指导教师：杨涛　　提交时间：2016/3/18

CVPR2015 Paper Translation

No：01
姓名：崔鑫
学号：2013302545
班号：10011304

ChaLearn 2015 人体识别竞赛：动作识别与文化识别

摘要

在以前的系列人体识别（LAP）挑战竞赛中，chalearn 在 2015 举办了两场比赛中提出：行动/互动点在 RGB 数据文化事件识别。我们在人类活动对 RGB 数据序列识别第二轮。在文化事件识别方面，已经有了几十类。这包括场景理解和人的分析。本文总结了两者的挑战和所获得的结果。chalearn 比赛的细节可以在网站 http://gesture.chalearn.org/.

简介

人体在静止图像和图像序列中的自动分析，也被称为看人不断地进步，不断改进新的方法。这推动国家先进技术发展。相关的应用是很多的，如人机交互，人机交互，通信，娱乐，安全，商业和体育，同时具有重要的社会影响的残疾人和老年人的辅助技术。

2015 年，ChaLearn 组织新的竞争和行动/互动发现和文化事件识别大车间。连续识别，自然的人类信号和活动的识别是非常具有挑战性的，由于多模态的视觉线索（例如，手指和嘴唇，面部表情，身体姿势）的运动，以及技术的限制，如空间和时间分辨率的性质。此外，文化事件的图像构成一个非常具有挑战性的认识问题，由于服装，对象，人的构成和背景的高变异性。因此，如何结合和利用这些知识从像素构成一个具有挑战性的问题。

这激发了我们的选择，组织一个新的工作室和一个竞争，这一主题，以维持计算机视觉界的努力。这些新的竞争是我们以前的工作室的 CVPR 2011, CVPR 2012, ICPR 2012, ICMI 2013, and ECCV 2014 的新变化。我们继续使用我们的网站 http://gesture.chalearn.org 升级，而在数量竞争挑战条目评分采用 codalab 微软斯坦福大学平台在线（http：/ / codalab。org /），其中我们已经组织了国际竞赛，其中涉及到计算机视觉、机器学习问题。

在本文的其余部分，我们更详细地描述了我们所组织的竞赛和竞赛者取得的结果。

挑战题目和时间表

2015 的 ChaLearn 竞赛有一个定量的评定标准：行动/互动发现 RGB 数据和文化事件识别静止图像。这两个竞争的轨道的特点是以下：

行动/互动的识别：在总样本中，有 17 个人来提供 235 个简单动作。在选定的行动相关的肢体的运动和包括大多

数的相互作用在不同的动作。

文化事件识别：受到 PASCAL VOC 2011 －12 成功组织行为识别的挑战激励，我们计划举办一个比赛，包括了 50 种类对应不同的世界性的文化活动。在所有的图像类别，服装，人的构成，对象，照明，和上下文中做构成可能的线索被剥削的事件，同时保留固有的内部和类间变异的这种类型的图像。成千上万的图像下载和手动标记，例如相应的狂欢文化活动（巴西、意大利、美国），啤酒节（德国），圣费尔明（西班牙），胡里节（印度）和祇园祭（日本）等等。

