指导教师：杨涛　　提交时间：2016-3-14

# CVPR2015 Paper Translation

No: 01
姓名：张锦阳
学号：2013302488
班号：10011301

# 结合图像流和博客文章的探索与总结

Gunhee Kim
Seoul National University
gunhee@snu.ac.kr

Seungwhan Moon
Carnegie Mellon University
seungwhm@cs.cmu.edu

Leonid Sigal
Disney Research Pittsburgh
lsigal@disneyresearch.com

## 摘要

我们提出利用大量的共享照片流和博客，这两个在网络上最流行的数据来源，做联合的故事为基础的总结和探索的方法。博客由一系列图像和相关的文本构成;他们用简洁的句子和代表性的图像来描绘事件和经历。我们充分利用博客，以帮助实现照片流的收藏故事为基础的语义总结。反过来，博客可以通过在它里边展示连续图像的差值被增强。我们用统一的潜在排名的 SVM 架构制定了关于校准博客到图片流和图片流摘要的一个总结联合对齐的问题。我们解决了两个耦合潜伏 SVM 的问题，通过首先确定摘要和解决博客图片到照片流的对齐问题，反之亦然。对 10K 博客（120K 相关图像）和 6K 的照片流（540K 图片），通过新收集的大型迪斯尼乐园的数据集，我们证明了博客文章和照片流对总结，探索，语义知识转移和照片内插是互相受益的。

## 1. 简介

一般用户拍摄的照片可以被视为他们要记住的故事和要告诉我们他们的经验的个人陈述。图 1 展示的是最明显的例子之一，参观迪士尼乐园。在某一天，成千上万的人参观迪斯尼乐园，其中许多人照了巨量关于他们与家人或朋友特殊的经历的照片大。此外，一些比较热心的游客还愿意写游记博客，在他们的个人故事与行程，通过评论，印象，以及有关旅游景点有趣的事实展开。大多数博客包括有用信息的文本信息，和他们从他们旅行中大量图片中挑选的几张最具代表性的图片。

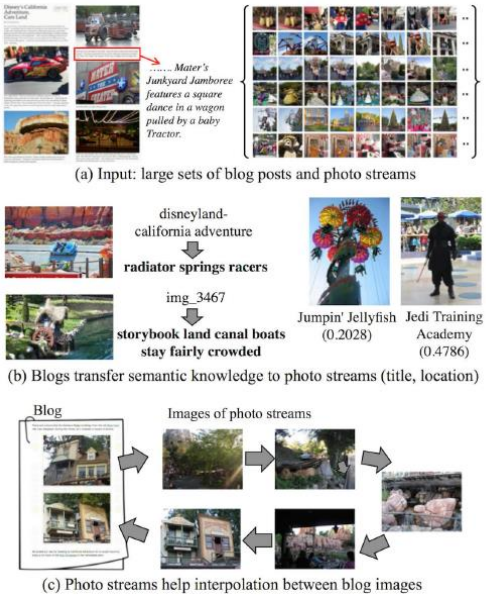


图 1.大集合图片流和博客之间为了联合概要和探索的动机因素。(a) 输入有两方面：一组来自迪斯尼乐园的照片流和博客文章，这是由多个用户在不同的时间捕获的。(b) 博客通过转换语义上的知识，有益于图像流的摘要：举个例子，图片自动命名和基于景

点的图像定位。(c) 图像流通过允许博客图像之间插补来增进博客内容。两个博客图片用一个景点入口来作为查询，结果可以证明景点内发生了什么事情。

在本文中，如图 1 所示，为了这个总结和探索，我们用一种互惠的方法提出利用巨量照片流和博客的优势。博客通常由一组图像和相关的文本组成；它们通常以讲故事的形式写出来，通过用讲练的句子和代表性图片摘录主要时间。因此，博客可以帮助实现一个关于故事为基础的语义总结，我们收集的大规模的和日益增长的图片流往往是没有结构，没有关联或者是语义上不明确的。从相反的角度，在博客的连续图像之间可以插入各种图像的路径，每一个博客都能从这一大组图像中受益。每个博客是一个人的经历和根据少量选择的图像写的。因此，图片路径的插入，通过图片流来实现，允许博客作者提供可供替代的通过其他参观者通过类似方法的到的图片路径。

为了实现共同总结和探索，我们首先要从 Web 收集大量相关的名胜（例如访问迪斯尼乐园）的事件的照片集流和博客文章。然后，我们共同执行两基任务，让他们互相受益：(1) 博客上的图片和照片流之间的校对 (2) 博客流的总结。这个校对任务可以发现一个博客照片在照片流中的图像对应。这个总结任务从图片流中采用最大的覆盖率和最小的冗余率选择了最有意思最重要的图像。由于博客的照片是由用户的精心选择的，令人振奋的图片流拥有的图片摘要可能通过的更多的语义含义匹配博客图片。相反的，图片流的摘要可以让对齐测试任务更快，更集中。我们制定的对准和总结为两套潜在排名的 SVM 问题的最优化。因此，从相片流的初始聚合开始，我们解决了当两个测试任务之一交替时，决定输出是另一个。

