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：常磊

学号：2012302500

班号：10011205

关于图像搜索的三角嵌入与民主聚合

Herve Jegou, Andrew Zisserman

摘要

我们这里考虑类似 SIFT 算法的一种设计思想：用一个一维向量去代表一张嵌入和聚合了一组局域斑块描述符的图片。但我们期望能够更加具体地去构建蔟代表向量，例如费舍尔向量或者 VLAD，虽然它们规模会偏小或者是中等。

我们写了两篇论文，都是旨在调整个人贡献在最后的总结中的大小。第一篇是关于一个新型的可以避免在编码过程中对绝对距离的依赖的嵌入方法。第二篇论文是关于一个“民主化”的可以进一步允许在聚合阶段让不相干的描述符产生相互作用的策略。

当利用标准的公共图像检索基准测量点以在我们的实验中显示出时，这些方法不仅弥补了利用中小型

向量代表图形艺术的缺乏，而且有了一个大幅度的性能提升。

1、 介绍

现在考虑用一组描述向量描述图片的问题，例如一组 SIFT 描述符，通过一个一维设定向量可以使得这样的组矢量的简单对比用 cosin 相似度反应原始集的相似性。这就是在文献中的许多论文花费相当大的篇幅来描述关于大规模的图像检索问题。解决方案第一步是由一组向量（特征集合）来描述，且其内每一条都代表图像的子部分,然后被转换为基于聚合策略的单个载体，例如视觉词汇包（BOW）表示[30]，BOW 和 MULTIPLE-[12,14]或软任务[25,32]，还有局部性约束线性编码[33]，VLAD 或者是费舍尔向量。类似的方法也用于大规模图像分类，但是我们这里主要讨论图像检索。

所有的这些方法都可以被分解成以下两步：嵌入步骤单独地映射集合中的每个矢量以一个高维空间；而聚合步骤产生从该组映射向量的的单个向量，例如使用 sum-或者 max-pooling。我们重温这两个步骤并且用每一个做出新颖的贡献。我们的总体目标是设计一个“民主”的内核，使得该组中的每个矢量的贡献几乎同样至设定相似度。这个目标在嵌入和整合阶段都被

分开处理。

首先，我们的目标是设计出嵌入步骤φ，使得对于任意一对向量（X，Y）描述两个特征点，假如特征点是匹配的那么该相似度φ（x）Tφ（y）为趋于一致，反之则让他们趋近于零，而且φ(x)⊤φ(y)的为了不相干的特征点振幅应该尽可能小。为达到这个目的，我们的第一篇文章是介绍三角嵌入方式(T-embedding)，这是一种只用方向编码而不是用幅度去输入关于一组锚点的向量。与此相反，我们以最相似的现有技术[13,17,26]，并抛弃我们认为并不可信的介于输入矢量和锚点之间的幅度信息。从这个角度来看，我们的方案可以被认为是一种用三角策略来定位的方法。

我们的第二篇论文是关于集合战略，显然由于考虑到一组向量的干扰，故将其先删除，然后在两组向量之间给予同样大小的比重。这涉及到一个优化问题来找到权重线性地平衡在最终向量表示每个映射向量的贡献，同时用改进的 Sinkhorn 算法[15,29]解决了此优化问题。这个方法在相对较短的表示上有特别明显的优势，它在取消映射向量之间的干扰具有不可或缺的作用。由于要在大范围图像检索项目上面向公众基础人群展示，我们这两篇论文都在以往的技术上做出了十分重要的改进：我们的嵌入算法优于为给定维数而

占用大篇幅文章的费舍尔算法，而且我们的聚合策略提供了类似的增益，这也是互补的所谓的幂律正常化。

本文结构如下。第二部分介绍符号并以此展开文章的描述。第三部分介绍了我们的三角嵌入，第四部分介绍了我们的集合策略。实验内容列在第五部分中。补充材料在附录中提供，并且第一页中项目源代码也是可用的。

2、 正文

让我们来设定两个集合x和y，并且card(x)=m，card(y)=n。每一个集合都是又一簇向量组成，例如与图像相关的局部描述符。我们首先考虑由Bo 和 Sminchisescu [4]2中的框架派生出来的以下组成形式：

$$\mathrm{k}(\mathrm{x},\mathrm{y}) = \sum_{\mathrm{x}\in\mathrm{X}} \sum_{\mathrm{y}\in\mathrm{Y}} \mathrm{k}(\mathrm{x},\mathrm{y}) = \psi(x)^T \psi(y) \qquad (1)$$

这里的的k(x,y)是介于变量里的单独向量的内核。右边的式子表明我们要更加细致地考虑代表变量的向量，由此两幅图像的比较就可以建立在它们的变量表示$\psi(x)$和$\psi(y)$之中的内在结果。匹配内核也可以写成如下形式：

$$\mathrm{k}(\mathrm{x},\mathrm{y}) = 1_n^T K(x,y) 1_m \qquad (2)$$

这里$1_n = \underbrace{[1,\dots,1]}_{\times n}$，同时我们定义一个$\mathrm{n}\times\mathrm{m}$矩阵：

$$\mathrm{k(x,y)} = \begin{bmatrix} k[x1,y1] & \dots & k[x1,ym] \\ \vdots & \ddots & \vdots \\ k[xn,y1] & \dots & k[xn,ym] \end{bmatrix} \quad (3)$$

此矩阵通常包含两个图像的局部描述符之间所有成对的相似性。对于任何内核k，我们定义其对应的归一化形式为：

$$k^*(x,y) = \alpha(x)\alpha(y)k(x,y), \quad (4)$$

这里正规化$\alpha(.)$是被定义为像$k^*(x,x) = 1, i.e.,\ \alpha(\mathrm{x}) = k(x,y)^{-\frac{1}{2}}$这样的形式。