竞赛是使用微软公司平台管理室。比赛日程安排如下：

2014 年 12 月 1 日

开始的定量的请愿行动/互动识别跟踪，发布的发展和验证数据。

2015 年 1 月 2 日

文化事件识别跟踪的定量研究之初，开发和验证数据的发布。

2015 年 2 月 15 日

为获取最终的评价数据的登记程序的开始。

2015 年 3 月 13 日

发布加密的最终评估数据和验证标签。参与者开始训练他们的方法与整个数据集。

2015 年 3 月 13 日

释放的解密密钥的最终评价数据。与会者开始预测的结果，最终的评价标签。这个日期是提交代码的最后期限。

2015 年 3 月 20 日

定量竞争结束。最后评估数据提交预测的最后期限。组织者通过在最后的评估数据上运行它，开始了代码验证。

2015 年 3 月 25 日

事实表提交的截止时间。

2015 年 3 月 27 日

比赛的结果公布。

竞赛数据

本节介绍了为每个比赛提供的数据集及其主要特征。

A.动作和交互数据集

我们提供的 hupba 配方+数据集[15]带注释的开始和结束的行动和互动框架。每个动作/交互类别的一个关键帧例子如图 1 所示。数据集的特点是：获得 9 个视频图像（RGB 序列），共有 14 个不同的演员出现在序列。它们图像序列已被记录使用一个固定的摄像头，具有相同的静态背景。

请来 14 人来采集 235 个行动/互动样本。

每个视频（RGB 序列）是在 15 帧的速度记录，每个 RGB 图像被存储在 BMP 文件格式的解析公式。

其中有 11 动作类,含孤立和协同行动：波，点，拍，蹲，跳，走，跑，握手，拥抱，亲吻，打。行动样本中有较高的类内变异。

演员出现在不同的姿势和执行不同的动作/手势不同的视觉外观的人的四肢的演员。因此，人类的姿势、自我遮挡和服装和皮肤颜色的变化都有很大的变异性。

关于执行动作和相互作用的长度差异较大。几分心行为的 11 类也存在。

其中运动轨迹的数据链表在 Table1 中。照片案例的数据集在 Figure1。

B.文化事件识别数据集

在这项工作中，我们引入了第一个数据集的基础上的文化事件和第一文化事件识别的竞赛。在这一节中，我们讨论了一些最密切相关的工作。

动作分类的竞赛[8]属于 pascal-voc 挑战这是视觉对象类别的识别和检测基准。特别是，在 2010 个 10 类的行动分类的挑战进行了介绍。这个挑战包括在一个静止图像中被一个人执行的行动的预测。在 2012 有 2 个不同的变化，（1）这取决于如何被确定在一个测试图像（我）的人（我）由一个紧包围盒周围的人；（2）只有一个单一的点位于身体上的地方。

社会事件检测[11]这项工作是由三的挑战及其元数据的图像常见的测试数据（时间戳、标签、地理标签为其中的一小部分）。第一个挑战包括在德国的测试集合中发现的技术事件。在二次挑战中，任务包括在汉堡（德国）和马德里（西班牙）在测试集合中找到所有的足球活动。第三个挑战旨在论证和在测试集在马德里公共场所发生抗议事件的愤怒者运动。

在这项工作中的社会媒体事件识别[1]作者介绍了社会媒体事件识别的问题。他们提出了一种增量聚类算法，将社会媒体文件分类成一组越来越多的事件。

表 2 显示了我们的文化事件数据集和其他在国家的艺术。行动分类数据是最密切相关的，但图像和类别的量是小于我们的。虽然图像和类别的数据集[11]和[1]是大于我们的数据集，这些数据集是不相关的文化事件，但在一般的事件。一些事件在考虑这些数据的例子是足球比赛（足球比赛发生在一月的罗马抗议事件（2010），在马德里的公共场所发生的愤怒者运动），等等。

C.数据集

文化事件识别挑战的目标是调查的基础上的一些线索，如服装，人类的构

成，对象，背景等的识别方法的性能，这一目的，文化事件数据集包含显着的变化，服装，行动，照明，本地化和背景。

文化事件识别数据集由 2 个图像搜索引擎（谷歌图像和冰图像）收集的图像。为了建立数据集，我们选择了 50 个重要的文化事件，我们创建了几个与这些事件的名称查询。为了提高检索到的图像的数量，我们结合一些关键词的事件的名称（节日，游行，事件，等等）。然后，我们删除重复的网址和下载的原始图像。为了确保下载的图像属于每一个文化事件，一个过程被应用到手动过滤每一个图像。接下来，所有精确的重复和近重复的图像被删除从下载的图像集使用的方法，在[3]。虽然我们试图删除所有重复的数据集，有可能存在一些剩余的重复，没有找到。我们相信这些数目是足够小的，所以他们不会显着影响研究。所有这些预处理后，我们的数据集是由 11，776 图像。图 2 中所示的绿色文化事件的数量由国家选择。