为了进行评估，我们抓取的数据集迪斯尼乐园，其中包括来自 Flickr 的 6K 照片流约 540K 的图片和与 BLOGGER 和 WordPress 相关的有 120K 图像的 10K 的博客文章。虽然我们主要讨论在迪斯尼乐园的背景下提出的方法，我们的方法可以扩展到任何问题领域，只需要很少的改动，因为我们的 NLP 前处理（例如关键词提取）是无监督。唯一的要求是该域必须有的照片流和博客苏夫网络 cient 数量（例如城市或博物馆之旅）。

在实验中，我们集中展示了，博客文章和照片流确实是相互受益的。首先，我们证明了博客文章有助于实现大型成套照片流的以故事为基础的语义总结。此外，我们为两个任务的语义知识转移的算法，比较我们的算法与其他候选方法。因为博客是由一组图像和关联的文字组成的，一旦对准完成后，我们可以移交与博客图片相关的对齐的图片流的语义知识，但是他们中大多数有噪点或没有语义上的意义。具体来说，我们展示了博客文章提高了图像定位精度（即找到哪里拍摄了这些图片），以及自动图像命名（即用于图像创建描述性标题）。第二，我们表明，一个大集合的照片流致使连续的博客图像之间有了更好的轨迹插补。我们通过使用亚马逊的 Mechanical Turk，定量评估众包我们的轨迹插补方法的性能。

**相对于以前的工作。**在最近的计算机视觉研究，一些研究已进行组织和探讨非结构化的游客的照片集。旅游相集能够使使用者交互式地在一个三维空间浏览大量基于地理位置标注的旅游景点。旅客们的图片对在 Google map 上创建一个虚拟的全世界的标志性建筑起到至关重要的作用。这个 3D 形式的维基百科在各个地区用 3D 标注模型和维基百科的形式建立了语义导航。从工作的另一个方面，处理关于户外活动的网络相集创造的

关于故事式的图表的问题。对比这些工作，主要的差异是，我们明确地利用博客的文字和相关图像，组织一般用户的照片流。这样，我们就可以实现语义含义和故事为基础的总结。虽然 3D 维基百科使用在线的与用户图片流对应的文件，但它集中精力于一些小数目的结构良好的样例。相反的，我们使用公众数据通过大量无结构化的博客文章的创建。

最近，一直专注于利用共同的视觉和文本数据，以解决具有挑战性的计算机视觉问题。一些明显的问题是，它可以在两个互补的信息源之间协同影响，包括从来自自然语言描述图像[12, 17, 30]和视频[2, 6, 19]，联合的探测，场景的分割和从图像中通过文本化的信息[4]得到的对象类型。然而，我们的工作的新颖特征是双重的。首先，我们使用博客作为数据源的文本，二是我们的目标是以故事为基础的照片流和博客的的探索。

最后，在数据挖掘的研究，已经有几个以前的文章，以解决从 Web 日志数据[3, 5, 18]中的故事的提取。然而，大部分都是基于纯粹的文本信息。他们之中，工作[7]可能是最贴近我们的，因为他们也利用博客的照片，发现在几个大城市流行的标志性建筑（如北京，悉尼）。然而，[7]简单地利用博客作为照片储存库，而我们关闭了博客文章和照片流之间联合探索的循环。因此，我们可以执行一些额外的任务，包括来自博客语义知识转移到博客的照片之间的图像和照片轨迹插补。

贡献。我们的贡献有三。

1) 据我们所知，我们的工作在共同利用大型成套博客文章和照片流互相收益的总结和探索是独一无二的。我们发现，博客在语义知识转移以及照片流的故事为基础的总结是非常有用的。同时，一大组图片流帮助在任何连续博客图像之间合理差值。
2) 我们提出的方法用于共同解决统一排名的 SVM 框架下的对齐和总结测试。当另一个的的解决方案调理时，我们交替解决这两个问题中的一个。
3) 为了进行评估，我们收集了迪斯尼乐园的数据集，它由 10K 博客文章与 120K 相关的图像，和 540K 的图像 6K 照片流组成。我们证明了博客文章和照片流确实对总结和探索有所帮助。

## 2. 输入数据和预处理

### 2.1 图片流和博客文章

我们的方法的输入是一组照片流 $\mathcal{P}=\{P^1,\cdots,P^L\}$ 和一组博客 $\mathcal{B}=\{B^1,\cdots,B^N\}$ 为一个景点的主题，如迪斯尼。每张照片流是一组某一个用户在某日按顺序拍摄的图像，通过 $P^l=\{p_1^l,\cdots,p_{L^l}^l\}$ 表示。我们假设每一个图片流pli和时间戳tli是相关的，基于此，每一个图片流在时间上被排序。另外一些图片流可能包含GPS信息gl i和文字信息sli(例如标题和标记)。每一个博客包含一对图片和文字块(定义为 $B^n=\{(I_1^n,T_1^n)\cdots,(I_{N^n}^n,T_{N^n}^n)\}$)。我们用I表示所有博客的图片信息。我们讨论自然语言处理的细节在2.3节的博客文章里。

请注意，使用自然语言处理技术是无监督。一个可选的域的特定输入对名字实体的提取来说是一个词汇列表。由于我们对本土化的景点的信息特别感兴趣，我们添加迪斯尼乐园的景点和地区名称的词汇表，引用旅客的地图。在另一些地域，词汇列表可

