示在此，我们报告了不同的算法的运行时越来越多的观察。我们使用的视频资料与静态相机看着忙碌的停车记录在一天的时间很多。每个框架都有一个分辨率 240、我们认为 320×达 38000 帧 3 图 6。（右）显示运行的时间。非精确 ALM [7] 和从 de la Torre 和黑 [9] 很快成为算法不切实际的，由于其内存使用这两种算法的运行；内存不足超过 6000 观测。TGA，对另一方面，大致呈线性数据表。在这实验中，我们计算 100 部分组成，这是足够的捕捉到背景中的变化。比较，我们还表明，EM PCA [28] 运行时间。TGA 只不过是比 EM PCA 稍慢，这是一般公认为最实用的算法之间于普通的 PCA 上的大型数据集。这是很明显的一个鲁棒 PCA 算法可以实现的运行时间与普通的 PCA 4 在每个迭代的 EM。PCA 比 TGA 计算价格，我们发现所需要的迭代次数比通常较低 TGAEM PCA，这解释了类似的运行时间。38000 图像，TGA (50%, 100) 使用 2 的 134 次迭代每个 - 组件，以平均 50 次迭代。EM PCA，对另一方面，203 的平均使用 729 次迭代 3 -。为了进一步强调 TGA 的可扩展性，我们计算整个星球大战 IV 电影 20 大部件。这是 179415 帧分辨率为 352×153。计算这 20 个组成部分（见 [15]）花了 8.5 个小时一个英特尔至强 128 GB 内存 e52650。

5. 讨论

主成分分析是一种基本的工具减少数据分析维度。以前的工作解决了鲁棒性在数据载体，但没有规模 vectorelement 水平的大型数据集。这是麻烦大数据应用的可能性异常增加的数据采集自动化。在本文中，我们介绍了格拉斯曼平均（GA），这是一个简单的和高度可扩展的方法，子空间估计与高斯数据主成分分析。我们进一步表明这种方法可以鲁棒通过使用强大的平均，产生强大的格拉斯曼平均（RGA）。对于一个给定的应用程序，我们只需要定义了一个强大的平均产生一个合适的强大的空间估计。我们开发的修剪格拉斯曼平均（TGA），这是一个强大的子空间估计在工作向量元素的水平。这种复杂性为相同可扩展的 EM 方法 [28]，和实证的结果表明，TGA 不太慢而更强大。一个可扩展的鲁棒 PCA 算法的有效性打了许多新的应用。我们已经表明，TGA 执行以及在计算机视觉中的不同的任务，在那里可供选择的算法产生不良的结果或失败在所有运行。我们已经表明我们如何可以计算强大在整个电影的台式电脑部件理的时间。这可能使表示和搜索视频的新方法。此外，PCA 无处不在使 RGA 相关超出计算机视觉；如大发现在生物学数据，物理和天气预报。最后，我们的方法是基于标准的构建块快速修剪平均值。这使得它非常适合加速通过并行化和在线计算。