数据集可以在以下网址下载：https://www.codalab.org/competitions/2611。一些额外的细节和文化事件数据集的主要贡献如下：

第一个数据集来自全球各地的文化事件。

超过 11，000 张图片，代表 50 个不同的类别。

高帧内和类间变异。

对于这种类型的图像，可以利用不同的线索，如服装，人的姿势，人群分析，对象和背景场景。

评价指标将是识别精度。

图 3 显示了一些样本图像和表 3 列出了 50 个选定的文化事件，他们所属的国家和被认为是这一挑战的图像的数量。

文献中没有类似的数据集。例如，在 ImageNet 竞赛不包括文化活动分类为这个特定的轨道。在 PASCAL VOC 2011 - 12 作用分类的挑战，图像的数量是相似的，在 11，000，但类别数目增加 5 倍以上。

协议与评价

这部分介绍了协议和评价指标的两个轨道。

A.行动/互动追踪评估程序

评价行动/互动识别的准确性，我们使用 Jaccard 指数，越高越好。因此，对公式的作用和相互作用类别标记为 RGB 序列公式，Jaccard 指数被定义为

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}}, \qquad (1)$$

公式是行动/互动式序列式地面真理

和公式是这样的一个动作序列公式和公式的预测，1-values 二进制向量对应的帧中的 formulath 行动正在进行。参与者都是基于平均 Jaccard 指数对所有序列的所有类别中进行评估，其中运动类是独立的而不是相互排斥（在某一帧一个以上的动作，互动，手势可以活跃）。

在误报的情况下（如推测一个行动或互动不标记在地面的真理），Jaccard 指数 0 为特定的预测，也不会在平均 Jaccard 指数计算计数。换句话说，在地面真理和预测中出现的行动/互动类别的交集是相等的。

图 4 显示了 2 个动作的计算实例。请注意，在识别的情况下，不同类别的地面真理注释可以重叠（在序列中出现在同一时间内）。此外，虽然不同的行为的序列中出现的序列，在同一时间，行动/相互作用被标记在相应的时间（可能重叠），没有必要确定的演员在现场。

图 4 中的例子显示在动作序列类的不同实例的平均 Jaccard 指数计算（单红线表示地真相注释和双红线表示的预测）。在图像的顶部，一个可以看到地面的真理注释，行动行走和战斗在序列公式。在图像的中心部分是预测评估获得 Jaccard 指数 0.72。在图像的相同程序的底部与动作战斗和获得的 Jaccard 指数是 0.46 行。最后，平均 Jaccard 指数计算得到的值为 0.59

B.文化事件追踪评估程序

对于文化活动的轨道，参与者被要求提交给每个图像的每一个事件的信心。参与者均采用平均精度评价（AP），灵感来自于度量用于 pas-cal 挑战[7]。按如下计算。

首先，我们计算精度/召回曲线的精度单调递减的版本。这是通过设置的精度为召回公式的最大精度得到的任何召回公式。

然后，我们计算该曲线下的区域，通过数值积分。为此，我们使用了很好的梯形法则。让公式表示我们的精度/召回曲线的功能，梯形规则的作品，在这条曲线下的近似如下：

挑战结果和方法

这一节中，我们总结的方法，提出的最高层的参与者。八个小组提交了他们的代码和预测的最后阶段的竞争，第二行动/互动和六的文化事件。表 4 包含了最后一个团队的等级和分数的两个轨道，以及用于每个团队的方法是在本节的其余部分。