以从公开的信息中很容易地构建，而且没有任何算法自己参与的匹配。所以，我们的方法是通用的，并且可以被应用在任何被采集的事件和标题上。

## 2.2 图片描述

我们使用矢量量化密集的特征提取，这是在最近的计算机视觉文献的标准方法之一。我们密集的提取 HSV 颜色，SIFT 和面向边缘（HOG）的柱状图，分别在每一个图像的的 4 和 8 个像素的规则网格上。然后，我们通过采用 K-means 随机选择的描述符来构成每个要素类型 300 的个视觉单词。最后，最近的字被分配到网格的每个节点。正如图像或区域的描述，我们构建 L1 化空间金字塔直方图统计每个视觉词的频率在三级正规网格。我们通过串联的颜色 SIFT 和 HOG 特征的两个空间金字塔直方图定义图像描述符 v。

## 2.3 博客文章的自然语言预处理

在博客中的内容是通常是高度与嵌入图像相关的，因此可以作为一个丰富的语义信息源。然而，以这种方式使用博客文本的数据有几个自身的挑战：（一）很难定位到底是哪部分文字对应于哪个图像，（二）博客文字是复杂的，去理解自然语言描述的算法也是一个挑战（三）非专业写作往往含有大量的词汇和语法错误。此外，我们不能指望用句子来形容这些图片博客所有图像或事情。

我们这里的目标是描述每一个博客 Bn，通过一组图片，相关联的元数据，和相应的置信度：{(In 1 ,mn 1,vn 1)⋯ ,(In Nn,mn Nn,vn Nn)},当 mn i ={ln i ,kn i }，ln i 是一个名字实体的列表，kn i 是一个通过博客 n 中的图片 i 的相关联的内容中提取的关键短语的集合，并且 vn i ∈ [0,1] 是一个置信度向量的分数的长度|ln i | + |kn i |。

**命名实体提取** 为了从博客内容中提取命名实体（即主要景点名称，在我们的设定中），我们使用基于 CFR 的实体命名辨识器的线性链，在在标准的 NER[14,27]训练主体(CoNLL 2003 22)上训练。我们只选出位置相关的实体，并且对管辖范围内的景点，找到最接近的匹配（例如迪斯尼乐园）。单词的置信度作为一个景点的根据被标记，所以关于 NER 标签的后验概率会通过自身的距离等级对应的最接近的匹配被处罚。

**关键短语的提取** 为了关键短语的提取，我们使用一个叫 RAKE(快速自动提取关键字)[20]的无监督的方法，它通过测量词的使用频率和关键字的临近关系[1]计算单词蕴含的相互关系的得分来预估关键短语。

**图片关联的置信度** 我们工作的一个主要挑战是计算我们如何通过博客的图片来关联所提取的文本信息（例如地点和关键词）。假定一个文本块更靠近这个图像，则属于图像的概率较高，我们采用基于图像到文本距离的简单试探法。例如，属于某一确定的图像的置信度得分可以计算为包含位置的名字的文本块的置信度总和，由图像到文本块的距离的处罚：

$$\mathbf{v}_i^n = \sum_{t \in T} \left\{ h_t(\mathbf{I}_i^n) - pen(t, I_i^n) \right\} \qquad (1)$$

当 vin 参考被一个位置 lin 关联的关于一个图像 lin 的置信度分数，T 是一个属于博客的文本块的集合，ht(ln i )是文本块的一个置信度分数，t ∈ T，包含一个地理位置 ln i ，并且 pen(.)是一个惩罚函数，它根据文本块和图像的距离降低这种组合关系。我们使用 $pen(t, I_i^n) = d(t, I_i^n)/|T|$ 是文本块 t 和图像 lin 在每一个文本块和图像被当做队列中的一个元素时的索引距离。

## 3. 方法

为了探索博客和图片流之间的联系，我们要解决两个子问题：（i）从博客图像到照片流的对齐，（ii）照片流的汇总。对准通过建立一个双向影像图实现 G = (V,E)，其中顶点集由博客和照片流的图像构成(i.e. V =I ∪P)，并且边的集合 E 是他们之间当前的对应关系。我们用 W ∈ R|I|×|P|表示邻接矩阵，在|P|是所有照片流中照片的数目时。因此，对准的目标降低到对 W 的估计。另一方面，总结的目标是对每一个照片流 PI ∈ P 预测出最好的集合 SI ⊂ PI，我们使用 S 表示一个摘要的集合 S = {S1,⋯ ,SL}。现在列出对于对齐的显示 A1-A4 和概要 S1=S2。

(A1) （稀疏性）我们让 W 具有几个非零元素。我们只保留少数强匹配，以避免不必要的复杂的排列在博客图像链接图片流中太多的图像处。

(A2) （相似性）如果一个博客图像 i 相对于 k 更类似于相片流 J，则 WIJ>WIK。

(A3) （概要）我们更喜欢对齐一个博客图像通过摘要 SI ∈ S。如果 $j \in S^l$ and $k \notin S^l$，那么 Wij > Wik 是被赞成的。

(A4) （连续性）连续的图片在相同的图片流中匹配图片是被鼓励的，它使得图片差值测试更加容易。

(S1)（对准）因为博客图片都是代表性的和有选择性的，图片的摘要应该尽可能从博客图片中有许多反向的对齐链接。那就是，W∗SI ∈ R|I|×|SI|，对 SI 来讲，从所有博客图片的一部分稀疏矩阵是不稀疏的。