2.1. 结构：嵌入和聚合

我们将结构k分成两步，并分别称之为嵌入和聚合。嵌入步骤ϕ：$\mathbb{R}^d \to \mathbb{R}^D$，映射任意$\mathrm{x} \in \mathrm{X}$为

$$\mathrm{x} \mapsto \phi(\mathrm{x}) \quad (5)$$

聚合步骤是通过函数ψ从一列嵌入向量$\{\phi(\mathrm{x1}), \dots., \phi(\mathrm{xn})\}$中来计算单矢量。给这个函数做一个简单的总结，在此情况下我们表示ψ_s为：

$$\psi_s(x) = \sum_{x \in X} \phi(x) \quad (6)$$

这个对ψ的简单定义被隐性地用在(1)中。在这种情况下，$\mathrm{k(x,y)} = \langle \phi(x) | \phi(y) \rangle$。匹配核心k在通过计算其在聚合向量中元素的点积：

$$\psi_s(x)^T \psi_s(y) = \sum_{x \in X} \sum_{y \in Y} \phi(x)^T \phi(y) \quad (7)$$

其中，每个可能的匹配(x,y)有助于整体组相似性，并

且每个分量的权重都是$\phi(x)^T\phi(y)$。

这个构思，被认为特别是Bo和Sminchisescu [4]和 Tolias等人提出，包含了很多的实现途径。让我们首先来考虑嵌入步骤。对于视觉词汇包里面的c的绝对值大小|c|=D，某个属于X的单个描述符x被映射到这个D维矢量。非零位置的确定是基于最近分配规则。通过多个视觉词[14]的分配，当软分配[23,25,32]为统计到重心的距离时所赋予的不同权重时，多个组成成分被赋值为1。类似局部线性编码途径，费舍尔向量[22]或者是VLAD[13]也提供了可供选择的对ϕ的定义。幂律正常化[11,13,23]通过后处理聚合向量来改变函数ψ。

注意：所述嵌入步骤类似于编码步骤通常认为是在文献[13],并且公式(6)接近于汇集步骤[13]。我们使用其他术语来避免混淆，因为在我们的这种情况下，所有的操作上施加的每个描述符的基础都是被包含在嵌入阶段。在这方面，函数φ已经包含部分池化操作，包括基于几何的池化，例如一个空间金字塔。因此，在这个构想中，∅(x)的维数通常和集合X中最后一个代表向量相同。

2.2．匹配内核中的干扰

集合向量化的基础是通过投影一组描述符向量

到一个单矢量，其优点是可以产生一组与线性代数兼容的向量代表，这些在SVM和量子化理论中出现过，但是只是提及到少量。然而，这个步骤给了普通描述符在最后表示中不公平的重要性。更加精确地，通过比较$\psi_s(x)$至它本身，一个给定的向量x映射至相似性集合$\psi_s(\mathrm{x})^T\psi_s(x)$的贡献是通过给定的$\phi_s(\mathrm{x})^T\psi_s(x)$：

$$\phi(x)^T\sum_{x'\in\chi}\phi(x') = \|\phi(x)\|^2 + \phi(x)^T\sum_{x'\in\chi:x'\neq x}\phi(x') \quad (8)$$

这个公式表明了φ的两个重要属性：

1、 等式左边$\|\phi(x)\|^2$隔离出的匹配符，其贡献很大程度上（即二次性）取决于它的常态。

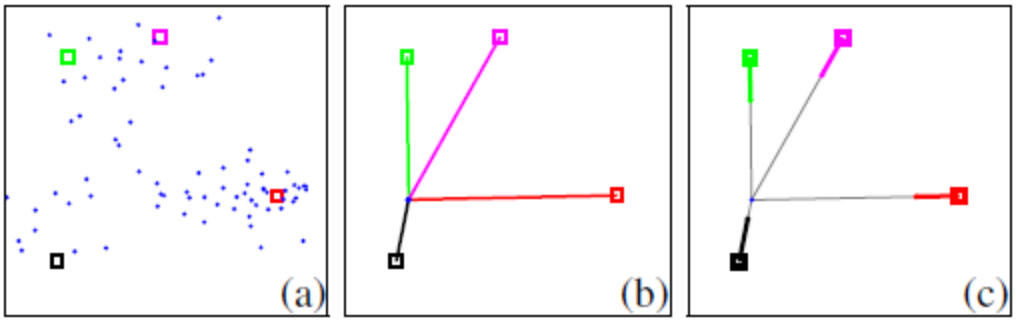

图1.在d=2和|C| = 4时嵌入法插图。
（a）分布和学习锚点；（b）与给定的矢量x相关联的残余矢量；（c）标准化残差R(x)。

2、 由于与其他向量的关系导致等式右边对于x的作用来说是一种“噪音”污染。

本文旨在解决上面两个问题，且此二者都是在嵌入函数ϕ(第三节)的设计和聚合函数ψ(第四节)中得到解决。

3、 三角嵌入

在这个部分中，我们来介绍T-embedding(三角嵌入法)函数ϕ_Δ。它提供了在(8)中提出的几个理想的特性意见，特别是两个无关特征之间的内积几乎是零，除此之外当特征值是相对于来自相同的分布得出其他特征之也是足够接近的。

3.1 框架

给一个向量x在D维单位球上的分布，我们考虑一个集合$\{c_1, \dots, c_{|C|}\}_i$，$c_i \in \mathbb{R}^d$，其中下角标|c|指的是锚点。这个集合是典型地通过K-means学习得到，并且和视觉词汇表相似。然而在我们的上下文中它更与为锚图而提出二进制编码的目的联系更为紧密。

