指导教师：杨涛　　提交时间：**2016/3/16**

CVPR2015 Paper Translation

No: 01

姓名：牛洋

学号：**2013302579**

班号：**10011305**

用于自动图像标注的独立特征估计

Amara Tariq Hassan Foroosh
The Computational Imaging Lab., Computer Science,
University of Central Florida, Orlando, FL, USA

摘要：

自动图像标注在图像的搜索、归档和整理系统中扮演这举足轻重的角色。而在没有一个注释工具之前，这样的系统必须依赖于用户输入或大量文本的网页上的图像，以获得其文本描述。在这种情况下，用户提供的图像信息可能不足活着存在大量混杂的信息，同时网页上的所有文本可能并不是一个图像所伴随的描述文字。因此，开发有效的工具来正确和足够的自动标注图像是极其重要的，在这个过程中，图像的内容和上下文起着重要的作用。一个合适的量化的背景下的图像可能会减少视觉特征和适当的图像文本描述之间的语义差距。在本文中，我们通过张量分解提出了一种无监督的特征独立的图像环境量化。将自动标注图像过程中出现的先验知识和预先估计的环境合并，而对于预测注释的评价提供了证据证明了独立特征环境估计方法的有效性。

1. 介绍：

图像搜索和检索系统很大程度上依赖于提供的图像文本描述，以满足用户的文本查询。这样的系统可以大大受益于以生成准确简洁图像描述为目的的图像标注体系。随着描述图像的低级别的视觉功能（HOG，均值与标准差的颜色通道，边缘过滤器等）不能代表准备和可理解的图像文本内容的连接，自动图像标注成为一个极具挑战性的问题。低级别的视觉特征和图像的文本描述之间的距离被描述为语义鸿沟。图像的上下文有助于减小这个间隙。上下文信息可以用作自动图像标注过程中的先验知识，以弥补图像低表示的图像内容与文本描述之间的差距。

图像上下文估计是标注图像信息任务中必不可少的。元数据或任何额外的数据提供的图像可能提供可理解的上下文，但这样的附加信息的可用性并不实际。在理想情况下，上下文应该从图像本身来估计。上下文表示可以适用于某些形式的图像表示。例如，图像的视觉特征可能在识别过程中被用作先验知识场景的细节，即图像的内容。而且，上下文估计任务是与视觉特征和语义鸿沟的内在问题绑定在一起的。

在本文中，我们提出了一个功能独立和无监督的情况下的估计过程。建议的过程不依赖于图像额外的信息或任何形式的视觉特征。它涉及到张量 Tucker 分解，通常用于视频处理。我

们提出了一种能把图像转化成合适张量的独特方法，这种方法能够提供用于单个图像的有用的上下文信息。我们纳入在自动过程中估计的上下文图像标注作为先验知识。在两个非常流行的数据集上（iapr1 和 ESP 的游戏）提供了令人鼓舞的证据证明了我，我们上下文估计策略的有效性。

本文的其余部分安排如下。我们在第 2 节展示了一个关于图像注释和张量分解相关的调查，问题的构想在第 3 节，上下文估计策略将在第 4 节糖提出，在 5 节中，我们描述了上下文信息的注释方案。在第 6 节解释了我们上下文估计策略背后的直觉，而第 7、8 节分别展示了该策略的评价和结论。

2．相关工作：

人们在先前将自动图像标注这个问题已经研究得很好了，而且此前已经已经有人提出了解决办法。有一种流行的框架概念来源于从自然语言到机器语言的翻译处理的相关模型。通过假设一个用训练数据解决期望的生成模型，这些方法可以用一些视觉表示的图像估计句子的联合概率。这些框架计算高效，程度准确，目前我们已经提出好几种方法，这些方法的性能比相关性模型为基础的框架更好，只是在成本上增加了计算复杂度。这些方法依赖系统的一些迭代优化，设想图像编码信息的最邻近的点适当标记图像[9，17，3，29]。计算机视觉的目标识别工具已经在注释程序[21，15]。这种方法通常工作在非常有限有限的词汇集上。