(S2) （覆盖面和多样性）摘要 SI 应该尽可能包含一些冗余的图片，并且不要错过任何图片流中重要的图片。

正如上面描述的，计算 W 需要 A3 中的结果图片的摘要 S，反过来，计算摘要 S 也需要 A1 中的 W。所以，我们初始化 S，并且备份更新的 W 和 S，知道他们收敛。我们用隐变量制定对齐和摘要作为两组排序的 SVM 问题。(例如 [9, 15, 28])，这一方法的主要优势在于它的灵活性；可以方便地添加额外的约束，同时使用完全相同的方法和优化策略，这个可以更高效的解决随机下降梯度的问题。对准和聚合的细节将在第 3.1 节和 3.2，分别进行讨论。

**初始化** 我们首先得到一个初始化的图片流的集合，用 S 表示，通过应用 K 均值的图像描述符的聚类。我们设置 K= 0.3|PI|，|PI|是图片流的大小处。我们使用一个相对高点的 K 值作为初始化的摘要去选择图片。随着迭代过程，当减少冗余是，S(t)被更新，包括一个各式典范图像的集合。

### 3.1 博客和照片流之间的一致性

从博客图片到图片流对齐的目的是发现一种对应关系（例如 计算 W ∈ R|I|×|P|）。我们假设初始化 S(o)是可用的。

我们基于潜在排名的 SVM[8, 28]设计我们的校对优化方法，为了满足先前讨论的 A1-A4 的约束，最小化了一个基于损失的正规余量。我们解决单独每个博客的优化。使用 Wn ∈ R|In|×|P|表示一个博客 Bn 的稀疏矩阵的一部分。

$$\min_{\mathbf{W}^n,\xi} \frac{1}{2}\|\mathbf{W}^n\|_1 + \frac{\lambda_A}{M}\sum_{i=1}^{M}\xi_i \qquad (2)$$
$$\text{s.t. } \forall(i,j,k)\in\mathcal{C}_d\cup\mathcal{C}_s\colon \mathbf{W}^n_{ij} - \mathbf{W}^n_{ik} \geq \Delta(\sigma_{ij},\sigma_{ik}) - \xi_i$$
$$\forall(i,j,k)\in\mathcal{C}_c\colon \mathbf{W}^n_{ij} - \mathbf{W}^n_{ik} \geq \Delta(\#_{ij},\#_{ik}) - \xi_i$$

λA 是一个标准化的参数，M = |Cd| + |Cs| + |Cc| 是示例约束的数量。Eq（2）的目的是为约束 A1 使用 L1 的标准乘法替代 L2，鼓励稀疏矩阵 Wn。通过 A2-A4 三个约束编码设置了 C，包含了一个巨大数量的图片元组（I,j,k），i 取样自 Bn，j，k 来自图片流。

首先，Cd 是为 A3 的相似约束，当一个博客图片 i 的相似度对 j 而言比

k 大，Wn i,j ≤ Wn j,k 时进行处罚。我们用Δ(σij,σik)=|σij −σik|作为损失函数，σij 是 i 和 j 之间的相似度。

第二，Cs 是对（A3）的摘要约束。我们喜欢用总结的 S 图像进行匹配。所以，如果 j ∈ S 并且 $k \notin S^l$，那么 Wn i,j ≤ Wn j,k 是处罚过的。

最后，Cc 是对 A4 的连续性约束，它在同样的图片流中促进了博客中连续图片的匹配。假设 j ∈ Pl 并且 k ∈ Pm。我们定义#ij =Pp∈P l(σi−1,p +σi+1,p)/2|Pl|，着表明 i 相邻的元素 (i − 1, i + 1) 和 Pl 之间的相似特征。通过损失函数Δ(#ij,#ik) = |#ij −#ik|，如果 i 的邻居相对于 Pl 和 Pm，比 Pl 更相似，那么 Wn i,j > Wn j,k 是被鼓励的。

**约束代** Eq（2）的优点是早一个约束下的各种关联的对齐的目标更灵活，可以更容易地分割。然而，如果我们考虑三元组的所有可能组合 C 的尺寸可能非常大。举个例子，最大的 Cd 是 |Bn|·PL l=1|P l| 2，其中 L 是图片流的数量。

因此，我们生成如下的 C。首先，我们发现了对每一个博客 K 最近的图片流 N(Bn)，使用 Hausdorff 距离度量，并且普遍的约束的例子从 N(Bn)中来。我们设置 K = c · log(|P|)，其中 c=4。然后根据容许的计算资源，我们混合 C（如每个博客 5~10k）。对于每一个博客的图像，我们使用的是加权随机抽样不放回抽样三重约束。例如，我们通过从 j ∈ S 选择一个图片，为每一个博客图像生成一个 Cs。如果 σij > σik，我们生成一个三元组(i,j,k)。我们重复添加三元组，直到达到固定的设定规模。

**优化** 我们优化 Eq.(2)通过使用一个在线随机坐标下降算法在[23, 24]中。由于数据集的博客和照片流是大规模和可能不断增长，这是对使用随机梯度下降的制定是有好处的，其收敛速度快比一个批处理算法更好。我们补充提出了详细的推导和伪代码。