与目前存在的大部分工作相反，我们着眼于方向性的信息，并丢弃绝对距离的锚点。这个策略跟割线流形有关[9]。这是为规避“带宽问题”的关键，也即在绝对距离的依赖性上通常是不可靠的。其结果是，我们的新颖矢量表示被三角理论隐式地定义。这是通过考虑一组归一化残差矢量取得：

$$r_j(x) = \left\{\frac{x-c_j}{\|x-c_j\|}\right\}，\text{其中j=1,...}|c|， \qquad (9)$$

其保留了x和c_j之间的角度信息，而丢弃了绝对幅度。图1说明了这种三角测量的策略。对于所有的j，

我们假设x ≠ c_j,这是如果c_j向量被k-均值(k-均值重心，作为不同的向量的平均水平，严格处于但单位球内)方法获得，那么ι_2-归一化SIFT向量就会得到保证。级联R(x) = $\left[T_1(x)^T, \ldots, T_{|c|}(x)^T\right]^T$是一个中间d维代表，且其值D=|c|×d。但这不仅冗余而且给了主向量过多的比重。我们随后美白（基于特征值中心，旋转和缩放）的代表性[17]。更确切地说，表示由Σ与随机变量R(x)的相关联的协方差矩阵，我们的T-嵌入由R(x)中所得：

$$\phi_\Delta(x) = \Sigma^{-1/2}(R(x) - R_0), \qquad (10)$$

这里$R_0 = \mathbb{E}_x[R(x)]$和Σ都是建立在训练集上由经验测量所得。

从图2可看出在原始空间，与每个特征向量相关联的值，也即，输出描述符$\phi_\Delta(x)$中的各成分组成。通过类比PCA或者Laplacian Eigenmaps[3]，最大的特征值与“低频”有关：相应的组分变化缓慢作为输入的描述符功能。与此相反，与具有小的特征值相关联的特征向量对应于高频率：一个给定的输入功能的微小变化都有相应的输出组件上产生较大的影响，正如从图2中可以看出其幅值变化从极大值(左面)变化至极小值(右面)。

结果是，我们的嵌入算法是在不同的分辨率层面

上来比较描述符。这也是在现有工作下像金字塔匹配核心[8]问题和语法树[21]问题，且其是通过使用不同的量子化手段施以不同的分辨率变化。在我们的例子中，并没有量子化产物：第一个组成部分反映的是粗糙的位置，而最后一个则是更加局域化。为了改善该描述符的定位，我们丢弃具有最大本征值相关联的对d第一组件。就像在附录A中所讨论的那样，这减少了不相关的描述符之间的余弦相似性的方差。因此嵌入式描述符$\phi_\Delta(x)$的最终维度是$\mathrm{D}=\mathrm{d}\times(|\mathrm{C}|-1)$。

3.2. 高效的计算和内核匹配

像在第二部分中介绍的那样，我们现在来考虑从我们的T-嵌入法中继承核心匹配。利用线性方程(6)和(10)，我们计算一个矢量集合χ中的含有n个向量作为显式集合表示。

$$\psi\left(\Phi_\Delta(\chi)\right)=\sum_{x\in\chi}\phi_\Delta(x)$$

$$=\Sigma^{-1/2}\left(\sum_{x\in\chi}R(x)\right)-n\Sigma^{-1/2}R_0$$

在四核笔记本上使用高效Matlab实现以上经典代表参数$(\mathrm{d}=128,\mathrm{n}=3000,|\mathrm{C}|=16)$耗时20ms。然而，在做聚合操作前归一化$\phi_\Delta$去确保$\phi_\Delta(x)^T\phi_\Delta(x)=$

1。特别是在第四节提出来的聚合技术对这个非常重要。这种情况下计算速度就慢了很多，因为每次$R(x)/|R(x)|$都要与$\Sigma^{-1/2}$做一次乘积运算。

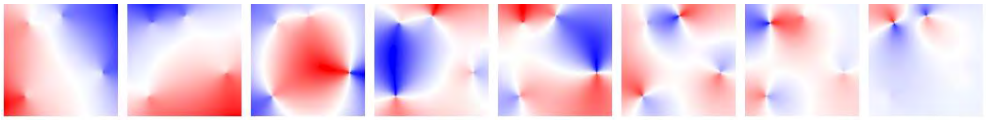

图2.与图1中相同的例子中函数ϕ_Δ的图形表示。$\phi_\Delta(x)$由8部分组成，其中在每个地图的空间位置表示2-D输入矢量x，幅度代表了输出量（红=负值，蓝=正值，白色=0）。与两个最大的特征值相关联的组件（左面）平滑的分布，同时丢弃的表示在最后。通过观察其他组件所送达的地方。

3.3 性能分析

通过这个框架构建，除了在点c处奇异外，我们的T-嵌入满足好几个理想特性。首先，函数ϕ：$\mathbb{R}^d \to \mathbb{R}^D$： $x \mapsto \phi(x)$是一一映射的。其次，除了在质心的位置，函数是处处连续的(正如上面标注出来的一样，在一个单位球中矢量分布(就像SIFT描述符) 和通过K-方式学习的锚点，因为其重心严格处于单位球内部，故它在单位球中处处连续，)。此属性确保输入矢量的微小变化不产生在输出空间的急剧变化。同样的，嵌入功能也处处可微。

注意到VLAD，LLC和费舍尔向量也都是一一映射的，但是词汇包(BOW)却不是这样的。在这三种嵌入方法中，只有费舍尔是连续的。

除了这些正规的数学性质外，我们的嵌入方法还有另外一个关键的性质，这将在后续中提到，其实也

就是两个不相关的向量的内积$\langle \phi_\Delta(x)|\, \phi_\Delta(y)\rangle$有很大的概率接近于零。相反的，在适用SIFT描述符的时候，相似区域与嵌入描述符有更大的可能大于0。