现在已经提出一些系统用辅助信息来进行图像标注，通常这些系统将新闻数据集附带在图像的附带信息中[7，8]．辅助信息提供每一个图像上下文，有助于减少视觉特征与文本描述的语义差距。塔里克等人建议在没有辅助信息的情况下使用图像场景分析估计上下文[27]。

在上下文估计过程中辅助信息与视觉特征之间存在依赖，而我们消除了这种依赖。

张量已被用来作为一种天然的表示视频、文本文档集合和图像 [4，14，28，1，10]。张量分析与分解算法已被应用到任务中，如在域中的动作识别和运动检测视频分析[23，19，30，26]。科尔达等人提出张量分解方法，详细研究并尝试应用[13]。

在本文中，我们提出了一种新策略，用独立的图像和张量分解作为上下文信息的源头来形成有用的张量，这种新策略可以被用到图像标注过程中。

3．配方和符号问题：

我们的目标是估计有用的上下文信息作为图像标注过程中图像的文本标记。假设训练数据来自合理注释和预测系统。训练数据集是由图像和它们的文本描述组成。让每一个训练样本用 I 标记，而

且符号是固定不变的，比如 V，表示所有 I 的集合。N 表示 V 的大小，这些有用的测试图像均可标记为 Io，需要附加说明的图像设为 V，标记系统的目的就是产生一个词汇子集{w1,w2, ...,wB}，每一个 wB 表示一个需要标记的图像 Io。

我们认为，上下文信息被编码在图像训练结构组内，这些图像组和组内其他图像存在着一定的“联系”。一组成员共享的这个“关系”实在 4.1 节中定义的。每一个测试图像编码的上下文信息都有一些关联。

4. 上下文信息估计：

上下文信息的估计在所提出的系统中是一个三步走的过程。第一项任务是形成一组有限数量的训练数据，每一组数据之间都有一些“关系”。下一步是涉及张量的形成和分解，由每组图像组成。在最后一步中，测试图像的上下文是通过修改张量和对比新的分量和上一步中生成结果而估计得到的。

假设每个训练图像 I∈Ⅱ的文字描述都被编码为长度 N 的一个向量 V。每个向量 V 的记录 Vn 表明图像的文本描述词汇的对应词的存在或不存在。

4.1 语境组

在第一步，我们需要构建图像的组，称为上下文组，这样每个组所有图像彼此有某种“关系”。我们的系统这个过程中有两个要求：1）每个图像组应该是某些上下文的表示，并对图像有适当注释预测，2）一组内的图像彼此应具有足够的视觉相似性，以便该组可作为上下文视觉特征形成的基础。

训练集的图像的文本描述是可用的。假设图像的文本描述预测其视觉表现是很直观的。我们计算的 TFIDF 表示每个图像 I 的文本描述，用 V′ 表示一个长度为 n 的向量。如果 NnI 是第 n 个字第 n 次出现在图像 I 的文本描述中，Nn 是第 n 个词汇在图像数据集中的累积频率的描述，第 n 个记录 V 是：

$$v_n^I = \frac{N_{nI}}{N_n} \qquad (1)$$

图一：由文本描述的相似性为基础形成的一个实例

图二：由文本描述的相似性为基础形成的一个实例

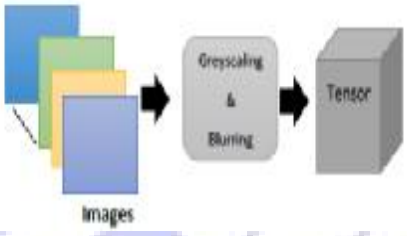
训练集的图像以他们的 TFIDF 向量之间的余弦相似度为基础被聚集起来。TFIDF 表示特性确保这一过程中文本描述中有相同独特单词的图像放在