A.动作/交互识别方法

MMLAB

此方法是在[13]中提出的系统的一个改进，它由两部分组成：视频表示和时间分割。对于视频剪辑表示，首先提取改进稠密轨迹与猪，霍夫，mbhx，和 mbhy 描述符。然后，对每一种描述符,参与者训练GMM和用Fisher向量这些描述符变换到一个高维的超空间向量。最后，在整个视频剪辑中，用集合池来聚合这些代码，并用功率二级规范对其进行规范化处理。对于时间的识别，作者采取了时间的滑动方法沿时间维度。为了加快检测的处理速度，作者设计了一个时间的整合直方图的小鱼矢量，在任何时间窗口中，有效地评价池的小鱼矢量。对于每一个滑动窗口，作者使用汇集的的的的的的的向量表示，并将其注入到用于动作识别的支持向量机分类器中。此方法的概要如图 5 所示。

FKIE

该方法实现了一个端到端的生成方法，从特征建模到活动识别。该系统结合密集的轨迹和时间结构模型的动作识别的基础上一个简单的语法的行动单位的基础上的。作者修改原来密集的轨迹实施等。[19]避免使用一个轨迹来避免邻近的兴趣点（如图 6 所示）。他们使用一个开源的语音识别引擎，用于分析和分割的视频序列。因为一个大的数据语料库通常需要训练这样的系统，图像被镜像到人为地产生更多的训练数据。最终的结果是通过投票的输出的各种参数和语法配置。

B.文化事件识别方法

MMLAB

该方法融合了五种事件识别 ConvNets。具体来说，他们微调 Clari 辉净预训练的 ImageNet 数据集，亚历克斯网预培训的地方 googlenet 预训练数据集，在 ImageNet 数据集和数据集的地方，和 VGG 19-layer 网上 ImageNet 数据集。从这些五 ConvNets 预测分数加权融合的最终结果。此方法的概要如图 7 所示。

Upc-Stp

该解决方案是基于相结合的全连接的功能（FC）两卷积神经网络层（ConvNets）：一个 pretrained 与 ImageNet 图像和一个好的 chalearn 文化事件识别数据调整。一种线性支持向量机的训练，每个功能相关联的每一个功能层，并与一个额外的支持向量机分类，从而形成一个分层的体系结构。最后，作者提炼他们的解决方案，通过加权的输出的足球俱乐部的分类，从训练数据的时间建模的事件。特别是，高分类分数的视觉特征的基

础上被处罚时，他们的时间戳不匹配事件特定时间分布。此方法的概要如图 8 所示。

Mipal_snu

这种方法的动机是，训练和测试的判别区域，将提高性能的分类。受[9]的启发，他们首先提取地区的建议，这是文化事件识别的独特区域的候选人。工作[18]被用来检测可能有意义的不同大小的区域。然后，补丁使用深卷积神经网络（美国有线电视新闻网），其中有 3 个卷积层和池层，和 2 个完全连接的网络。在训练后，每一个图像补丁的测试图像的类的概率分布。然后，测试图像的类概率性确定为熵阈值后平均斑块的概率。

讨论

本文介绍了 ChaLearn 看着人们 2015 的挑战包括竞赛的主要特点（1）RGB 行动/互动识别（2）文化事件识别。2 个大型数据集设计，手动标记，并公开提供给参与者的性能结果进行了比较。分析最后参与测试集的团队所使用的方法，并将其上传的模型，得出若干结论。

对行动和交互识别案例 RGB 数据序列，所有车队使用改进的稠密轨迹[19]为特征，使用 PCA 降维。从分类的角度来看，无论是生成和歧视已被使用的团队，虽然支持向量机获得了更好的结果。这是本次比赛的第二轮，提出的方法优于第一轮比赛。然而，由于在竞争的发展阶段，只有两个决赛获得了更好的结果比基线和冠军的得分已经为 0.5385，它表明，有改进的空间，和行动/互动识别仍然是一个悬而未决的问题。