图像片数据集的定量分析。我们收集的余弦相似的相关和不相关的图像块经验统计。为达到这个目的，我们使用由Brown提供的Liberty和Nortredame数据集等等[34]。每个数据集包括来自多个图像的大约500K的图像块并分为150k的集群，其中一个集群对应于同一个物理场景点。这些数据集通常用于学习图像块信息[28，34]，但在这里我们使用它们用于学习和评估嵌入，并且简单地使用RootSIFT[1]作为图像块信息。在数据库Liberty里学习各种方法。我们在Nortredame通过考虑150k的对匹配描述符(每簇一对)和相同数量的无关描述符(我们采取从两个不同的簇中分别选取一个描述符)。图3展示出余弦相似性在相关或者不相关图像块中用原始描述符描述的情况，之后，它们分别映射到费舍尔矢量编码(没有聚合操作)和我们的T-嵌入算法中。费舍尔向量和T-嵌入都增加了无关和相关描述符之间的对比度。因此，T-嵌入要好于费舍尔。首先，在无关图像块中平均相似度比起ϕ_Δ是要更加接近于0的，只有少量无关对是与此不符的，此外，z

在费舍尔算法中，，相当大的比例（注意对数刻度）

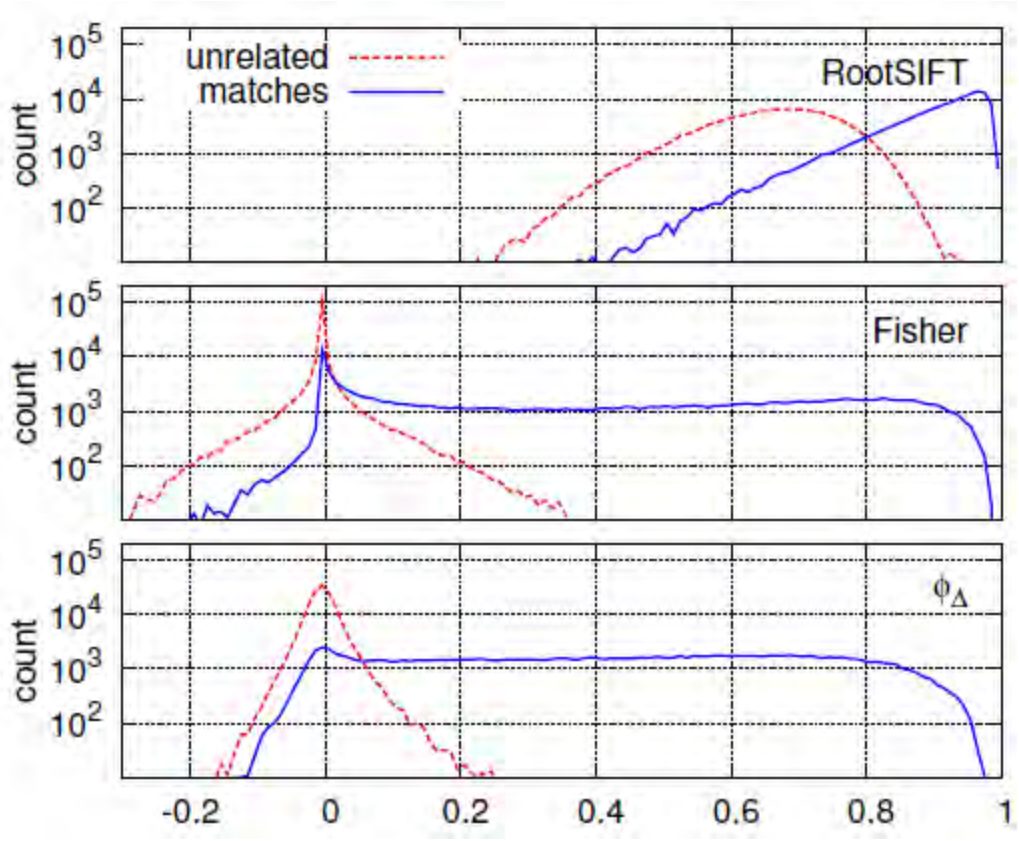

图3. 相关的（普通）和无关（虚线）补丁之间的余弦相似的直方图，对于RootSIFT描述符（顶部），是用费舍尔内核嵌入（中部，|c| = 16），我们的T-嵌入（底部，|c| = 16）。计数采用对数刻度。

的正确匹配都给出了接近0的相似度。这一比例在T-嵌入中相比要低得多。一个高ϕ_Δ余弦相似性有两个局部描述符x和y，可靠地反映了我们在相应的图像块中视觉相似性的信心。作为这种发现的副产品，它有可能在基于绝对ϕ_Δ余弦相似性时确定如何去接近图像块，作为可视化示出在图4中通过在一个给定的图像中检测类似的模式（脉冲串[11]）。映射至T-嵌入中相似性度量的质量在附录A中与ROC曲线一起来评

估。对图像块描述符的[28,34]的监督学习有助于改善在所有情况下的分离操作。

4、 民主聚集

我们的T-嵌入通过对无关描述符给一个余弦相似度几乎是0的条件来减少（8）中的干扰，同时对真实的匹配结果提供一个相对较高的积极评分。然而在这个阶段描述符仍需独立的来考虑。下面，我们进一步明确地分析和减少它们在聚合阶段的限制干扰。