同一组中。每个图像组将能够唯一为这些单词提供证据。例如，如果这个词“天空”通常发生在所有训练的描述中数据，它是“通用”的数据集的词汇的一部分而不属于一个独特特征的图像组。

另一方面，如果“雪”是一些图像的标签，该组这些图像可以唯一提供标记“雪”。此外，在一起分组的图像文本表示之间有很高的相似性。因此，我们有理由认为一组图像之间有合理的视觉相似性。我们采用了一个迭代的分层聚类过程和截止阈值控制每次迭代的组数。大群组在随后的迭代中进一步分离，以保持上下文组尺寸分布尽可能均匀。

4.2. 张量的生成和分解

一个语境张量 Tc ∈ RX×Y × Z 被构造出来用于第一步中形成的图像组中的每一个图像。一组中图像的被调整到一个固定的高度 y 和宽度 x，转换为灰度，通过高斯模糊滤波器加工，再连在一起形成张量。张量的三维，即 X，Y，Z 分别代表图像的高度和宽度，图像索引。



图三：张量生成：一组图像叠放在一起形成一张量

注意，这个过程的目标是要估计上下文的整体标记，同时上下文由同一组中的图像描述的独特词语编码。这种上下文标记应该是不敏感的视觉细节，以便当一个新的图像关联到任何上下文标记时进行评估，它专注于新图像和成员图像之间整体的相似性而不是对图像的局部细节。因此，一组图像的模糊高斯处理过滤器，以消除边缘生出的尖锐效果。

下一步是上下文张量的分解，通过 Tucker 分解找到一个紧凑的语境组的标记。Tucker 分解是一个受欢迎的技术来突出张量 Tc ∈ RX×Y ×Z 到较小核心张量 S 和三个矩阵 P,Q,和 R，例如

$$\mathcal{T}_c \approx \mathcal{S} \times_1 P \times_2 Q \times_3 R = \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{z=1}^{Z} g_{xyz} p_x \circ q_y \circ r_z, \quad (2)$$

其中 P ∈ RX×K, Q ∈ RY ×K, and R ∈ RZ×K 是正交矩阵，S ∈ RK×K×K 是核心张量，K ≤ min(X, Y,Z)。×i 表示 i 型张量之间的乘法，其结果也是一个张量 A = B × i α ∈ ⇒ $(\mathcal{A})_{jk} = \sum_{i=1}^{I} \mathcal{B}_{ijk}\alpha_i.$

我们采用秩 1 分解，即 K 值设置为 1。在这种情况下，P，Q，R 向量长度分别等于图像的宽度，图像的高度和语境组大小。向量 R ∈ RZ×1 我们系统中最重要的，这个向量表达了在张量 Tc 中相邻图像之间的相似度/不相似度，因为在一个上下文组中的图像都是连接在一起的，即，它们都是视觉上相似的，因为他们都有非常相似的文字说明，但是在向量条目中仍

然有小的不同。向量 R 是上下文组之间的标记。

4.3 上下文预估

下一步是量化测试图像的背景在不同语境下的关联性。每一个测试图像表示为 Io。没有可用的文字来吗描述 Io。我们在先前的一节解释过了，上下文标记是一个长度为 R 的向量带着它条目上的小差异，这是由同一组内视觉上相似的图像 tucker 分解的结果张量 R 组成。如果是不相干的实体，例如一个测试图像 Io，插入到这个张量在任何位置，此时 I，它会干扰向量内 Io 和周围的索引 I。干扰的程度将与 Io 和该组的成员之间的差异性成正比。

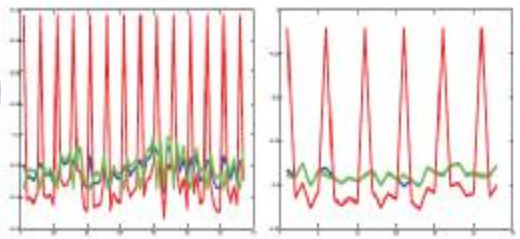


图 4：将视觉类似的和不同的图像插入到一个张量进行秩 1Tucker 分解并进行比较，蓝色曲线：原 Tucker 分解向量 R，绿色曲线：新 Tucker 分解向量 R′，并将同一上下文组视觉上相似的图像在分解向量的情况下，插入上下文张量，红色曲线：新的视觉上不相似的图像在分解向量的情况下插入张量。

