中国科学: 信息科学 2021 年 第 51 卷 第 8 期: 1287–1301
SCIENTIA SINICA Informationis
论文

大场景多对象的深度社交分组网络

李坤[1], 李万鹏[1], 孙晓琨[1], 方璐[2*]

1. 天津大学智能与计算学部, 天津 300350
2. 清华大学电子工程系, 北京 100084
* 通信作者. E-mail: fanglu@tsinghua.edu.cn

收稿日期: 2021–01–22; 修回日期: 2021–03–23; 接受日期: 2021–04–07; 网络出版日期: 2021–08-09

国家自然科学基金 (批准号: 61860206003) 和天津市应用基础与前沿技术研究计划 (自然科学基金) (批准号: 18JCYBJC19200) 资助项目

摘要 在计算机视觉中, 群体分析越来越受到人们的关注, 对图像中复杂人群进行分组是群体分析领域的基础技术需求. 现有的人群社交分组方法只针对固定人数的小范围场景, 不能处理真实世界中的大场景图像. 本文提出首个面向十亿像素大场景图像的基于深度学习的细粒度人群社交分组框架, 由一种图引导的全局到局部的划分策略与一个学习隐函数表示社交对交互模式的深度社交分组网络组成. 该框架可在大范围场景图像上实现准确的人群分组. 本文方法同样适用于小场景图像, 在小场景图像数据集上的实验结果表明, 本文提出的框架相比于现有方法取得了显著的性能提升. 相关代码与训练数据即将开源.

关键词 群体, 大场景图像, 深度学习, 社交分组, 图引导

1 引言

随着计算机视觉技术的不断发展, 行人检测、轨迹追踪等以个体为主体的计算机视觉技术 [1,2] 取得了显著成功, 然而面向群体的研究工作还不够充分, 故本文将关注点从个体转移到群体, 聚焦于如何自动估计静态图像中个体间关系以达到人群分组目的这一具有挑战性的问题上.

早期的群组检测方法使用具有位置特征的连通图 [3], 或是对位置特征进行简单的聚类 [4]. 考虑到同小组中大部分个体是相互注视的, Robertson 等 [5,6] 采用了将个体位置与头部朝向相结合的研究方法. 最近, 社会信号处理领域对该问题又提出了新的解决方案, 研究者们引入 F-formation 空间 [7] 的概念作为人群划分的基本条件. 由于 F-formation 空间的高约束性, 基于此的人群分组技术达到了最先进的性能. 基于 F-formation 空间的人群分组方法可分为两类: 对 F-formation 空间的霍夫投票法 (Hough vote for F-formations, HVFF) [8] 与以个体距离为先验的图匹配算法 [9]. 然而, 以上人群分组方

引用格式: 李坤, 李万鹏, 孙晓琨, 等. 大场景多对象的深度社交分组网络. 中国科学: 信息科学, 2021, 51: 1287–1301, doi: 10.1360/SSI-2021-0024
Li K, Li W P, Sun X K, et al. Deep social grouping network for large scenes with multiple subjects (in Chinese). Sci Sin Inform, 2021, 51: 1287–1301, doi: 10.1360/SSI-2021-0024

法均面向视场角有限、行人数量适中的小场景 [10, 11] 或是对象交互类型有限、交互模式简单的封闭场景 [12], 且人群分组的性能也有待提高. 大场景图像的视野范围更广、行人数量更多、个体间交互动作更丰富且群组关系更加复杂: 既有两三人组成的细粒度群组, 又有由细粒度群组构成的粗粒度、多粒度群组. 对大场景图像的分组分析, 更有利于群体研究的发展, 满足公共安全、智慧城市等领域的应用需求.

要对宽视场大场景图像中的远距离行人进行分析, 需要保证远距离行人的清晰度, 即图像的超高分辨率. 面对宽视场大场景图像采集与分析对于图像分辨率的需求, 清华大学搭建了十亿像素级阵列相机 [13~15]. 基于此阵列相机, Wang 等 [16] 构建了国际首个十亿像素动态大场景多对象数据平台 PANDA, 填补了大场景下高密度人群数据平台的空白, 为探索计算机视觉新方向提供了不可或缺的数据基础. 该数据集由采集到的丰富真实场景组成, 场景中的行人行为符合日常生活习惯, 并且每张图像都标注了脸部朝向、行人被遮挡率、行人轨迹、行人小组 ID 等信息. 本文以该数据为基础, 研究宽视场超高分辨率大场景图像的人群分组方法.