## 3.2 照片流概述

对于每一张照片流 PL，目的是发现总结的集合$S^{l*} = \text{argmax}_{S \subset P^l} s^l_S$，其中 $s^l_S$ 是对任何集合$S \subset P^l$的排名得分，S 表示排名分数的子集 S⊂PL。尽管所有可能子集的大。小是指数。（如$2^{|P^l|}$）我们将通过限制子集探索的次数，提出以下易处理的近似算法。我们计算基于所述相似潜排名 SVM 排名分数 S1 到满足约束条件的第 3 节中的（S1）- （S2）。

$$\min_{\mathbf{s}^l,\xi} \frac{1}{2}\|\mathbf{s}^l\| + \frac{\lambda_S}{M}\sum_{i=1}^{M}\xi_i \quad (3)$$
$$\text{s.t. } \forall (S^l_i, S^l_j) \in \mathcal{C}_p : \mathbf{s}^l_i - \mathbf{s}^l_j \geq \Delta_S(S^l_i, S^l_j) - \xi_i$$

其中 λS 是常规参数，并且 M = |Cp|。损失函数是$\Delta_S(S^l_i, S^l_j) = |\kappa(S^l_i, P^l) - \kappa(S^l_j, P^l)|$，一对$(S^l_i, S^l_j)$的集合

$$\kappa(S^l_i, P^l) = \sum_{p \in P^l} \max_{s \in S^l_i} \mathbf{q}_s \sigma(p,s) - \alpha|S^l_i| + \beta\nu(S^l_i). \quad (4)$$

通过在工作部分的启发在线游客图片的总结[25]，函数$\kappa(S^l_i, P^l)$被如下定义。公式（4）的第一个任务是一个加权的 K-均值目标，以提高在（S2）中的摘要报道。Qs 从博客中作为一个对应链接的被拿出的图片的摘要，会被鼓励.（例如 更多图片有反向链接，更高的 qi 值被分配）我们计算如下的权重向量$\mathbf{q} \in \mathbb{R}^{|P^l|\times 1}$。我们首先在 Pl 的图片之间建立一个相似的图表 Sl，其中连续的图片被一个链关联，并且权重通过相似度的因素被计算。然后建立一个集成的相似矩阵 $\mathbf{U} = [\vec{0}\ \mathbf{W}_{P^l}; \mathbf{W}^T_{P^l}\ \mathbf{S}^l]$，$\mathbf{W}_{P^l} \in \mathbb{R}^{|\mathcal{I}|\times|P^l|}$ 代表相片流中所有图片的相似的的投票结果。我们从 U 中计算页排名向量 v，并且最后的|Pl|尺寸部分 v 变成 q。公式（4）的第二项为简洁的图片过多的惩罚的总结，第三项是 SL 是形象描述的

变化，鼓励多样性。参数 α 和 β 可通过交叉验证进行调谐。

如前面所讨论的，优化公式（3）的 关键难点是可能的 S ⊂ PI 的集合的数量是指数级的。为了应付这个问题，我们通过使用算法 1 的贪婪算法，如在结构的 SVM 广泛用于生成约束接近为子集选择的问题[15，29]。我们的想法是，我们选择对的子集作为约束汇总考生，利用贪婪算法，我们在其中反复的图像添加到一个子集，使目标的最大增幅。由于我们的子集选择可以看作预算最大覆盖问题的一种特殊情况，贪婪方法允许（1 - 1/ e）的近似结合[25]。对于公式（3）优化，我们使用了类似的网上随机坐标下降求解[23]。由式（2）唯一的区别是ℓ2 的惩罚，这是与常规的 SVM 是相同的。

**Algorithm 1:** Greedy algorithm for constraint generation.

**Input**: (1) A photo streams $P^l$. (2) A size of subset $K$
**Output**: (1) A pair of subsets $(S_i, S_j) \in \mathcal{C}_p$.
**1**: Initialize $(S_i, S_j) \leftarrow (p_i, p_j) \in P^l$ by randomly sampling two images from $P^l$ according to **q**.
**for** $i = 1 \cdots K$ **do**
  **2**: $S_i \leftarrow \text{argmax}_{p \notin S_i} \kappa(S_i \cup \{p\}, P^l)$. Repeat for $j$.
  **3**: If $\kappa(S_i) < \kappa(S_j)$, swap $S_i$ and $S_j$.

### 3.3 插值

一旦获得对准 W 和总结 S 的结果，我们便进行如下连续的博客图像之间的插值。我们首先建立每个照片流 PL，在此我们通过第一阶马尔可夫链特性的相似性的边缘的权重表示在照片流的图像的序列的对称邻接矩阵 SL。那么我们所有的照片流的结合 SL 构建块对角矩阵（公式）。我们再创造一个集成的相似矩阵（公式），它可以被看作是所有博客的图片和照片都流之间一个很大的相似矩阵。

如果给出两张博客图像（公式），我们可以应用日元的算法 V 至产生 k 个最短路径之间（公式）。最后一步，我们只挑选了图像摘要 S。尽管 V 规模是非常大的（公式），路径规划速度很快，因为 V 是极其稀疏，并且连续的博客图像图形非常接近。

## 4. 试验

我们集中于展示，在总结、探索方面，照片流和博客相互之间有所帮助。关于论证博客对于照片流效用性，有两个语义上的知识传输任务：图片定位和图片的自动命名。然后，我们用博客报道照片流的总结结果。最后，应用一大套照片流评估博客照片之间的轨道插入。