用上下文估计测试图像的关联性，它被插入到一个固定的时间间隔的位置，比如 L，通过为 Io 在这些位置交换图像，而存在于相应的上下文张量中。新的矢量 R 通过 Tucker 分解计算得到。R 和 R′ 的差异是一个关于 Io 的关联性的相反的程度，并由 Tc 相应的上下文组表现出来。我们通过每一个可能的上下文组估计出 Io 有条件概率分布，比如：

$$P(T_c|I_o) = \frac{\exp(-(R'-R)^T\Gamma^{-1}(R'-R))}{\sqrt{2\pi|\Gamma|}} \quad (3)$$

Γ 是协方差矩阵，被假定去形成一个 γ I 当 I 是一个恒等矩阵的时候，γ 可以通过部分数据的保留部分被选择出来。此概率分布将测试图像的关联与可用的上下文组编码，轮流通过每个组的独特的词语设置来编码其关联性。

如前所述，词的频繁发生是在形成上下文组的过程中占较轻的比重。为了为这样的词汇服务，我们可以形成一个“全体”的语境组，由所有训练图像组成。分配给每一个测试图像 Io 相同的条件概率和“全体”的上下文组，P(Tc|Io)被重整以便于它的总和为 1。让 α 表示重整化比重，由部分训练数据通过交叉验证来估计得到。

注意在三个步骤的过程没有视觉特征。相反，在一个全面估计背景下通过 Tucker 分解图像过程得到的概率分布得到原始训练数据的文本标签。

5. 上下文敏感图像自动标注

在这一节中，我们将提出我们的自动图像注释策略，结合上下文信息，在第 4 节估计出作为分布 P(TC | IO)。

我们的注释模型来自于相关性模型的启发，用于机器翻译中的联合概率来估计语言之间的不同。该模型在加权期望在得到联合概率和图像的视觉特征过程中诱导出依赖性上下文。

5.1 数学模型

每个图像被假定为一个视觉单元的数，即 R = { R1，R2，…，Ra }。这些视觉单位是通过一个网格划分每个图像固定尺寸。每一节的颜色和质地品质网格形式代表该节的向量。以集合为代表的图像的文本描述设为W = { W1，W2，..，Wb }，每一个 Wb∈V，此时 V 是词汇集。集合的大小设为 w，B 被假定为所有相同测试图像。

上下文信息引入假设存在一套所有上下文集合，即 T 使得每个 Tc∈T 对应一个语境组，反过来，一个语境张量（4.1 和 4.2 的定义）。每个训练图像 I 属于其中一个 Tc。每个测试图像 Io 有一定的条件概率分布于所有的 Tc，用 P(Tc | Io)标注。上下文也用 P(Tc | Io)编码。塔里克等人提出了一个类似于上下文估计场景分析的案例[27]。因此,我们采用类相关性通过上下文加权的模型,即 P(Tc | Io),用来估计单词的联合概率的单词和 Io 视觉单元。

1. 通过测试图像 Io 的可能性概率选择上下文类别 Tc ∈ T，即 P（TC | IO）；
2. 通过概率从训练集中选择图像 I；
3. for a = 1, 2, ...,A
   （a） 从可能的 PR(.|I)中选择一个视觉单元 Γa
4. for b = 1, 2, ...,B
   （a） 从可能的 PVTc (.|I)选择一个单词 ωb