宽视场超高分辨率大场景图像下的人群分组将面临如下挑战: (1) 传统的分组方法不能直接处理高分辨率、宽视角的大场景图像, 深度学习方法对输入图像具有像素限制, 需要一种既能维护图像信息整体性又能保证分组准确性的大场景人群分组方法. (2) 大场景图像中包含大量行人, 宽视角下的行人粒度具有明显差异, 群组数量与目标大小的不确定性成为大场景人群分组任务的难题, 需要一种既能适应不同群组结构又能满足任意粒度的人群分组方法. (3) 在人数较多的大场景中, 行人间必然会存在严重遮挡、视觉平行等特殊情况, 需要一种既能满足个体完整性又能自适应处理特殊情况的人群分组方法.

针对以上挑战性问题, 本文提出一种全局到局部模式的大场景图像细粒度人群社交分组框架, 该框架适用于无限制场景图像, 能够对图像中的复杂人群进行细粒度划分. 首先, 框架中包含一种图引导的局部社交对检测方法, 与现有简单行人组合的原始方法相比, 本文的方法能够灵活地对任意群组结构的人群进行分组. 其次, 为了准确判断任意情况的社交对关系, 本文设计了一个多支路协作的端到端深度社交分组网络 DSGnet (deep social grouping network). 该网络以社交对的两种信息图作为输入, 学习各种交互情况所对应的特征表示, 能够自适应地对不同情况下社交对中是否存在交互行为进行判断. 最后, 所有局部社交对的交互情况将作为图像中全局人群分组的划分依据. 实验结果表明, 本工作提出的分组框架适用于原有小场景公共数据集与大场景 PANDA 数据集, 并具有较好的分组效果. 图 1 展示了本框架的工作流程与分组效果. 我们将开源相关代码与训练数据[1)], 为对本工作感兴趣的研究者们提供便利. 本工作的主要贡献如下.

- 提出首个基于深度学习的大场景图像细粒度人群社交分组框架, 使用图引导的划分策略与深度学习网络, 将社交分组问题形式化为连续的隐函数表示, 突破传统方法不能处理大场景图像的技术瓶颈.
- 针对群组数量不定、个体尺度差异问题, 提出一种动态变化的全局到局部的划分策略, 提高处理全局视图中任意群组数量、任意尺度目标问题的灵活性, 同时降低了对多人数群组情况的处理规模.
- 针对密集人群中存在的复杂空间关系问题, 提出一种空域与深度联合特征引导的社交关系估计方法, 所设计的深度网络可以很好地解决具有严重遮挡、视觉平行等特殊社交对组成情况下的分组问题.

2 相关工作

计算机科学中对复杂人群分组化是高度跨学科的一项任务, 在分析群体社会活动时一定会利用到

1) http://cic.tju.edu.cn/faculty/likun/projects/DSGnet/DSGnet.html.