在文化事件识别的情况下，目前的趋势，在计算机视觉文学，深层次的学习架构是目前在大多数的解决方案。由于大量的图像所需的训练卷积神经网络，团队使用标准的预训练的网络作为输入到他们的系统，其次是不同类型的分类策略的复杂性和一些国家的最先进的方法计算的要求，使他们不可行这类比赛，时间是一个硬约束。然而，研究 GPU 计算的方法，已经被许多球队在两个轨道，使这些方法，对最后的结果有很大的影响。

致谢

我们要感谢这些比赛的赞助商：微软研究院、巴塞罗那大学、亚马逊、inaoe，visada，和加利福尼亚。这项研究已经研究项目 tin2012.38187-c02-02 部分支持，tin2012-39051 和 tin2013-43478-p. 我们衷心感谢英伟达公司捐赠与用于创建的文化事件识别跟踪基线的

Tesla K40 GPU。

参考文献

References

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in socialmedia. In Proceedings WSDM, 2010.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. CoRR, abs/1405.3531, 2014.

[3] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In ACM International Conference on Image and Video Retrieval, 2007.

[4] S. Escalera, X. Baro, J. Gonz ´ alez, M. Bautista, M. Madadi, ` M. Reyes, V. Ponce, H. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. ChaLearn Looking at People, European Conference on Computer Vision, 2014.

[5] S. Escalera, J. Gonzalez, X. Bar ` o, M. Reyes, I. Guyon, ´ V. Athitsos, H. Escalante, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclarof. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. 15th ACM International Conference on Multimodal Interaction, pages 365 - 368, 2013.

[6] S. Escalera, J. Gonzalez, X. Bar ` o, M. Reyes, O. Lop ´ es, ´ I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In ChaLearn Multi-Modal Gesture Recognition Grand Challenge and Workshop, 15th ACM International Conference on Multimodal Interaction, 2013.

[7] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. IJCV, 88(2):303 - 338, 2010.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and

A. Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2):303–338, June 2010.

[9] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pages 580–587. IEEE, 2014.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In In CVPR, pages 2169–2178, 2006.

[11] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In Proc. MediaEval 2012 Workshop, 2012.

[12] S. Park and N. Kwak. Cultural event recognition by subregion classification with convolutional neural network. In In CVPR ChaLearn Looking at People Workshop 2015, 2015.

[13] X. Peng, L. Wang, Z. Cai, and Y. Qiao. Action and gesture temporal spotting with super vector representation. In L. Agapito, M. M. Bronstein, and C. Rother, editors, Computer Vision - ECCV 2014 Workshops, volume 8925 of Lecture Notes in Computer Science, pages 518–527. Springer International Publishing, 2015.

[14] A. Salvador, M. Zeppelzauer, D. Monchon-Vizuete, A. Calafell, and X. Giro-Nieto. Cultural event recognition with visual convnets and temporal models. In In CVPR ChaLearn Looking at People Workshop 2015, 2015.

[15] D. Sanchez, M. A. Bautista, and S. Escalera. HuPBA 8k+: ´ Dataset and ECOC-graphcut based segmentation of human limbs. Neurocomputing, 2014.

[16] J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. Neural Processing Letters, 9(3):293-300, 1999.

[17] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. International Journal of Computer Vision, 104(2):154-171, 2013.

[18] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. International Journal of Computer Vision, 104(2):154-171, 2013.

[19] H. Wang and C. Schmid. Action recognition with improved trajectories. In IEEE International Conference on Computer Vision, 2013.

[20] L. Wang, Z. Wang, W. Du, and Q. Yu. Event recognition using object-scene convolutional neural networks. In In CVPR ChaLearn Looking at People Workshop 2015, 2015.

[21] Z. Wang, L. Wang, W. Du, and Q. Yu. Action spotting system using fisher vector. In In CVPR ChaLearn Looking at People Workshop 2015, 2015.