### 4.1 数据集

**照片流数据** 通过查询关于迪士尼乐园的关键词，我们可以从 Flicker 上下载照片。然后，我们可以手动的放弃那和迪士尼乐园无关的或者少于 30 张照片的照片流。结果，选择出了 6026 个生动照片流的 542217 张独特的照片，每个照片流都是一系列由同一个摄影师在一天之内拍摄的照片。

**博客数据** 表格一总结了博客的数据集。首先，在谷歌中，我们通过改变查询项目，在 blogspot, wordpress,Typepad 三大知名网站中查阅了 53091 篇独特的微博帖子和 128563 张有关联的照片。然后根据 Travelogue,Disney,Junk 这些公园中的代表，博客被手动地分为了三大组。那些带有 Travelogue 标签，表明了我们兴趣的博客通常都用多样的迪士尼照片描述故事和活动。迪士尼的标签一般被用在那些迪士尼相关的博客上而不是 Travelogue，例如，迪士尼电影，卡通或者商品。为了实验，我们使用了 Travelogue 博客，其中包含了 10075 条帖子和 121251 张相关图片。

### 4.2 图片定位的结果

**任务** 我们评估是否博客可以

使语义上的知识填充那些丢失信息，或带有混乱附属物的照片流。尤其地，评估博客能够辅助解决照片定位问题。迪士尼包含有各种各样的区域（例如：未来园），而且其中的每一个都包含一系列景观，该任务只在找到哪出景观被拍摄到了。对于地面实况，可以使用专业的贴标签机，利用 GPS 信息对 3000 张照片流进行注释，然后从中随机抽取 2000 张进行定位测试，并重复该实验十次。

表格二 图片定位精确度 我们报道 了

| Method | Top-1 Attr. | Top-5 Attr. | Top-1 Dist. |
|---|---|---|---|
| (JointRSVM) | **9.12%** | **22.83%** | **28.81%** |
| (KNN+KM) | 7.34% | 18.62% | 24.27% |
| (DTW+KM) | 4.05% | 15.31% | 21.03% |
| (VKNN) | 5.16% | 16.85% | 22.12% |
| (VSVM) | 4.63% | 15.80% | 20.63% |
| (Rand) | 0.93% | 4.63% | 5.56% |

Table 2. Image localization accuracies. We report top-1 and top-5 attraction accuracies, and top-1 district accuracies. Despite (JointRSVM) being only weakly supervised with blog text, it has twice the accuracy of fully supervised (VKNN) and (VSVM).

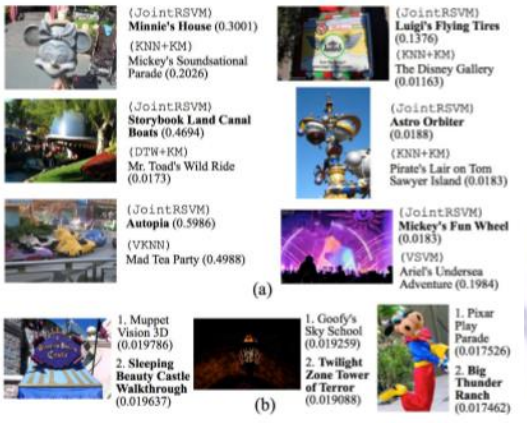


图 2.本地化的例子。（一）我们的方法，并在每种情况下的最佳基线，与置信度值之间的结果比较。（二）典型的似是而非的故障。正确的答案（粗体）估计作为第二个最好的。从左至右依次为：（一）景点相似的外观，（二）图像太暗，及（iii）字符只有弱与位置相关联。

第一到第五的景观精确度和第一街区精确度。尽管博客文章的指导很微弱（JointRSVM）,它还是存在两次全面的指导（VKNN and VSVM）。

图形二 图片定位的例子 （a）理论和每一案例中最优基准线对比结果（b）典型的未遂失败事件。正确答案（粗体字）预测和第二个最好的一样。从左至右：（i）相似景观外观（ii）照片太暗（iii）特征和位置联系微弱

从位置层面出发，我们使用了从两个迪士尼公园（Disney California Adventure and Disney Park）的 18 个街区中挑选的 108 个景观和饭店，并且在游客地图上找到了这些景观。在补充中，我们描述了应如何应用理论和基准线的细节。

**基准线** 我们将理论与四条基准线进行了对比，检测了两项基于视觉的理论—只关注照片的视觉效果。通过查询 108 个景观名称，我们在谷歌和 Flicker 上下载了 100 张顶级照片。使用这些照片作为训练数据，我们学习了分别用 VSVM 和 VKNN 指示的线性 SVM 和 KNN 分类器。这些对比可以用来证明博客数据使用的合法性。我们还实施了两条采用相同博客数据但运用了不同算法的基本准则。KNN+KM 使用 KNN 搜寻博客和照片流之间的队列，K-意味着收集照片流的总结。DTW+KM 开发 DTM 寻找队列，同样的 K-意味着总结收集。

表格三 图片命名结果 左：显示了那些我们预测的比原始标题合适的失败之作 右：根据投票给我们的失败之作的数量呈现检测样本的数量，因此有 156 个样本我们的标题全票通过。

我们使用 JointRSVM 指示我们所有的模型，因为位置关键词从博客向照片流中转移，我们采用 3.2 中的理念，用命名的实体位置引出。我们还报道了机会表演以体现定位任务的复杂性。