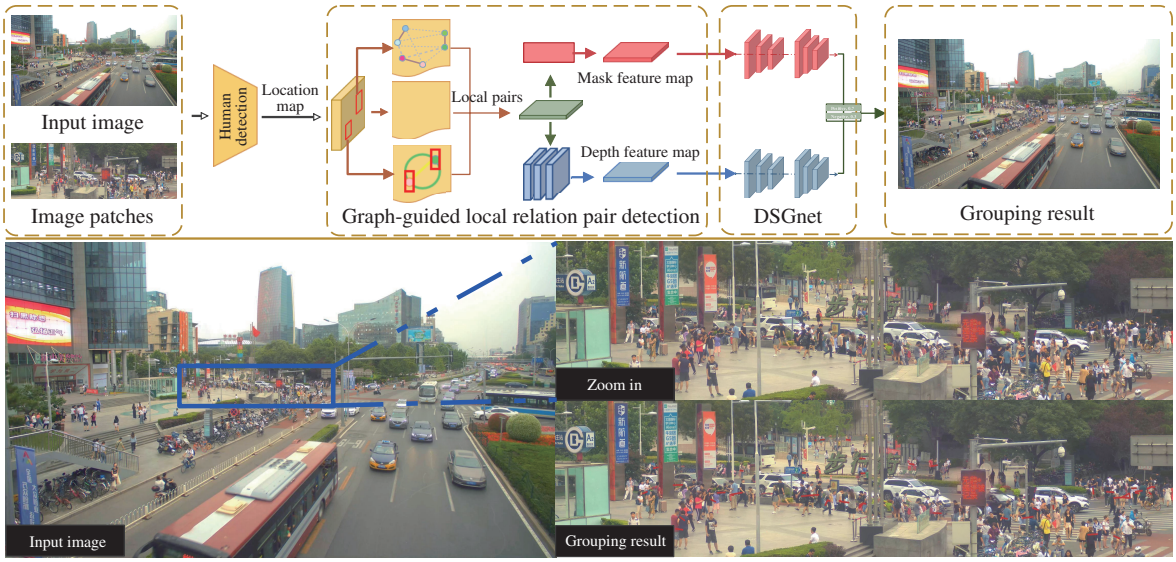

图 1 (网络版彩图) 面向大场景图像的细粒度人群社交分组框架及结果: 下方子图为十亿级像素图像输入、局部放大图以及本工作分组结果

Figure 1 (Color online) The framework and results of fine-grained crowd social grouping for large scene images: the subimage below shows the gigapixel-level image input, partially enlarged picture, and the grouping result of the framework

社会科学和认知科学的相关知识 [9]. 群组的概念是宽泛的, 它可以是在某时刻准备做相同事件的一群目标, 例如等待红绿灯过马路的两侧行人、车站进出口两个不同方向的人群等, 也可以是多数场景中相互交谈或是存在动作交互的聚集群组, 本文的研究目标面向后者. 即群组中的个体彼此紧密联系, 并将群组以外的个体排除在外. 分组问题根据是否利用时间信息可分为基于视频的动态分组和基于图像的静态分组两类.

早期的动态群组检测工作 [3, 4] 是通过在会议场景等密闭空间下利用传感器等设备, 对个体轨迹与动作进行建模来完成的. 该类方法解决了个体间具有遮挡或图像质量低情况下的分组问题, 但这类方法的场景是受限的, 不能直接移植到现实世界具有多样性的场景中. 为了解决这个问题, Hongeng 等 [17] 尝试使用半隐式马尔可夫 (Markov) 模型进行大规模事件检测, 用逻辑运算符表示个体的动作, 通过这些逻辑运算符的组合将相同事件组合成多线程同步事件解决人群分组问题. 视觉问题中人体轨迹追踪已经达到了很好的效果, Cheng 等 [18] 提出了一种根据运动轨迹对人群进行划分的方法, 在轨迹中引入了高斯 (Gauss) 过程以适应小组中人们活动的变化性, 提高了基于个体活动进行人群分组方法的鲁棒性. Ni 等 [19] 用个体因果关系、成对因果关系和群体因果关系对群体活动进行编码, 这些因果关系分别描述了不同个体运动轨迹之间的局部相互作用以及推理关系, 被用于解决人群分组任务. 上述方法主要关注于预测个体在场景中的行为. 另一方面, 预测整个场景中各个群体的活动标签也是分组问题的一种可行方法. 早期的该类方法通常使用手工提取特征并应用概率图模型 [20~26] 或语法模型 [26, 27] 进行群体活动识别. 近年来, 随着深度学习中循环神经网络 RNN 的出现, 这种既能够学习信息表示又能够对顺序数据中的时间关系进行建模的方法, 提升了动作识别问题的性能. Ibrahim 等 [28] 使用基于 RNN 多阶段的长短期记忆网络 LSTM 模型来表征个体级别的动作, 并结合个体级别特征生成组级别的动作表示, 利用动作表示划分复杂人群, 达到对人群进行分组的目的. 最近, 图卷积网络被用于学习 “演员关系图” 中个体的联系, 该方法能够同时捕获演员之间的外观和位置从而完成群体活动识别的任务 [29]. 但上述基于循环神经网络对个体动作进行建模的方法缺乏通用性, 它们专

注于面向特定的交互动作并且需要结合时间信息进行连续预测.