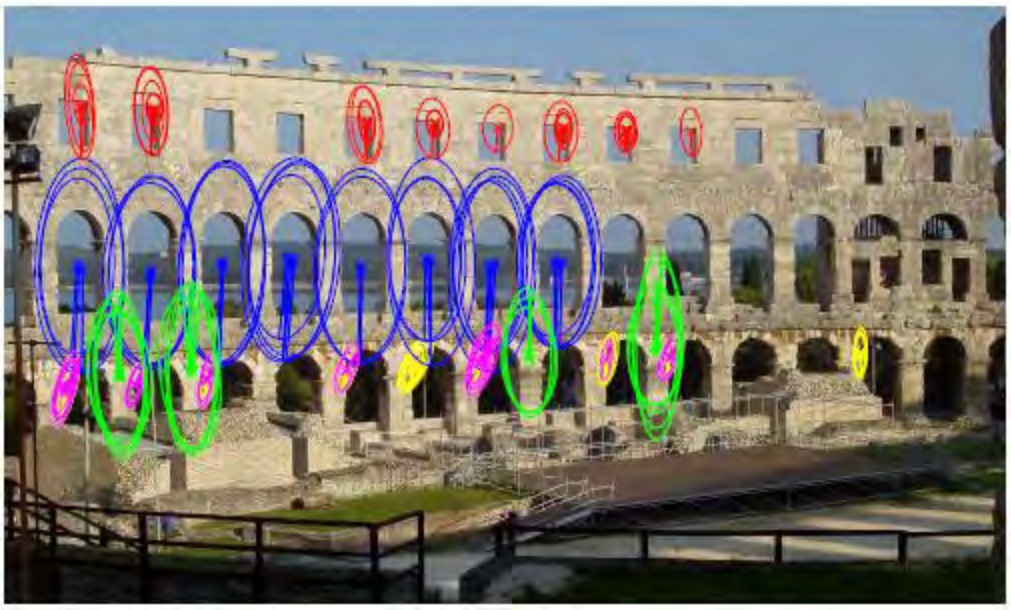

图4. 通过阈值处理，检测自相似的结构嵌入式描述符之间的余弦相似。为了产生它，我们只是描述映射到阈值处理之间的革兰氏矩阵K（带阈值0.5），并进行了相关的图形的连接成分分析。我们展示的五大组成部分（一种颜色为一个）。

匹配核心k当且仅当被定义为民主的，对于任何集合χ使得card(χ)=n，响应矩阵k满足

$$k(\chi,\chi)1_n = C1_n, \qquad (13)$$

这里缩放因子C或许是（或者不是）取决于χ。换句话说，一个民主核心确保在χ里面的所有向量对于集合

自身相似性都有一样的权重。在这剩下的这部分中，我们目前的优化问题是旨在聚合阶段可以任意地产生民主内核。然后我们讨论收敛问题并提出一种策略来达到收敛的效果。

4.1．民主化

内核像在（1）里面通常不是民主化的。为实现民主化这个性质，我们修改包括与之关联的每一个向量上的线性附加权重，例如

$\mathrm{k(x,y)} = \sum_{x\in\chi}\sum_{y\in Y}\lambda\chi(x)\lambda Y(y)k(x,y)$。　　（14）

每个量$\lambda\chi(x)$（$\lambda Y(y)$）仅取决于x和集合χ（分别有y和Y）。现在考虑集合$\chi = \{x_1, \ldots, x_n\}$，相应的权值$\lambda_i$，其中i=1,..,n，可通过求解来确定，实现的时候，该组方程

$\forall x_i \in \chi, \lambda_i \times \sum_{x_j\in\chi}\lambda_j k\left(x_i, x_j\right) = C$　　（15）

其约束条件为∀I，λ_i>0。问题总结成矩阵形式为

$\wedge \mathrm{K} \wedge 1_n = C1_n$，　　（16）

这里∧= $\mathrm{diag}(\lambda) = \mathrm{diag}(\lambda_{1,\ldots}\lambda_n)$是对角线严格的正矩阵。注意，式子（14）相当于定义了一个新的匹配的内核$k'(x,y) = \lambda\chi(x)\lambda Y(y)k(x,y)$。考虑在第二部分中介绍的匹配核心为“嵌入+聚合”这种特殊情况。方程（14）可以重写为：

$\mathrm{k(x,y)} = \sum_{x\in\chi}\sum_{y\in Y}\lambda\chi(x)\lambda Y(y)\phi(x)^T\phi(y)$　　（17）

$$= \left(\sum_{x\in\chi} \lambda_\chi(x)\,\phi(x)\right)^T \left(\sum_{y\in Y} \lambda_Y(y)\,\phi(y)\right),$$

其中可以看出，民主化是相当于定义一个替代函数ψ，记为ψ_d。更精确地说，我们通过加权求和取代（6）中的聚集函数ψ_s：

$\psi_d\big(\Phi(\chi)\big) = \sum_{\phi_i\in\Phi(\chi)} \lambda_i\phi_i$。

加权矢量通过ℓ2-标准化，以产生归一化匹配内核。

4.2. 修改Sinkhorn缩放算法

值得注意的是，这个问题类似于投影到一个双随机矩阵[29]:它等价于如果C=1且k是正数。在额外的假设条件（矩阵K具有全面支持并完全不可分解[15]）下，该Sinkhorn算法收敛到一个唯一的解决方案且满足∀i，λ_i>0。它是一种通过进行交替正常化的行和列的定点算法。我们采用由奈特[15]分析对称变体，削弱每次迭代的影响，基于最近提出的方法[14]，通过使用幂指数小于0.5为更平滑的收敛性。附录B给出了此优化策略的伪代码。Sinkhorn是一种可以快速收敛的算法。出于对效率的考虑我们经过10次迭代就停止了。实践表明，迭代更多次数并没有更多的好处。