**结果** 表格二体现位置精确度 利用博客文章的理念胜过基于视觉的算法，我们的理念是很精确地，尽管只有微小的指导性，仅成功了两次基于视觉的全面指导。DTW+TM 表现不太理想是因为博客中的照片排序与照片流中的不匹配，我们注意到这项任务是很有挑战性的，只有低于 1%的机会。在很多案例中，即使对于一个专家而言也很难定位出与内容相符的位置，有可能匹配出不同的位置信息（比如，米老鼠事实上可能在任何位置被发现）。图表二展示了位于最优基准线前三行与定位的对比，同时，最后一行展示了典型的未遂的失败之作。

## 4.3 图像自动命名的结果

**测试** 自动图像命名就是为图像生成一个描述性标题的任务。网上图片经常没有标题或是相机自动分配无意义的代码。（例如 IMG1136.jpg）在这个测试中，我们量化多少 Amazon Mechanical Turk (AMT)的照片已存在的照片的命名得到改善。我们随机抽样 500 张照片流，并通过传送获得的语义的关键短语 2.3 节中的，生成标题。通过我们的方法发现以上的排列链接。我们展示了土耳其图像查询 Iq，并且原有的和估计的标题在乱序中。然后，我们问土耳其人选择哪一个图像更好。我们获得网络连接的答案已经为每个不同的土耳其人进行了查询。

**结果** 表 3 汇报了结果。即使考虑一定程度的 AMT 标签不可避免的噪声干扰，我们的产量是通过 AMT 注解是有意义的。我们的算法获得选票 64.2%，这证明我们的假设，即博客帮助改善互联网上的有噪声的图像的语义理解。请注意，一些测试图像用户分配到高品质的标题，但不是很多。

图 3 示出了具有实际的和估计的标题的查询图像的一个例子。前两行目前情况下，我们在那里标题是比原来的更好，并且在最后一行显示故障情况下，它原来的标题比较好。

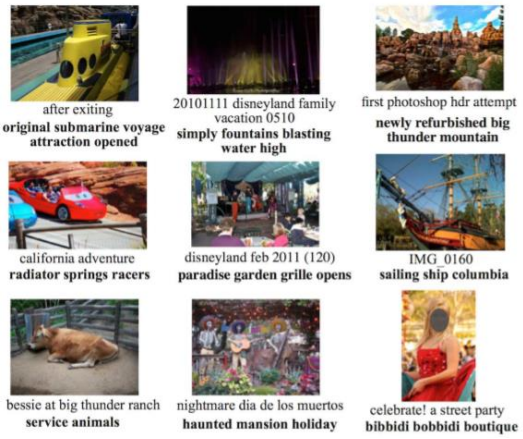


图 3.起标题的例子。 前两行说明情况，其中我们自动生成的标题（粗体）比原件更好（细体）。我们的 titles 发现并指定正确的景点名字（例如 radiator springs racers）或者替代无意义的标题，换成更语义丰富，更有意义的一个。（例如 IMG0160 为到哥伦比亚航行的帆船）第三行显示似是而非失败的案例。在所有的没有意义的原始标题有更高的质量。 从左到右 :（i）我们的标题是正确的，但少了特殊性（例如牛的名字贝茜）（ii）图像的框架导致我们预测朝向一个隐没的地方(iii) bibbidi bobbidi 是一个许多小公主可以看到的独特的景点。

## 4.4 图片流总结的结果

图 4 显示了汇总结果三个选定的照片流的定性比较。我们展示了通过我们的算法在两次迭代选出的前 6 幅图

像(t = 0 ， t = 2)。我们的算法中在大多数情况下，在 2-3 次迭代后收敛。我们也展示均匀采样的图像的简单（UNIF）基线。在 t = 0 的结果表示其中我们应用初始基于内容的 K-均值聚类到的图像的视觉描述的摘要。该（unif）是选择是有语义意义的风险的。（例如在第三例子白开水图像)在 S(0）仅设有可以由使用低级别的限制。例如，如果一个相片流是非结
构化和包括许多不良的拍摄的照片，汇总可能包括这样令人不快的图像。另一方面，虽然我们的方法使用相同的低级别的功能，它可以很容易地发现代表图像由于通过博客写手语义意图和价值选择博客图像相似性的投票。

## 4.5 博客图片之间的插值结果

我们现在展示使用一大组照片流的博客图像之间插入的定量和定性的结果。地面的实景是不可用的;因此，我们通过 AMT 进行用户研究。我们首先随机抽样 300 对连续图像在博客作为测试集记定义为(I1 q,I2 q) ∈ IQ。我们运行算法和基线产生 I1 q 和 I2 Q 之间的图像的最可能的序列。在 AMT 我们展示 I1 Q 和 I2 q 和那一双用我们的方法和基线的一个预测图像序列。我们问一个土耳其人来选择最有可能的结果。我们从获得网络的连接对（I1 Q，I2 Q）回答查询。我们用两个基线的共同使用的博客文章和照片流比较我们的算法（联合 SVM)：(KNN+KM)和(DTW+KM)。表 4 示出的我们的方法和两个基线之间的成对的 AMT 偏爱测试的结果。该数字表示选择我们的预测最有可能是 I1 Q 和 I2 Q 之间来回答的平均百分比。虽然问题是相当主观的，可我们的算法明显优于基准。我们的方法（联合 SVM）的结果是优选的 61.9％和测试用例比（KNN+ KM）和（DTW+ KM）基线 66.5％的。

