指导教师：杨涛　　提交时间：2016/03/12

CVPR2015 Paper Translation

No: 01
姓名：王文萱
学号：2013302568
班号：10011305

关于群体场景理解的深度学习属性

Jing Shao1 Kai Kang1 Chen Change Loy2 Xiaogang Wang1
1Department of Electronic Engineering, The Chinese University of Hong Kong
2Department of Information Engineering, The Chinese University of Hong Kong
jshao@ee.cuhk.edu.hk, kkang@ee.cuhk.edu.hk, ccloy@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.h

摘要

在计算机视觉方面，对群体场景的理解是一个基本问题。在这项研究中，我们开发了一个多任务的模型以便共同学习，并且结合其表观和运动特性来理解群体场景。我们建议将群体的运动通道作为一个深度模型的输入，这样通道设计的灵感来自于群体系统的泛型属性。为了更好地展示我们的深度模型，我们构建了一个具有从8,257个群体场景中提取的10,000个视频新的大型WWW群体数据集，并且建立了一个拥有94个特性的属性集于WWW上。我们进一步测量在WWW上的用户表现并且和策划中的深度模型进行比较。大量的实验表明，我们的深度模型与基于基线的强群体拥挤方式相比，在交叉场景属性识别中表现出显著的性能改进，并且在多任务学习中深度学习特性方面表现出更加优越的性能。

1. 简介

在过去的十年中，群体分析领域对于拥挤场景的理解有着一个了不起的进化，其中包括群体行为的信息[38, 24, 26, 33, 3, 48,46, 45, 15, 27, 41,40, 44, 47]，群体的追踪[2, 32, 49]和群体的分类[1, 7, 16, 42]。这些进展引发了关于群体数据集的创建以及具有新型和鲁棒性特性的分析群体固有属性的模型。大多数顶尖的学者[7, 38, 3, 48, 24, 27, 15]对于群体的理解是场景的具体特定化，即群体模型是从一个特定场景和对于其他场景的简单概括。属性

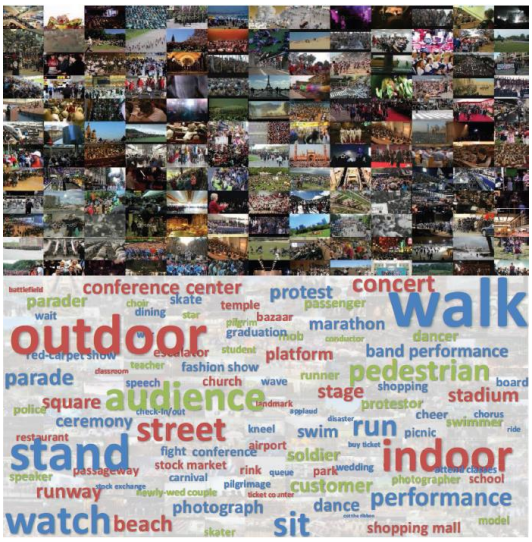

图 1. 一个关于 WWW 群体数据集及其属性的概述。红色表示位置（在何处），绿色代表主体（是什么），蓝色指示事件/行为（为什么）。每一个单词的面积正比于其在 WWW 数据集中的出现频率。

是对于特别场景的通用特点的有效描述。

近年来，研究对象的属性体现出事物[11, 20, 4]、面孔[19, 25]、行为[13, 23, 39]和场景[31, 28, 12, 30]作为一种替代或补充直接表示吸引了大量的注意，作为通过多属性描述目标事物而非被歧视分配到一个特定的类别，此限制对于描述目标事物的特性过于严苛。除此之外，科学研究[5, 8]表明，不同群体系统的相似原装可以通过一些常见的属性或特征表现出来。事实上，属性可以表达出更多的信息，如在群体视频中可以通过回答“这是何人”、“人群在哪里”、“为什么在这里产生群体聚集行为”来描述一个视频，而且这不仅仅是定义一个场景或者事件标签。例如，一个基本属性可以大致表示描述这个群体视频作为“指挥”和“合唱团”伴随着“观众”的“掌声”在“舞台”上表演，这样便与一个类似“合唱”的分类标签进行了鲜明的对照。近期，一些作品[33, 45]已经开始在群体属性分析上做出努力。但是，在他们进行的工作中使用的属性是有限的（仅仅只有四个[33, 45]），而且在多样性的场景中他们的数据集也很小。

在本论文中，我们引入了一个大规模的群体视频数据集，用于设计去理解（WWW）群体数据集1中群体场景的人物识别和定位。它包含了从8,257个群体场景中得到的10,000个视频。就最先进的知识而言，WWW的群体数据集是迄今为止最大的群体数据集。在WWW群体数据集中的视频全部来自于现实世界，是从不同的地方收集而来，并且由多种摄像头捕捉。我们进一步将94种有意义的属性定义为高层群体场景的表现，如图1所示。这些属性被从网上而来的群体视频中标签信息所引导。它们覆盖了常规的群体地点、主题、行为和事件。