4.3. 一般情况，收敛问题及其解决方案

在任意内核K(.,.)的情况下，假设与Sinkhorn收敛所需一般来说不满足的（矩阵k非负且完全不可分解

[15]）。因此，一个正解不必要存在。对有负值内核来说任何优化算法都或许会产生负权重，尤其是如果$\sum_j k(x_i, x_j) < 0$就会发生。而这种情况是不希望出现的，因为这意味着关于在集合中新的或者已经删除的向量的权重计算是很敏感的。我们通过采用下述的预处理步骤解决这个问题。

积极实施。在经过$\ell 2$-标准化$\phi_\Delta x$后，所有向量的比重是相同的，我们通过在k中设置所有的负值变为0以解决收敛问题。与新矩阵K^+用Sinkhorn算法进行权重计算可得结果为正，这是因为所有的行与列的相加结果是正的。由此产生的嵌入ψ_d不是严格意义上的民主内核，但趋向于更加“民主”。

4.4. 讨论

考虑在（8）中的右边部分。如果嵌入完全移除（相同集合间的描述符没有联系），然后我们的民主化是一个校准，这样所有的标准$\|\phi(x)\|$都变得一致。附录C也表明，我们的策略是等效于在视觉词汇包向量的情况下，而不需要倒排文件频率使得加权平方根逐元正常化。

5．实验

本节介绍的结果是我们的民主内核。这些新颖的成分形成我们的方法，唤作T-嵌入和民主聚合，且可

以单独使用。因此我们分别评估其影响，先只用T-嵌入进行实验（a）；（b）用民主化聚集法来应用到费舍尔嵌入中；而（c）用我们的两种方法。在整个本节中，我们仅使用归一化的内核k^*，这意味着该图像矢量是归一化为具有单位欧几里得范数。

5.1. 数据集和评估协议

我们采用的公共数据集和相应的评价协议，可经常在大规模图像搜索的情况下使用。所有的这些学习阶段，也即，为我们的T-嵌入算法而用K-均值聚类算法和学习进行投影，使用独特的图像采集而进行离线操作，并不包含对图像的索引或者是图像查询。

Oxford5k[24]是由5062张建筑物图像和对应于牛津11处截然不同的55张查询图像。搜索质量是通过在55处查询图像的平均值计算来衡量的。图像会被任意注释为相关的，不相关的，或者是垃圾，这就表明一个用户是否把图像考虑为相关与否是不清楚的。在以下建议中的协议，垃圾图像从排名中删去。在这Oxford5k试验中，所有的学习步骤是在Paris6k数据集[25]中进行。Oxford105k是Oxford5k与100k张参照图像，这是为了在一个大的范围中评估搜索质量。

INRIA Holiday [12]。该数据集包含1491张不同地点、

不同对象的照片，其中500张用来查询。搜索质量是通过mAP来测算的，且查询在排名列表中已经删除。为获取这些词汇，我们使用了由Holiday提供的独立数据集Flickr60k。对于Holiday和Oxford我们利用我们的方法进行了三次实验，并且报告了平均水平。

5.2. 实施注意事项

局域描述符是用Hessian矩阵仿射检测器[19]提取并用SIFT算子来描述。我们使用的是在前面文章中已经使用过的描述符。在我们的所有实验中都采用了RootSIFT变体。

幂律正常化。图像包含“视觉脉冲”[11]，例如在图4中可以看出，这意味着在同一幅图像中大量的描述符几乎是相同的。这些描述符在相似性中占有主导地位，甚至在“民主化”内核中也是。作为普通的后处理步骤[11,23]，我们在矢位图像代表中应用幂律归一化，随后用ℓ2-正常化它。该处理是由一个常数α参数化来修改控制的指数的值，这样
$a := |a|^{\alpha} sign(a)$。我们一般标准化地设置$\alpha = 0.5$以确保在这些方法中由公平的比较性。请注意，这部分还包括对该参数的具体分析。

旋转和规范化（RN）。幂律归一化能够抑制视觉脉

冲，但是频繁地共同出现会损害对该相似性的度量[6]。在VLAD，这个问题由增白矢量而得到解决[10]。然而在增白实施阶段需要输入大量的数据而且最小特征值产生伪影。这就造成这样的处理只有当表示非常短的时候才可以使用。作为一种替代方法[27]，我们在学习图像矢量（在学习组中）时利用PCA转置矩阵在旋转数据后应用幂律归一化，也即没有增白阶段。这产生一个类似于增白的效果，但是更加稳定并且比依赖PCA特征值。为了避免全特征值分解和在学习阶段过多的图像需求，我们计算第一个1000个特征值向量并利用革兰氏施密特正交化来对空间进行提示（正交补这些第一特征向量）来产生完整的基础。经过这次旋转变换，我们应用定期幂律正常化然后通过选择基础捕获这两个现象的第一个组件共同解决了突发和共同出现的问题。

5.3. 方法和参数的影响

与已经存在的技术相比较，我们的方法没有引进其他参数，除此外例如在Sinkhorn中设置的迭代次数对性能也是没有影响的。主征值的绝对值大小为c，并且描述符α与幂律归一化相关联。对oliday描述符的分析在图5里展示。Oxford5k的分析在中展示（补充材料）。得出的结论是在这两个数据集里都是相同