图 5 示出由我们的算法和两个基线投影图像序列的实例。在每一组的左边，我们显示了两个查询博客图像;我们用我们的方法和行两个基线显示了估计的图像。因为


图 4.照片流汇总定性比较。我们发现用我们的方法（在初始化后 2 次迭代）创建的摘要的前六名图像和基线（UNIF）。结果变成两次迭代经过语义含义。


图 5.博客图像之间的插值的例子。左侧的图像对示出连续博客查询图像，右边照片序列是由

不同的算法的内插结果。

询博客图像和所检索的图像是不相交的，每个算法只能充其量从其它用户的照片流返回类似（但不相同）的图像。在大多数情况下，我们的插值是更连贯和有代表性的查询图像（Darth Vader 在底部右侧）。

实验结果清楚地表明，我们可以用稀疏选择博客图像之间利用照片流连接的细节。例如，在博客一个或两个图像可被选择用于给定的景点。它们可以是代表快照，但未能捕捉到场景或经历，我们的内插可以在其中填充它。

## 5. 结论

我们提出了一种利用照片流和博客的大集合优势的方法，通过联合故事为基础的进行总结和探索。为了实现这一目标，我们针对两个潜在排行的 SVM 问题进行交替优化调整和总结。在博客和 Flickr 迪斯尼乐园新收集的大型数据集照片流，我们发现，博客和照片流在各个任务中都互惠互利。

## Reference

**[1] A. Aizawa. An Information-Theoretic Perspective of TF-IDF Measures. Info. Proc. Manag., 39(1):45–65, 2003. 3**

**[2] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In CVPR, 2013. 2**

**[3] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning Down the Noise in the Blogosphere. In KDD, 2009. 2**

**[4] S. Fidler, A. Sharma, and R. Urtasun. A Sentence is Worth a Thousand Pixels. In CVPR, 2013. 2**

**[5] A. S. Gordon and R. Swanson. Identifying Personal Stories in Millions of Weblog Entries. In ICWSM Data Challenge Workshop, 2009. 2**

**[6] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition. In ICCV, 2013. 2**

**[7] R. Ji, X. Xie, H. Yao, and W.-Y. Ma. Mining City Landmarks from Blogs by Graph Modeling. In ACMMM, 2009. 2**

**[8] T. Joachims. Training Linear SVMs in Linear Time. In KDD, 2006. 4**

**[9] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-Plane Training of Structural SVMs. Mach Learn, 77:27–59, 2009. 2, 4**

**[10] G. Kim and E. P. Xing. Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction. In CVPR, 2014. 2**

**[11] G. Kim and E. P. Xing. Reconstructing Storyline Graphs for Image Recommendation from Web Community Photos. In CVPR, 2014. 2**

**[12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and Generating Image Descriptions. In CVPR, 2011. 2**

**[13] A. Kushal, B. Self, Y. Furukawa, D. Gallup, C. Hernandez, B. Curless, and S. M. Seitz. Photo Tours. In 3DIMPVT, 2012. 2**

**[14] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML, 2001. 3**

**[15] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning. In WWW, 2009. 4, 5**

**[16] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why We Blog. CACM, 47(12):41–46, 2004. 1**

**[17] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing Images Using 1 Million**

**Captioned Photographs. In NIPS, 2011. 2**

**[18] A. Qamra, B. Tseng, and E. Y. Chang. Mining Blog Stories Using Community-based and Temporal Clustering. In CIKM, 2006. 2**

**[19] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating Video Content to Natural Language Descriptions. In ICCV, 2013. 2**

**[20] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic Keyword Extraction from Individual Documents. In Text Mining: Applications and Theory, pages 1–20. John Wiley and Sons, Ltd, 2010. 3**

**[21] B. C. Russell, R. Martin-Brualla, D. J. Butler, S. M. Seitz, and L. Zettlemoyer. 3D Wikipedia: Using Online Text to Automatically Label and Navigate Reconstructed Geometry. In SIGGRAPH Asia, 2013. 2**

**[22] E. F. T. K. Sang and F. D. Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. CONLL, 2003. 3**

**[23] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In ICML, 2007. 5**

**[24] S. Shalev-Shwartz and A. Tewari. Stochastic Methods for L1 Regularized Loss Minimization. In ICML, 2009. 5**

**[25] I. Simon, N. Snavely, and S. M. Seitz. Scene Summarization for Online Image Collections. In ICCV, 2007. 2, 5**

**[26] N. Snavely, S. M. Seitz, and R. Szeliski. Photo Tourism: Exploring Photo Collections in 3D. In SIGGRAPH, 2006. 2**

**[27] C. Sutton and A. McCallum. Piecewise Pseudolikelihood for Efficient Training of Conditional Random Fields. ICML, 2007. 3**

**[28] C.-N. J. Yu and T. Joachims. Learning Structural SVMs with Latent Variables. In ICML, 2009. 2, 4, 5**

**[29] Y. Yue and T. Joachims. Predicting Diverse Subsets Using Structural SVMs. In ICML, 2008. 5**

**[30] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the Visual Interpretation of Sentences. In ICCV, 2013. 2**