从建模的观点而言，我们对于深度学习群体特性是否可以超越传统手工艺品特性十分有兴趣。因为视频，我们除了表观信息之外拥有了运动信息，我们可以同时从表观和运动方面来检查深度学习而得到的群体特性。和直接输入简单的一帧或多帧的深层神经网络相比，，我们建议使用运动特征渠道作为深度模型的输入。从设计的深度模型而得到的实验结论而言，我们揭示出具有中心属性的群体数据集使我们可以在传统的群体场景理解中做出更好的工作，并且在群体场景事件检测、群体视频检索、群体视频分类方面拥有巨大潜能。我们进一步设计一种用户研究用于衡量人类可以识别群体属性的精确性，并且何种类型的数据可以使得用户达到最高的精确度。本次研究具有必要性，且对于我们的实验提供参考评估至关重要。具体而言，观察人类的感知（提供不同数据类型时）结果和计算模型所得结论息息相关。

我们的贡献具体如下：

1）具有群体属性注释的最庞大的群体

	CUHK [33]	Collectiveness [45]	Violence [14]	Data-driven [32]	UCF [1]	**WWW**
# video	474	413	246	212	46	**10,000**
# scene	215	62	246	212	46	**8,257**
# frame	60,384	40,796	22,074	121,626	18,196	**> 8 million**
resolution	multiple	670 × 1000	320 × 240	640 × 360	multiple	**640 × 360**
source	Getty Images, Pond5, surveillance	web, surveillance	YouTube	web	BBC Motion Gallery, Getty Images	**Getty Images, Pond5, YouTube, surveillance, movies**

表格1. WWW和其余现存数据集的对比。WWW提供最大数量的视频、场景和帧数。

数据集——我们建立了一个拥有从8,257个场景中获取10,000个视频的庞大的群体数据集。设计了94个群体相关的属性并且对于数据集中的每个视频都通过注释加以描述。定义这样庞大的群体理解属性集实属首次。

2）对于群体场景的理解深度学习特征——我们开发了一个多任务的深度学习模型，为了共同研究外观及运动特征并且对它们进行有效结合。我们特别设计了群体运动通道作为深度模型的输入，而不是如同大多数现存的[17]视频分析工作，直接将多帧的数据作为学习运动特性的输入。运动通道的灵感来自于群体属性的通有性质，其在生物和物理领域被深入研究。伴随着多任务的学习过程，在深度学习特性时属性之间的相关性可以被很好的捕获到。

3）广泛的实验评估和用户研究探索WWW数据集——它们在静态表观线索和运动线索如何表现不同方面提供了有价值的见解，并且补充了三种属性类型:“地点”，“任务”和“原因”。它同时表明,那些针对于人类群体研究的特征比先进的手工特征要更加有效。

2. WWW群体数据集建设

多数的现存公开群体数据集[6, 9, 22,38, 48]仅仅包含一个或者两个特定的场景，甚至最大的那个[33]也仅仅提供从215个群体场景中获取的474个视频。相反的是，我们计划的WWW数据集提供从8,257个各式场景中获取的超过8百万帧的10,000个视频，因而我们对群体理解领域可以提高卓越的综合性的数据集。如此来源广泛充足的视频同时可以提升多样性和完备性。我们将WWW数据集和其他的公开可用的群体数据集对比，具体体现在表格1中。通过对比这些列在表格中的项目，我们的数据集在规模大小和多样性方面均优于其余数据集。

2.1 群体视频建设

收集关键字。为了获取一个大规模并且具有综合性的群体数据集，我们选择了一系列的与常规情景（如：街道、体育场和溜冰场）和群体事件（如：行军、合唱和毕业）有关的关键字，从而可以提升搜索的有效性。

为了使目标具有普适性，我们不包含由特定地点产生的关键字，但是使用一些可以描绘常规地点的普遍关键字用以代替。例如，我们选择了“地

标”作为关键字，而非使用一些特定地点的名字，如“时代广场”和“中央车站”。“地标”可以吸引游客群体的常识已广为人知。除了关于功能地点如“车站”、“餐馆”和“会议中心”的关键字，我们还包含一些特定类型的地点，如“自动扶梯”和“舞台”。即使这些可以被视作物体，它们也因为和群体具有高度相关性而广为人知。

收集群体视频。这些收集起来的关键字被用来从一些公用的视频搜索引擎如Getty Images、Pond5和YouTube上搜索视频。为了增加检索群体视频的成功率，我们在大多数的关键字上都加入了“群体”或者“团体”，除了一些已明确表示出群体的关键字（如：“合唱”和“马拉松”）。除了以上的三种来源外，我们还进一步从23部电影中收集了469段视频。为了控制视频的质量，我们删除了一些有着模糊运动、合成群体和极短时间的视频。另外，所有的复制视频亦经过过滤。