的。为了补充这些曲线，表1显示了通过对Ox-ford5k、oxford105k和Holidays中固定词汇量中一步步显示出我们方法的影响。

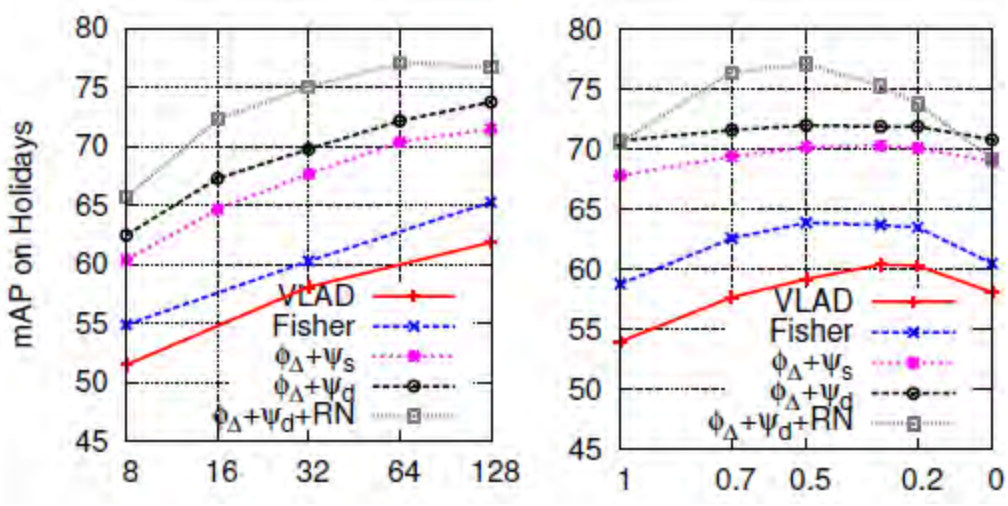

图5. 在Holiday中为不同图像向量表示性能参数的影响：VLAD、费舍尔和我们的使用聚集和ψ_s到T-嵌入ϕ_Δ中，还有民主化聚集ψ_d（使用或者不使用RN）。左边：词汇规模为|c|的函数；右边：描述符α的幂律归一化函数。注意，$\alpha = 0$具有二义性。

词汇量大小。对于所有的陈述，包括Tembedding和民主的聚集，表现为词汇量的增函数。作为参考，我们给出了改进Fisher基准线的表现。注意通过我们的嵌入方法对混合词汇量大小可以提供一个大的增益。我们的聚合方法给出了一个互补的增益。在词汇量较大时改善往往比较小：这是意料之中的，因为对于较大的词汇量间描述符之间的相互作用比小的时候更不重要。像|C|>128，民主聚合方法得到的好处还不值得计算开销。

我们的聚合策略在利用ϕ_Δ有了一个非常重要的飞

跃。就像期望的那样，在没有幂律归一化可以应用的情况下表现也是十分差强人意的。还有就是我们对描述符α的分析揭示了我们的聚合策略ψ_d对于幂律归一化有了重要的补充，这两个算法都改善了算法评分。

幂律归一化与RN。我们的方法对于幂律归一化重要性较小（右边曲线$\phi_\Delta + \psi_d$都更加平滑），除非我们采用RN：当我们使用（标准化）参数α=0.5时这个归一化策略有了一个重要的改善。

降维操作。为获得较短的表示，我们在RN归一化后通过我们的嵌入方法保持第一组D'组分。表1报告用于指出不同短矢量的表现，其中D’由128到1024。虽然由于维数的下降而导致性能上有所下降，但我们最好的部署$\phi_\Delta + \psi_d$ + RN仍优于5120维D’=512的费舍尔向量算法。

5.4. 与现有技术状态相比

基准线。我们认为作为基本线最近的工作都是针对同一应运场景和类似表示，也即，通过一个用向量表示的图像可以随后被还原[13]。我们与最近发表的关于中等规模向量的工作进行比较[2,13]。我们还比较我

method ↓	dim. red. to → D'	mAP Holidays	Oxford5k	Ox105k
Fisher baseline	–	63.9	50.7	44.9
Fisher + ψ_d	–	63.8	52.0	45.9
$\phi_\Delta + \psi_s$	–	70.4 ±0.3	58.9 ±0.3	52.3
$\phi_\Delta + \psi_d$	–	72.2 ±0.2	61.2 ±0.4	55.9
$\phi_\Delta + \psi_s$ + RN	–	74.5 ±0.4	63.3 ±0.9	55.5
$\phi_\Delta + \psi_d$ + RN	–	**77.1** ±0.7	**67.6** ±0.2	**61.1**
$\phi_\Delta + \psi_d$ + RN	→ 1,024	72.0 ±0.2	56.2 ±0.1	50.2
$\phi_\Delta + \psi_d$ + RN	→ 512	70.0 ±0.6	52.8 ±0.4	46.1
$\phi_\Delta + \psi_d$ + RN	→ 256	65.7 ±0.3	47.2 ±0.2	40.8
$\phi_\Delta + \psi_d$ + RN	→ 128	61.5 ±0.7	40.0 ±0.1	33.9

表1.我们的方法对性能的影响。首先，我们评估了将费舍尔和民主的聚集结合起来的影响。然后我们考虑T-嵌入ϕ_Δ与和函数（ψ_s）还有民主聚合函数（ψ_d），并展示RN给予的助推。最后我们将为做减少表示量后显示结果。

们重新实现的（改进）版本VLAD和Fisher向量集成RootSIFT。这个基准其本身接近或者通过组合最有效成分而优于现有技术状态。

结果。表2显示出我们的方法优于通过大幅度比较所有数据集中的数据。在最近的一篇文章[2]使用较大的在mAP、Holiday和Oxford5k词汇量增益为+11.8%。与我们使用相同词汇量规模的改进的费舍尔基准进行比较，Holidays在mAP上有+13.2%的增益，Oxford5k有+16.9%，而Oxford105k则有+16.2%。甚至当把维数D’下降至1024维时，我们的方法仍然大幅度地优于用少量向量代表的其他方法。只有当维数下降至D’=128维时其性能略低于在Arandjelovi´c and Zisserman[2]发表