该系统的目标是最大限度地提高下面的公式给出 Io 的 r 和 w 的联合概率。

$$P(\mathbf{w},\mathbf{r}|I_o)=\sum_{T_c\in\mathbb{T}}P(T_c|I_o)\sum_{I\in\mathbb{I}}P(I|T_c)\prod_{b\in B}P_{V_{T_c}}(w_b|I)\prod_{a\in A}P_{\mathbb{R}}(r_a|I) \tag{4}$$

假设训练图像 I 的图像描述在文本组的词汇集上承多伯努利分布，即 VTc。因此，PVTc (wb|I) 这是该分布的第 wb-th 个成分。

$$P_{V_{T_c}}(w_b|I)=\frac{\mu\delta_{w_b}+N_{w_bc}}{\mu+N_{T_c}} \tag{5}$$

Nwbc 表示有单词 ωb 的描述的上下文组 Tc 的成员数目，NTc 是 Tc 的成员总数，如果图像的描述中出现 ωb，则将 σwb 设为 1，不然设为 0。μ 是一个经验选定常数。

4.3 节阐述了在 P(I|Tc)被用来估计为下一步的功能时，通过方程 3 估计 P(Tc|Io)

$$P(I|T_c)=\begin{cases}1/N_{T_c}, & \text{if } I\in T_c\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

PR(ra|I)是给定一个训练图像 I 时产生的视觉单元 ra 的密度估计，此

时 ra 是一个视觉单元，即它属于测试图像 Io。这个密度估计使用高斯核。假定如果训练图像 I 是由视觉单位集{ I1，I2，…，IA }，那么

$$P_{\mathbb{R}}(r_a|I) = \frac{\exp(-(r_a - i_a)^T \Sigma^{-1}(r_a - i_a))}{\sqrt{2\pi|\Sigma|}} \quad (7)$$

这个方程使用带协方差矩阵Σ的高斯密度核，这个矩阵还可以为了方便被当做βI，此时 I 是恒等矩阵。β使点 ia 周围保持平滑，还可以凭经验选择数据集。注意这一估计意味着在 I 和 Io 之间空间一致的重要性，因为它在同一个网格位置比较视觉单元，用下标α表示。

该模型结合了为每一个测试图像 Io 估计的 P(Tc|Io)形式的上下文信息，而且我们的实验显示此信息提高注释系统的性能。

6．张量分解的含义

在这一节中，我们提出了一个简短的分析张量的形成和分解过程的复杂性及其对图像的上下文估计的影响。

张量的形成和分解的思想在文本挖掘与视频分析领域得到广泛的探索。张量为视频提供了一个全面的表示，视频的每一帧都是张量的一个'切片'。三维中的两维分别代表的帧的宽度和高度，而第三个维度代表时间。因此，张量非常适合视频的时间分析。我们在这项工作中的任务是要来的一个图像的全面张量形成策略而彼此没有时间联系。在我们的这个案例中，第三个维度只是表示图像指数。

三阶张量的 Tucker 分解是矩阵的主成分分析的高次扩展【13】。这是一个基于秩的估计，可得到三个矩阵的张量和一个预先指定大小的核心张量的分解。张量由文件作者和关键词信息组成，假定三个维度代表张量的单词，作者和关键词的张量。三个分解的矩阵 U,V 和 M 分别代表词和词组，作者与作者群体，关键词与关键词组。词组，作者群体和关键字组的数量由核心张量的大小指定。核心张量编码表示组之间联系的情况。