2.2 群体属性的注解

为了提供一个具有多样性的群体场景视频收录，数量庞大的可能属性被用来描绘不同的情节、主题和事件。这种可以推断出属性的能力使得我们可以通过提问“哪里有人群？”、“人群里有谁”和“这里为什么有这么多人”来描述一个群体聚集场景。重要的是，当我们面对一个全新的群体场景时，我们仍然可以使用以上3个关键问题（例如：新婚夫妇[人物]在沙滩[地点]举办婚礼[原因]）来描绘。

此外，在这些属性之间也存在着大量的可能交互作用。一些属性很可能和另外一些属性共现，即使一些看上去是独立的。例如，“街道”这种场景的属性更加容易和正在“行走”的事物“行人”一起出现，而且也很容易和事物“暴民”在有“战斗”时同时出现，但是却不会和事物“游泳者”相关，因为在“街道”上“游泳”是不存在的。另外，很多现存的属性是分组且分层次的，例如“户外的/室内的”便包含了绝大多数的其余描绘地点的属性。一些属性，像“运动场”和“舞台”也是属于“户外的”和“室内的”。

从网页标签收集群体属性。为了构建属性的分类，我们首先从Getty Images和Pond5收集了一些标签作为标签云的一种形式展示在图2中。通过仔细地检查总数达到7000+的恢复标签内容，我们发现从这些原始标签中产生独特而冗余的定义属性方式，因为它们大多数和我们关心的问题无关，甚至和群体无关，例如视频质量和环境条件。另外，一些高频率出现的（如：人类，成人，时间和种族）标签也很容易被丢弃。我们在清洗原始标签上付出了许多努力并且最终成功建立了一个有着94个群体相关属性的属性集，具体展示在图1中。它包含了三个方面的属性：（1）地点（如：街道、寺庙和教室），（2）人物（如：明星、抗议者和

溜冰者），（3）原因（如：步行、董事会和仪式）。可以在我们的项目网站查找到一个完整的列表。

群体属性的注解。我们雇佣了16位注解者来将WWW数据集中的属性分类标注。我们数据集中的所有属性都十分常见而且经常出现在我们的日常生活中，因而标注工作不需要特殊的背景知识。我们给每一位标注者均提供了一个每个属性同时含有正面和反面例子的数据集。他们的工作是对于每一个在实验室中的测试视频均选择出可能属性，其属性分别包含地点、人物和原因。在每一轮中，我们给每一位

图2. 原始的标签云（部分标记集）。字体越大的词表示其在数据集中出现的频率越高

标注者都展示了一个长达10秒钟的视频剪辑片段，并且要求他们在无时间限制的条件下从每一个属性单子中至少选择一个对应属性。

图3展示了在WWW群体数据集中的一些例子。就像图3中的第一行，并不是所有的视频剪辑片段都完整的描述了“何人”“在哪里”“在做什么”。因此，在标记之前，我们将“模糊不清”此选项加入到了每个属性单中。在所有的被标注的980,000个标签中，标注者选择“模糊不清”共有2855次，选择地点时占0.1%，选择人物时占0.2%，选择原因时占%0.4。在图3的第二行展示的两个视频证明了一个视频也可能有着为数不少的属性，也就是说，在一个单独的视频中可能存在多个事做不同的事情在不同的地点。

3. 群体属性方面的用户研究

在群体属性理解领域，表观和运动特性扮演着不同的角色。在这个部分中，我们在WWW群体数据集中进行了一项用户研究，调查当仅仅给出一类关键问题时人类的表现。这同时也作为一个关于比较我们在第5部分的实验经验的参考并且用于探索人类感知和计算模型之间的关系。

图3. 以上是一些WWW数据集的视频实例。第一行两个视频都有模糊不清的属性。然而另外两个第二行的视频在地点、人物和原因方面有着多重属性。

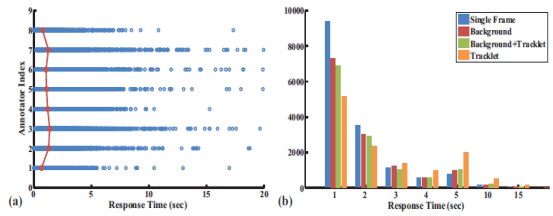

图4. 以上体现出用户的响应时间。图（a）中蓝色圆圈记录了所有标注者在标注工作时的反应时间，红色的线性标记了每位标注者的平均反映时间。直方图（b）展示了每一个问题的反应时间。

	Single Frame	Background	Background + Tracklet	Tracklet	Average
Accuracy	0.82	0.71	0.74	0.41	0.67