中报告的平均水平。

6.结论

本文的主要目的是减少局域描述符在结合时产生可以代表一幅图像的向量时的相互干扰。这由两个新的和互补的方法解决。第一个是T-嵌入，其可减少在利用无关描述符表示图像相似性时产生的冲突。第二种方法明确地抑制在聚合描述符时他们之间的干扰。由此产生的图像搜索表示甚至在我们的代表向量减少至1000份时也绝不逊色于国家最先进的编码方法，例如费舍尔内核法。

method ↓	$\|\mathcal{C}\|$	D	mAP Hol.	Ox5k	Ox105k
BOW [13]	20k	20,000	43.7	35.4	–
BOW [13]	200k	200,000	54.0	36.4	–
VLAD [13]	64	4,096	55.6	37.8	–
Fisher [13]	64	4,096	59.5	41.8	–
VLAD-intra [2]	256	32,536	65.3	55.8	–
VLAD-intra [2]	256	$\rightarrow$ 128	62.5	44.8	37.4
Our methods					
$\phi_\Delta + \psi_s$ + RN	16	1,920	69.5	53.1	45.6
$\phi_\Delta + \psi_s$ + RN	64	8,064	74.5	63.3	55.5
$\phi_\Delta + \psi_d$ + RN	16	1,920	72.3	57.1	49.5
$\phi_\Delta + \psi_d$ + RN	64	8,064	**77.1**	**67.6**	**61.1**
$\phi_\Delta + \psi_d$ + RN	16	$\rightarrow$ 128	61.7	43.3	35.3
$\phi_\Delta + \psi_d$ + RN	64	$\rightarrow$ 1,024	72.0	56.0	50.2

表2.基于Holiday、Oxford5k与Oxford105k数据集的中间向量维度与短小表示的比较。最后两行显示减少了我们的矢量从8064到1024或128部分后的性能。

致谢。此次工作是在ProjectFire-ID中完成，由ADR法

国研究代理机构支持，同时也通过ERC授予VisRec编号228180支持。我们衷心地感谢Karen Simonyan提供的特征值和石妙静在附录E中进行的基于UKB和Holidays+Flickr1M的补充工作。我们也感谢Florent Perronnin和Naila Murray进行的初步讨论，并鼓励大家阅读他们关于Max-Pooling的文献[20]。

参考书目

[1] R. Arandjelovi´c and A. Zisserman. Three things everyone should know to improve object retrieval. In CVPR, 2012.

[2] R. Arandjelovi´c and A. Zisserman. All about VLAD. In CVPR, Jun. 2013.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, Jun. 2003.

[4] L. Bo and C. Sminchisescu. Efficient match kernels between sets of features for visual recognition. In NIPS, 2009.

[5] Y. Boureau, F. Bach, Y. Lecun, and J. Ponce. Learning midlevel features for recognition. In CVPR, 2010.

[6] O. Chum and J. Matas. Unsupervised discovery of co-

occurrence in sparse high dimensional data. In CVPR, 2010.
[7] Y. Fu, M. Liu, and T. S. Huang. Conformal embedding analysis with local graph modeling on the unit hypersphere. In CVPR, 2007.
[8] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In ICCV, 2005.
[9] C. Hegde, A. C. Sankaranarayanan, W. Yin, and R. G. Baraniuk. A convex approach for learning near-isometric linear embeddings. JMLR, 2012. in submission.
[10] H. J´egou and O. Chum. Negative evidences and cooccurences in image retrieval: The benefit of PCA and whitening. In ECCV, 2012.
[11] H. J´egou, M. Douze, and C. Schmid. On the burstiness of visual elements. In CVPR, 2009.
[12] H. J´egou, M. Douze, and C. Schmid. Improving bag-offeatures for large scale image search. IJCV, Feb. 2010.
[13] H. J´egou, F. Perronnin, M. Douze, J. S´anchez, P. P´erez, and C. Schmid. Aggregating local descriptors into compact codes. In PAMI, 2012.
[14] H. J´egou, C. Schmid, H. Harzallah, and J. Verbeek.

Accurate image search using the contextual dissimilarity measure. PAMI, Jan. 2010.
[15] P. A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. SIAM Journal on Matrix Analysis and Applications, 30(1):261–275, 2008.
[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In CVPR, Jun. 2006.
[17] W. Liu, J. Wang, S. Kumar, and S.-F. Chang. Hashing with graphs. In ICML, 2011.
[18] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, 2004.
[19] K. Mikolajczyk and C. schmid. Scale and affine invariant interest point detectors. IJCV, 60(1):63–86, 2004.
[20] N. Murray and F. Perronnin. Generalized max pooling. In CVPR, 2014.
[21] D. Nist´er and H. Stew´enius. Scalable recognition with a vocabulary tree. In CVPR, 2006.
[22] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In CVPR, 2007.

[23] F. Perronnin, J. S´anchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In ECCV, 2010.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In CVPR, 2007.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In CVPR, 2008.

[26] N. L. Roux and F. Bach. Local component analysis. In Proc. Intl Conf. on Learning Representations, 2013.

[27] B. Safadi and G. Qu´enot. Descriptor optimization for multimedia indexing and retrieval. In CBMI, 2013.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. PAMI, 2014.

[29] R. Sinkhorn. A relationship between arbitrary positive matrices and double stochastic matrices. Annals of Mathematics and Statistics, 35:876–879, 1964.

[30] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In ICCV,

2003.
[31] G. Tolias, Y. Avrithis, and H. J´egou. To aggregate or not to aggregate: Selective match kernels for image search. In ICCV, 2013.
[32] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. PAMI, Jul. 2010.
[33] J. Wang, J. Yang, F. L. K. Yu, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification In CVPR, 2010.
[34] S. Winder and M. Brown. Learning local image descriptors. In CVPR, 2007.