对视频中每帧图像的每个个体进行轨迹追踪等相关分析具有较高的时间成本, 而在大场景图像中又包含很多个体, 故需要基于单图像的静态分组技术来高效地完成人群分组任务. 起初, Bazzani 等 [30] 认为视锥面相交、距离近的个体间普遍存在交互行为, 以该想法为基础提出了相互关系模式矩阵法 (inter-relation pattern matrix, IRPM), 该方法利用头部方向推断个体的 3D 视锥并作为该个体关注焦点 (focus of attention, FoA) 的近似值, 结合 FoA 和邻近度信息估计个体间是否存在交互. Hung 等 [31] 的优势集法 (dominant set, DS) 和 Tran 等 [6] 的交互组发现法 (interacting group discovery, IGD), 都将个体定义为图中的结点, 在优势集法中边间权重为两点的亲和力值, 而在交互组发现法中边间权重定义为两端结点的交互程度, 交互程度基于两个个体注意力椭圆的交集: 椭圆之间的重叠越多, 它们之间的交互程度就越大. 近来, 被广泛使用的静态分组方法有: 霍夫投票法 [8] 和图聚类法 (dominant-sets for F-formations, DSFF) [31], 两种方法都以 F 形为基础, 这种形式的定义为: 每当两个或两个以上的人维持一种空间和方向关系时, 就会形成 F 形. F 形是三种空间的整体组织: o- 空间、p- 空间和 r- 空间. 解释来说, F 形是一种空间模式, 可以表征两个或两个以上的人的群体, 群体中的个体聚集在一起进行交谈、进行社交、共享信息并相互影响. 霍夫投票法考虑每个人的位置与面部朝向, 通过推断 F 形的 o- 空间中心位置构建每个群组的 o- 空间, 根据不同 o- 空间的组成人员, 对人群进行分组划分. 图聚类法则是, 将个体的位置与身体朝向整合到一个集群并对该集群进行建图, 通过边缘加权图中的主导集确定每个群组的 F 形. Setti 等 [32] 对两种方法进行对比, 发现在高噪声的、复杂遮挡的情况下, 霍夫投票法达到了更好的效果. Russell 等 [9] 结合了两种方法的优点提出了 F 形图切割算法 (graph-cuts for F-formation, GCFF), 该方法以场景中个体的位置和方向信息作为先验进行对 o- 空间的直接表述, 成为解决静态分组问题的最优选择.

现有的静态群组检测方法受到组内个体位置组合方式的严重约束, 同时需要很多的先验信息作为输入, 不能自适应地处理真实大场景中具有严重遮挡或对象间距离较近的泛性情况, 且需要输入的先验信息难以精确获得. 为了解决上述问题, 本文提出了一种全局到局部模式的大场景图像细粒度人群社交分组框架, 使用静态图像中隐含的平面信息与空间信息完成分组需求, 与现有的静态分组方法相比本工作达到了最佳的分组效果.

3 本文方法

社交分组工作力求回答 “在这个场景中, 谁与谁是有关系的?” 这一问题. 传统的人群分组方法是在受限制的场景条件下开展的, 同时对先验信息的准确性有很高的要求. 本工作面向不受限制的场景, 同时每个场景中都包含大量的对象. 现有的静态图像分组工作 [6, 8, 9, 30~34] 都将个体间距离作为判断两人是否属于同组的主要依据, 这些工作都使用距离约束对图像中的个体组合进行划分, 超过距离阈值的个体将不会与目标个体进行分组判断. 本文将社交互动定义为: 相互认识的个体间共同行走、进行交谈等普通日常行为. 按照社会常识, 具有社交互动的个体间的距离存在范围性. 在人数较多的交互群组中, 局部个体组合的距离、交互动作信息又高度相似, 将全局群组划分为局部群组既能减少复杂度又不影响全局结果, 因此本工作选用由两个人组成的社交对作为局部群组的划分策略.