表2. 显示四类关键问题的选择精确度

我们提供给8位用户有着4种类型的数据，其中包括单一帧图像、背景、追踪图和有着背景的追踪图。相比较而来的真实体现在从整个视频中第3部分的注解集里。为了避免偏见的产生，我们为每位用户都提供了4种类型的所有数据和随机选取的10~15个属性。在标注之前，我们为每个标注者均提供了5~10个正面和反面的例子，从而帮助他们熟悉这些属性的含义。并且

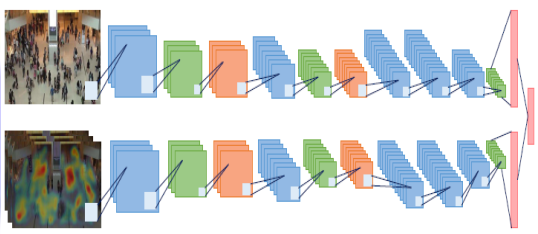

图5. 深度模型。有着相同深度值的表观和运动通道作为两个独立分支的输入。以上两个分支均包含多层的卷积（蓝色部分），大幅度合并（绿色部分），标准化（橘色部分）和贯通的连接（红色部分）。之后两个分支便融合成为一个整体层（红色部分）。

体验者会被告知他们的反应时间也将会被记录。

（1）*响应时间*：所有参与者的平均相应时间如图4所示为1.1094秒。图4还体现出，仅仅有多目标跟踪的识别更为复杂，并且人类识别仅有运动而无图片的群体属性来说更加具有难度。

（2）*精确度*：表2体现出仅有单一帧的体验者相比仅有多目标跟踪或者背景的可以获得更高的精确度。这意味着移动人群的表观特性和他们的姿势是很有用的，但是这些有可能在有背景图案的情况下变得模糊不清。我们可以发现背景和多目标跟踪可以做到互为补充。图6(a)体现出，仅有背景时，许多例子均可以被强有力的标记，并且很多与多目标跟踪的参与者的标记是相互关联的。极少的几个失败的例子中，最初的17个属性因其属性本属于“地点”而被多目标跟踪纠正。在接下来的23个属于“人物”和“原因”的属性中多目标跟踪都起到了更为显著的作用。图6(a)展现出对于属于“地点”的属性识别多目标跟踪表现较差。

4. 方法

我们为了学习从每个视频中的表

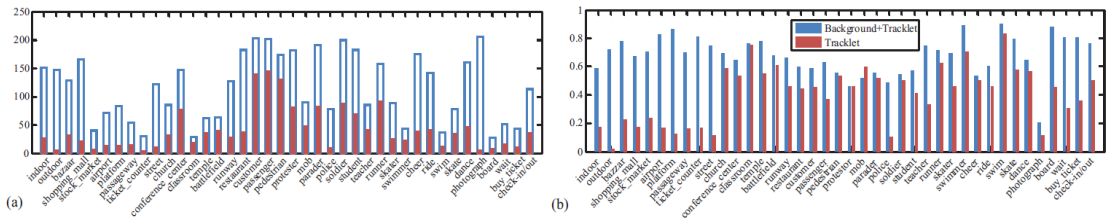

图6(a)错误标注的有着背景线索的例子（由蓝色条表示）并且注明经过多目标跟踪线索纠正的具体数量

图6(b)多目标跟踪线索和多目标跟踪线索+背景的精确比较。以上所有结果均来自于第3部分的用户研究

观和运动信息得到每个属性的特点而研究深度模型，并且将此学习模型用于不可视的群体视频属性的识别。

4.1深度网络结构和模型的确定

图5展示了深度模型的网络结构。这样的网络包含两个具有相同结构的分支。我们使用简单的符号来代表网络的各项参数：（1）Conv(N,K,S)表示卷积层，N表示输出，K表示核心，S表示幅度，（2）Pool(T,K,S)中用T表示合并层，K表示核心，S表示幅度，（3）Norm(K)用于表示标准化的局部反应，用了局部变量K，（4）FC(N)用N作为输出的全联通层，（5）在每一层的激活作用被ReLU使用校正线性单元和Sig型函数表示。之后，这样的两个分支便有了参数。

Conv(96,7,2)-ReLU-Pool(3,2)-Norm(5)-Conv(256,5,2)-ReLU-Pool(3,2)-Norm(5)Conv(384,3,1)-ReLU-Conv(384,3,1)-ReLU-Conv(256,3,1)-ReLU-Pool(3,2)-FC(4096)。输出便完全由两个索引至FC(8192)的分支联系。最终我们便可以得到FC(8192)-FC(94)-Sig从而产生94个可预测的属性。方程式(1)便表示缺少的网络功能交叉熵。网络外观分支的参数便可以通过事先训练好的想象网络侦测任务[29]模型来确定。

$$E = -\frac{1}{N}\sum_{n=1}^{N} t_n \log o_n + (1 - t_n)\log(1 - o_n) \qquad (1)$$