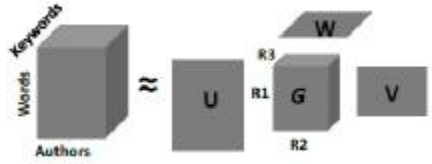


图 5：Tucker 分解：U=词×词组，v=作者×作者群，w=关键词×关键字组，R1，R2 和 R3 代表词，作者和关键字组

推荐的上下文估计策略采用秩-1 分解，即核心张量是一个标量和矩阵的是向量。这样系统根据其文本描述的相似性，已经知道属于一个张量的图像同一个组内。

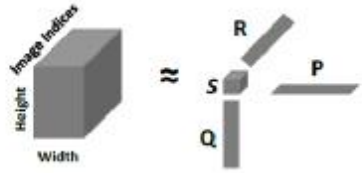


图 6.秩 1 Tucker 分解：S 是一个标量，P，Q 和 R 是向量，R=图像-索引×1,其中 1 表示了张量表示的单个上下文组。

这种类型的组信息在我们提到的系统的最终任务中是潜在有用的，例如图像注释。Tucher 分解的目的就是找出一个组中的单个元素是如何联系到该组的所有成员以便于系统判断测试项是否属于该组。理想的情况下，当所有图像相互之间都相似的话，图像的索引元素和向量会有小的变化。如果一个不相干实体插入，这个向量会扰动。扰动的量体现了对于本组的相异程度。如果不相干实体是测试图像，如在 4.3 节所见的一样，这个过程可估计出测试图像与本组图像的相似度。

6.1. 计算复杂性

所提出的背景估计的计算复杂度取决于用于 Tucker 分解策略。Tucker 分解流行的现有算法比如高阶正交迭代（HOOL）是基于交替最小二乘法（ALS）的。Phan 等人提出一种比 HOOL 简单的算法。ALS 方法不能保证收敛到全局最优解或一个固定的点。如果它在一定条件下收敛，那么它在有局部线性收敛速度。另外，牛顿几何微分方法给出了收敛保证与二次收敛速度，以及每次 OH3D3 给一个向量 T ∈ RH×H×H 和核心向量 S ∈ RD×D×D 的迭代的成本。

7.评价

在这一节中，我们将通过所提出的独立特征策略估计上下文合并的影响，这是基于自动图像标注的相关模型的。

7.1. 数据集

我们使用了两种流行的图像注释数据集，IAPR-TC 12 和 ESP game，用来评测我们的系统。IAPR 数据集拥有 19846 张游客所拍的图像，每一张图像都被语句注释过。

TreeTagger3 给令牌的标记化处理、词干提取、标注的部分用来描述图像。经常出现的名词被选到词汇表中。ESP 游戏数据集包括玩家在游戏中标记的图像。一个大小为 21844 的子集已经被普遍用于测试不同的图像注释系统。我们也尝试用相同的子集。这个对每个图像的描述已经以令牌/词汇形式形成词汇集。

在训练和测试集上，我们使用了相同的分割数据（90%用于训练，10%用于测试）。这两个数据集也被用于测试其他图像注释系统。IAPR 和 ESP 数据集不同的图像标注系统已分别普遍测试过的词汇集的 291 和 269 个最常见的单词。在我们的系统中，训练数据被分为许多背景组，每个组提供对于该组图像描述的情况。因此，词汇集不断从一个组变化到另一个组的样本，而不是被固定为所有数据的特定号码。但我们确定，大约相同数量的独特词汇（IAPR291，ESP 269）出现在最终的输出，即通过调整我们系统的参数来预测注释测试图像。在本文中，以上这些独特词汇的结果已被记录下来，以保持它们相媲美其他系统。

IAPR 和 ESP 游戏数据集的整体的词汇集的长度分别为 1002 和 2032。

我们的系统，在第一阶段，基于图像文本描述相似性来形成训练图像组，例如，IAPR 有 595 个上下文组，ESP 有 637 个上下文组。太小实体将被丢弃当它们对应整体词汇集不经常发生。训练图像修饰它的上下文组，如果到上下文标记 R 的平均距离发生了一定的偏差。这种图像缺少与其上下文组产生相似的条件。足够接近至少一个上下文组的测试图像是测试集的一部分。

7.2. 图像表示

我们在自动图像标注过程采用基于网格的图像表示。这中表示方法需要通过一个固定尺寸的网格划分每个图像。在我们的实验中使用了一个 5×6 的网格。每一个网格部分指定一个向量表示图像特定部分的颜色和纹理质量。在我们的实验中，这个向量的长度为 46，包含 18 个颜色特征（RGB，LUV，LAB 色彩空间每个通道的均值和标准差），12 个纹理特征（3 个尺度和 4 个方向的 Gabor 能量计算，4 bin HoG 和离散余弦变换系数。这组图像特征已经被许多以前的图像注释系统使用[16，6]。我们观察到增加超过 5 到 6 的网格尺寸并没有提高系统性能。