在这个方程中N=94表示了输出的神经原个数，tn（n = 1，. . . ,N)是目标分类数，on（n =1，. . . ,N)是输出的可能预测。

4.2运动通道

传统深度模型的输入是一个简单帧的位图（RGB通道）或者多帧[17]。在此篇论文中，我们建议将三个独立场景的运动通道作为表观通道的补充。一些著名的运动特性如光流不可以很好的在群体场景下描绘运动模式，特别是在复杂多样的场景下。科学研究展示出不同的群体系统可以共享一个相似的原则，并且可以由一些通有性质来定义。通过一些针对于群

体集团引进的独立量（如：集合性，稳定性和矛盾性）的启发，我们发现了这些性能也同时存在于整个场景空间而且可以从场景层次方面进行量化。经过我们的再生成，集合性表明了在整个场景下个体的程度可以作为一个集体运动的联合表现，并且不论整个场景是否可以保持其拓拓扑特性，稳定性较为体现，同时矛盾性可以通过每个最相近的临近的有趣观点中的相互作用/摩擦力进行衡量。图7中的例子直观的阐明了每个性能特性。

通过KLT特征点跟踪所有的描述均可以被多任务跟踪很好的定义，并且他们中的每一个均可以在WWW数据集上由75帧的视频进行计算。我们首先为整个多任务跟踪点集定义K-NN (K = 10)图。因为我们并没有提前察觉团队，在整个场景中描述符为我们提供了[45]更适合提取每个多任务跟踪点集合性的方案。随着相似的见解[33]，我们设计了描述稳定性的符号，通过计算和平均在K-NN图中每个不变的临近点的数量。这揭示了一个事实，一个稳定的群体需要保持一个相似的最近的临近点集合。这样矛盾的描述被定义在[33]，根据基于组的过渡优先性。因而这在我们的项目中是不适合的。此外，我们将这个描述符通过计算在K-NN图中相近的多任务跟踪相关速度计算。我们将平均每帧描述符为每个运动特性在时间域映射到输出三个运动地图，便可以很好地展现出深度模型的输入。尽管也许简单的一帧需要十个或者百个多任务跟踪，但是整个跟踪点却可以形成一个完整的连续的特征图。这个简明的通道建设过程在图7的第一行具体展示了。

如5.4部分所示，这些运动通道的改进之处在于可以提升属性识别的能力。

5. 实验结果

5.1设置

我们将所有无序的WWW数据集数据分离送入训练过程，企鹅人并且将实验比调至7:1:2。记录了以上三个集合的94个属性均被分为正面的和反面的，并且通过重叠场景来进行独立场景的属性学习，在所有的试验中，我们均将ROC曲线(AUC)作为评估标准。

5.2深度学习的静态特性评估

为了评估我们深度学习的静态特性（DLSF）仅仅从表观通道而得来的，我们选择了一个最先进的在场景分类对比方向已经被广泛应用的一系列的静态特性。文献显示出Dense SIFT [21] and GIST [28]在描述整体图像内容方向有着不俗的表现，当HOG [10]被广泛应用于行人检测时。他们都拥有成文应用于群体场景理解的潜能。为了捕捉到全球资讯，我们将颜色直方图应用于HSV的颜色空间和自身相似性(SSIM) [34]的描述。另外，我们也将原来的局部二值模式(LBP)应用于在群体场景中的量化结构。

我们从每个视频的第一帧可以提取出六类特

征量，同时可以建立静态特征直方图(SFH)伴随着一组无序的单词管道和K方法的聚集和线性码[37]。线性的SVM

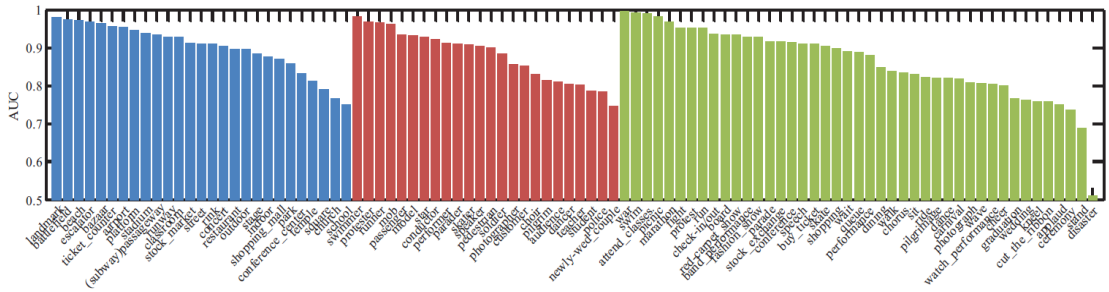

图8.每个从DLSF+DLMF获得属性的AUC值。蓝、红和率分别表示了位置、人物和原因的属性。

通过在每个属性上使用SFH被用于训练独立分类器。正如表3第二行所示，我们的DLSF方法要优于SFH基线。这样平均的AUC便可以提升6%。除去这总共的94个属性，它在其余64个属性（在最后一列显示）上有着更高的AUC。