Guillaumin 等人采用了整体和局部视觉特征的组合来改进他们的系统性能[9]。最近，chen 和 Verma 等人在他们的系统中使用了相同的功能。我们还进行了额外的实验与此相结合的功能集阿里观察到性能改进的情况。

请注意，这些功能在基于图像标注过程的相关模型中使用，与上下文估计程序无关。我们系统的初始阶段是独立的，只是在处理原始测试图像估计上下文信息。

7.3. 结果

在一般情况下，图像标注系统使用来产生每个图像的注释，且数量与测试数据的每个图像的平均值相同。一般使用的评价参数的平均值的精度并产生每一个单词和单词数（N+）。我们使用相同的评价标准。

	%age mean Precision	%age mean Recall	N^+
CRM[16]	21	15	214
MBRM[6]	21	14	186
BS-CRM[22]	22	24	250
JEC[20]	25	16	196
Lasso[20]	26	16	199
HGDM [18]	29	18	–
AP[25]	28	26	–
TagProp-ML[9]	48	25	227
TagProp[9]	46	35	266
FastTag[3]	47	26	280
2PKNN-ML[29]	54	37	278
context-RM	56	24	224
context-RM-B	61	24	242

Table 1. Performance evaluation for IAPR-TC-12 dataset

	%age mean Precision	%age mean Recall	N^+
CRM[16]	29	19	227
MBRM[6]	21	17	218
JEC[20]	23	19	227
Lasso[20]	22	18	225
AP[25]	24	24	–
TagProp-ML[9]	49	20	213
TagProp[9]	39	27	239
FastTag[3]	46	22	247
2PKNN-ML[29]	53	27	252
context-RM	55	21	226
context-RM-B	61	20	234

Table 2. Performance evaluation for ESP game dataset

表1和2显示我们的系统分别在IAPR-TC 12和ESP数据集与先前提出的

策略的性能比较。我们系统的两个变化已经被彻底测试过。双标记，即context-RM和context-RM-b，通过Guillaumin等人，我们的系统分别实现了基于网格的视觉特征和特征表现，基于注释来产生加权期望。表3显示了极高和极低出现的数据集单词样本。

7.4. 观察

正如2节所解释的，以前对于图像注释的不同探索使用了不同的策略，而每一种策略都有其优点和缺点。在表1和表2中，CRM和MBRM是两种关联模型注释技术，这是非常有效的算法并且运行适度。我们的注释策略也是基于相关性模型，但我们通过新的特征独立策略进行上下文估计。我们的注释预测框架运行起来比其他基于相关模型的系统更好。TagProp,FastTag and 2PKNN-ML指几个迭代优化或最近邻类型的框架，计算相当复杂，但是预测效果更加准确。我们的策略在预测的精度方面表现得比这类系统更好。我们系统的预测性能可与FastTag和TagProp-ML相比。大部分计算在于我们系统的前期处理阶段，涉及上下文估计。在其余部分我们系统的计算效率都比较高。我们的策略打败了基于贪心算法的系统如JEC和Lasso。

8. 结论

我们提出了一种采用张量分解的新的特征估计策略。张量先前已被建议作为一种天然的视频处理表示方案。我们的贡献是从单个图像形成张量来作为独特的解决方案，每个张量编码了有用的信息作为图像的上下文。所建议的上下文估计策略是特征独立。我们采用在自动标注图像标注过程中估计的上下文，一个通常存在于视觉特征和语义描述之间鸿沟的问题。我们的注释策略的性能提供了上下文估计过程的有效的数据。在将来，我们将继续探索秩高于1的张量分解，对于上下文组的形成以及上下文估计。

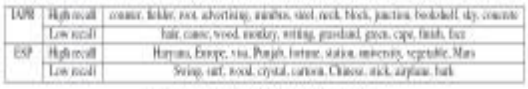


Table 3. Sample of words with low and high recall values

References

[1] B.W. Bader, M.W. Berry, and M. Browne. Discussion tracking in enron email using parafac. In Survey of Text Mining II, pages 147–163. Springer, 2008.

[2] B. CHEN, Z. LI, and S. ZHANG. On tensor tucker decomposition:The case for an adjustable core size.

[3] M. Chen, A. Zheng, and K. Q. Weinberger. Fast image tagging,2013.

[4] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. SIAM journal on Matrix Analysis and Applications, 21(4):1253–1278, 2000.

[5] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1

and rank-(r 1, r 2,..., rn) approximation of higherorder tensors. SIAM Journal on Matrix Analysis and Applications,21(4):1324－1342, 2000.

[6] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.

[7] Y. Feng and M. Lapata. Automatic image annotation using auxiliary text information. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008.

[8] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In IEEE 12th International Conference on Computer Vision, 2009.

[10] Z. Hao, L. He, B. Chen, and X. Yang. A linear support higher-order tensor machine for classification. IEEE Transactions on Image Processing, 22(7):2911－2920, 2013.

[11] M. Ishteva, L. De Lathauwer, P.-A. Absil, and S. Van Huffel. Differential-geometric newton method for the best rank-(r 1, r 2, r 3) approximation of tensors. Numerical Algorithms, 51(2):179－194, 2009.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003.

[13] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. SIAM review, 51(3):455－500, 2009.

[14] G. Kuhne, J.Weickert, O. Schuster, and S. Richter. A tensordriven active contour model for moving object segmentation. In Proceedings of IEEE International Conference on Image Processing, 2001.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating

simple image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[16] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In Advances in neural inforinformation processing systems, 2003.

[17] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. IEEE Transactions on Multimedia, 11(7), 2009.

[18] Z. Li, Z. Shi, W. Zhao, Z. Li, and Z. Tang. Learning semantic concepts from image database with hybrid generative/ discriminative approach. Engineering Applications of Artificial Intelligence, 26(9), 2013.

[19] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A survey of multilinear subspace learning for tensor data. Pattern Recognition, 44(7):1540–1551, 2011.

[20] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In Computer Vision–ECCV 2008. 2008.

[21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daum III. Midge: Generating image descriptions from computer vision detections. In Proceedings of Annual Meeting of European Association of Computational Linguistics, 2012.

[22] S. Moran and V. Lavrenko. Optimal tag sets for automatic image annotation. In Proceedings of the British Machine Vision Conference, 2011.

[23] K. Palaniappan, I. Ersoy, G. Seetharaman, S. R. Davis, P. Kumar, R. M. Rao, and R. Linderman. Parallel flux tensor analysis for efficient moving object detection. In Proceedings of the 14th IEEE International Conference on Information Fusion (FUSION), pages 1–8, 2011.

[24] A.-H. Phan, A. Cichocki, and P. Tichavsky. On fast algorithms for orthogonal tucker decomposition. In IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[25] M. Rubinstein, C. Liu, andW. T. Freeman. Annotation propagation in large image databases via dense image correspondence. In Computer Vision–ECCV 2012.

[26] C. Sun, M. Tappen, and H. Foroosh. Feature-independent action spotting without human localization, segmentation, or

frame-wise tracking. In IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[27] A. Tariq and H. Foroosh. Scene-based automatic image annotation. In IEEE International Conference on Image Processing, 2014.
[28] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In Computer Vision– ECCV 2002.
[29] Y. Verma and C. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In Computer Vision–ECCV 2012.
[30] Q. Zhao, G. Zhou, L. Zhang, and A. Cichocki. Tensor-variate gaussian processes regression and its application to video surveillance. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.