5.3深度学习的动态特性评估

我们同时也在表3中报道了和两个基线相比的深度学习的动态特性表现。在4.2部分中，一个直方图表示我们的目标运动描述。另一个直方图则是展示出在行为识别方向最前沿的结果在稠密轨道[36]中。两个基线特征都有独立分类器训练，由线性SVM相似性至SFH基线。

根据在表3中第三行的结果，DLMF比其他的两种基线分别在AUC高10%和5%。对于超过77%的属性，DLMF可以比基线达到更高AUC值。另外一方面，DLMF和DLF相比有着接近20%下比率。这与我们在表2中参与者研究的观察结论相一致，即运动线索与表观线索相比普遍在属性识别方面的效果要差一些。

5.4结合性深度模型的评估

在图5中体现了DLSF和DLMF两个深度模型的结合。和5个基线相比。第一个出现的两个基线是结合了静态特征(SFH)和两个动态特征(MDH和稠密轨道)。我们加入了一个基线[18]可以通过对输入视频的建模提取时空运动模式(STMP)作为时空立方体的结合。它同时结合了表观和运动线索。第四个基线是慢融合体制伴随着近期以多帧作为深度模型的目标[17]输入。这种为了视频分析的最先进的深度学习方法，它成功的在运动学分类领域取得了最好的表现[17]。这是一个针对于深度学习框架是否可以良好的学习群体特征的有趣研究。同时最后一条基线是两个流的的卷积网络结果用于行为识别[35]。我们使用光流图（也就是用2个地图还替换每一帧，每5帧表示一个视频）来代替运动通道并且保持其表观特性不改变。根据表3的最后一行，我们用DLSF+DLMF结合的深度特征在所有的基线方面均有优势而且

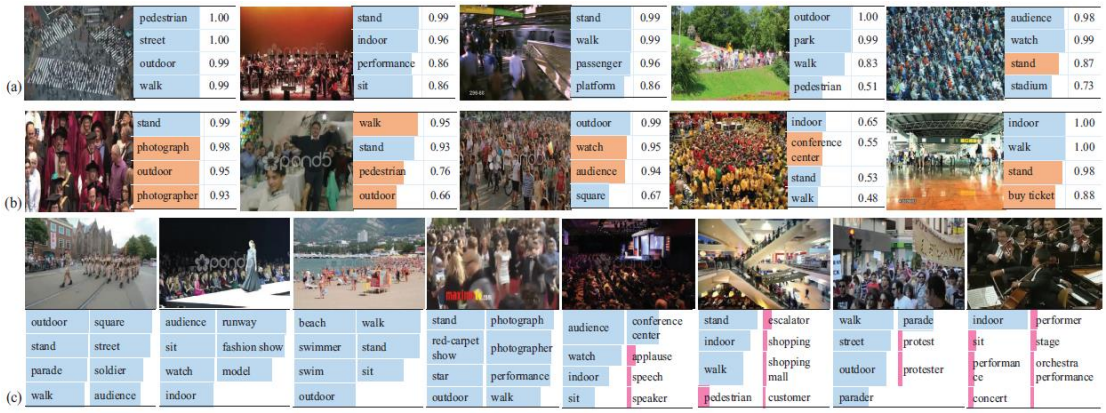

图9. 在(a)和(b)图中展示了良好和恶劣属性的预测样例。对于每一个图像而言，伴随着我们的DLSF + DLMF和最高预测分数的前四个属性已经显示出来了。这便是最高的预测分数。蓝色部分显示了正确预测的属性，橘色部分显示了错误预测的属性（失败警告）。在(c)图中所有的参与者都见每个图例标注了地面真实属性。如果在一个属性上的预测分数低于0.5，就会用红色标记出来，这表示了错误检测，其他地方则用蓝色表示。

Our Methods	mean AUC	Baselines	mean AUC	# wins
DLSF	**0.87**	SFH	0.81	67/94
DLMF	**0.68**	MDH	0.58	85/94
		DenseTrack [36]	0.63	72/94
DLSF + DLMF	**0.88**	SFH+MDH	0.80	78/94
		SFH+DenseTrack	0.82	72/94
		STMP [18]	0.72	89/94
		Slow Fusion [17]	0.81	74/94
		Two-stream [35]	0.76	89/94

表3. 和基线进行深度学习特征的比较。最后一列表明了在我们的试验中期望的深度特征比基线拥有更高的AUC值的属性个数。

STMP表现最差。缓慢融合[17]并没有使得手工艺更加出色的表现。为了捕获运动信息，这便是将多帧作为深度模型输入的原因。这也使得更大的网络结构拥有着更多的参数，也因此需要更大结构的训练数据。相似的，两个光流的结构[35]也包含了由十个运动通道产生更多的参数，并且选择性流动自身也不能很好的从不同场景中分辨出普遍特征。我们用三种运动通道作为深度模型的输入，这样便可以很好的总结运动信息和缩小网络规模。通过表3的总结结果，我们成功总结出了单独运动线索不能在群体属性识别中得到良好的学习运动属性(DLMF)，AUC的值提高了1%。一个结果。通过增加相对于深度学习静态(DLSF)的深度

详细的调查体现出相对于41种属性的AUC值均可以通过增加DLMF获得提升。这些属性绝大多数属于“人物”和“原因”。平均的AUC提升有15%。

定量评估。在图8中显示了每个有着DLSF+DLMF的AUC属性。不同的颜色分别从左至右代表了“地点”、“人物”、“原因”，并且这些结果均被降序排列。如“战争”此属性便得到了最高的AUC分数，然而“灾难”却是最低分。在训练过程中，最低分数往往出现在

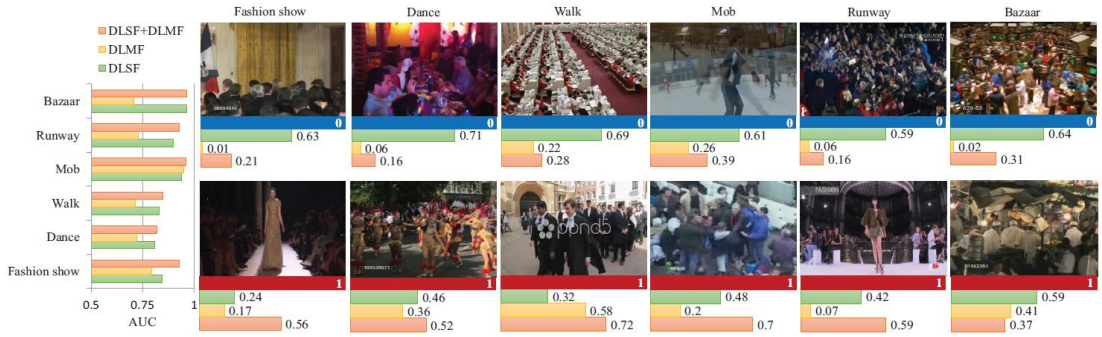

图10.表示分别被DLSF、DLMF和DLSF + DLMF预测的六个属性。第一行用蓝色条表示了反面实例，同时第二行用红色条表示正面实例。有着DLSF、DLMF和DLSF + DLMF预测值分数的被分别用绿色、黄色和橘色表示。

极少数的正面例子中。一些属性如“战场”、“暴民”和“战争”有着很强的相关性，及时它们分属于“地点”、“人物”和“原因”。它们均有着高AUC值。

定性评估。在图9(a)和(b)中展示了一些属性预测方面好的和坏的例子。在图9(a)标注的最后一个例子里展示了一个有着可能性高达0.87的“站姿”属性，当真实数据提示为“坐姿”。从人们的预测中得到的这个例子，这样要求了要确切区分“站姿”和“坐姿”。在图9(b)展示的第三个例子中，

综合的深度特征 VS 独立的深度属性

为了进一步核实我们拥有综合的DLSF和DLMF的深度特征拥有更好地输出，我们在图10中展示了6种有着定量(AUC值)和定性分析的结果。第一行给出了反面实例，而且表明了更高的预测可能性会产生更多的错误。相反的是，第二行中表现出更高的预测可能性会有着更高的准确度。普遍而言，最高的四个属性中有两个属性被错误预估。在第四个例子中“股票市场”被错误的认为是“会议中心”。这是因为人们在这个例子中不能很好的连续性移动，就如同“观众”或“观看表演”，同时它的表观线索如“观众”或“观看表演”便不能被很好地区分。图9(c)很好的展示了一些被错误检测的例子。提出的深度模型可以很好的识别有着独特运动和表观模式的属性，如海滩场景相关的属性。但是对于一些有着复杂和多样性表观和运动的属性而言却比较有难度，如一些和商场场景相关的属性。

DLSF提取了静态表观属性，因此在一些属性中指定运动模式会比较困难，如：“时尚表演”和“行走”。但是仅仅有着运动属性不能在有着相同运动模式的属性之间高效搜索。同样的，在第四列的反面例子中有着“溜冰”，但是给出的帧中体现出短截止图像是与“报名”或“打仗”更加相似。在DLSF或DLMF中组合模型融合外

观和运动通道以及补充丢失的线索，因此在所有样本的属性中显示出更加优越的性能。

5.5多任务学习

深度模型对于多任务学习是的一个全新的主意。我们将分别关于“地点”、“人物”和“语言”的三种不同的深度模型属性集而得到的结果进行对比。这称为单独任务学习。相比之下，以上讨论的深度模型我们称之为多任务学习。因为在不同类型的属性之间具有相关性，对三种属性集的联合训练含蓄的表现出普遍特征具有一起共享的相关属性。举个例子，一个“游泳者”应该在“沙滩”或者“行人”在“街道”走路。表4的前两列便展现出对于每种属性集通过单任务或者多任务学习所得到的平均AUC值。最后一列体现了在属性个数方面，多任务学习比单任务学习的表现更胜一筹。很明显的是通过多任务学习AUC值普遍从0.81提升至0.87。这样的精确度在大多数属性上都得以提升。

	Multi-task	Single-task	# wins
Where	0.89	0.84	22/27
Who	0.86	0.79	18/24
Why	0.86	0.79	36/43
Mean	0.87	0.81	76/94

表4.将单任务学习和多任务学习的平均AUC进行对比。最后一列是多任务学习在属性数量方面的表现要优于单任务学习。

6.结论

在此篇论文中，我们建立了一个从8,257个场景下提取10,000个视频的庞大规模群体数据集，并且计划了94个群体相关属性。这对于群体场景理解领域而言是一个重大的贡献。通过我们设计的深度模型学习了表观特征和运动特征。我们设计将运动通道作为群体系统通用属性的激励，而不用将多帧的数据作为现存视频分析工作的深度模型输入。通过多任务学习得到的群体特征，如在群体属性中具有相关性的部分也被捕获。关于群体特征的学习和群体属性的预测在未来的工作中具有很多的应用潜能，如群体视频检索和群体事件检测。

致谢

以上工作部分得到了来自香港的研究资助局（项目Nos.CUHK419412,CUHK417011,CUHK 14206114, 和CUHK 14207814），香港创新改革和技术支持规划项目(项目号ITS/221/13FP)，深圳基础研究计划(JCYJ20130402113127496) 的研究资金资助以及从NVIDIA公司得到的硬件支持。

参考文献

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*,2007. 1, 3
[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*. 2008. 1
[3] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR*, 2006. 1
[4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*. 2010. 1
[5] C. Castellano, S. Fortunato, and V. Loreto.Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591,2009. 1
[6] A. B. Chan, Z.-S. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008. 2
[7] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *TPAMI*, 30(5):909–926, 2008. 1
[8] H. Chat ́e, F. Ginelli, G. Gr ́egoire, and F. Raynaud. Collective motion of self-propelled particles interacting without cohesion. *Physical Review E*, 77(4):046113, 2008. 1
[9] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012. 2
[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6
[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1
[12] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10, 2007. 1
[13] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*. 2012. 1
[14] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR*, 2012. 2, 3
[15] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *CVPR*, 2009.1
[16] K. Kang and X. Wang. Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*, 2014. 1
[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2, 5, 7, 8
[18] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009. 6, 7
[19] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1
[20] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1
[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 6
[22] J. Li, S. Gong, and T. Xiang. Scene segmentation for behavior correlation. In *ECCV*. 2008. 2

[23] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 1
[24] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009. 1
[25] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013.1
[26] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010. 1
[27] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 1
[28] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 6
[29] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv preprint arXiv:1409.3505*, 2014. 5
[30] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 1
[31] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 1
[32] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Datadriven crowd analysis in videos. In *ICCV*, 2011. 1, 3
[33] J. Shao, C. C. Loy, and X. Wang. Scene-independent group profiling in crowd. In *CVPR*, 2014. 1, 2, 3, 5, 6
[34] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 6
[35] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 7
[36] H.Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 6, 7
[37] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 6
[38] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *TPAMI*, 31(3):539–555, 2009. 1, 2
[39] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 1
[40] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *CVPR*, 2015. 1
[41] S. Yi, X. Wang, C. Lu, and J. Jia. L0 regularized stationary time estimation for crowd group analysis. In *CVPR*, 2014. 1
[42] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015. 1
[43] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen. Rotationinvariant image and video description with local binary pattern features. *TIP*, 21(4):1465–1477, 2012. 6
[44] B. Zhou, X. Tang, and X.Wang. Coherent filtering: detecting coherent motions from crowd clutters. In *ECCV*. 2012. 1

[45] B. Zhou, X. Tang, and X. Wang. Measuring crowd collectiveness. In *CVPR*, 2013. 1, 2, 3, 5
[46] B. Zhou, X. Tang, and X. Wang. Learning collective crowd behaviors with dynamic pedestrian-agents. *IJCV*, 111(1):50– 68, 2015. 1
[47] B. Zhou, X. Tang, H. Zhang, and X.Wang. Measuring crowd collectiveness. *TPAMI*, 36(8):1586–1599, 2014. 1
[48] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *CVPR*, 2012. 1, 2
[49] F. Zhu, X. Wang, and N. Yu. Crowd tracking with dynamic evolution of group structures. In *ECCV*. 2014. 1