图 2 是本工作提出的大场景图像细粒度人群社交分组框架. 十亿级像素图像分块后的子图将作为检测网络的输入, 首先通过目标检测网络检测图中个体所处位置, 再对每张图像中被检测到的个体进行全局建图, 全局图中满足距离约束条件的个体组合组成社交对, 进一步操作社交对图像获得对应的掩膜图像与深度图像, 两类图像将作为社交分组网络的输入, 社交分组网络会对两类图像中的特征

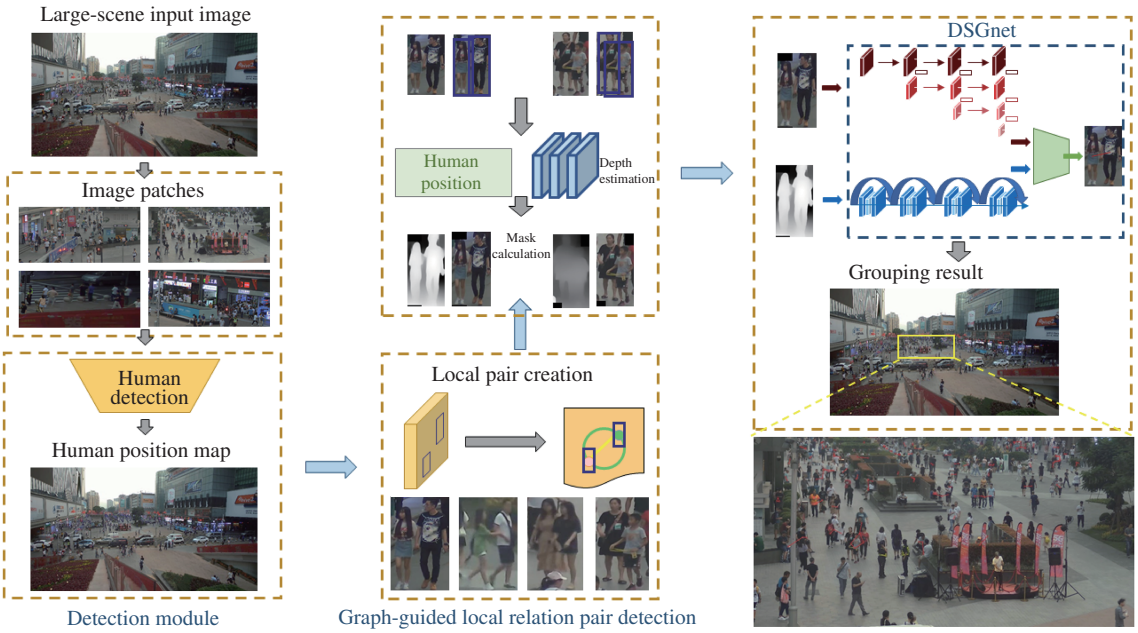

图 2 (网络版彩图) 大场景图像细粒度人群社交分组框架

Figure 2 (Color online) Fine-grained crowd social grouping framework for large scenes

进行提取与融合, 最终获得对应该社交对的分组结果. 下文将详细介绍每部分的具体细节.

3.1 目标检测模块

目标检测是一项基础的计算机视觉任务, 检测技术进步的同时也带动着视觉领域的发展, 社交分组任务的先导需求就是对图像中人的准确定位. 基于深度学习的检测算法是目前的主流方法, 其中按照检测流程又可分为两类: 两阶段检测法和一阶段检测法, 前者的检测流程为由粗到细, 后者则为一步完成. 现有的目标检测技术对输入图像具有像素限制, 不能直接处理十亿级像素的大场景图像, 故本工作采用对大场景图像原图进行分块简化的策略, 以 1024×2048 像素为单位对原图进行分块, 并在图像块与块之间定义 500 像素值的重叠区域以减少个体的部分缺失问题. 对所有大场景图像块的检测结果进行整合, 获得包含大场景图像中每个个体的位置图, 该位置图整体将作为后续模块的输入. 本工作选择 FasterRCNN [35] 作为大场景图像块的基础检测框架, FasterRCNN 是一种将特征提取、区域建议、感兴趣区域池化 (regions of interest pooling, Roi-Pooling)、非极大值抑制 (non-maximum suppression, NMS) 等集成到一起的端到端学习框架, 本工作面向复杂场景, 需要检测出场景中不同大小粒度的个体, 因此本工作选择基于两阶段检测的 FasterRCNN 作为本工作的检测模块. 在其中的区域建议模块, 本工作对经过卷积后提取到的大场景特征图的每个像素以不同长宽比生成 9 个锚框作为初选判定区域进行前景/背景判定, 并生成对应每个检测框的回归偏移, 在本工作中前景只包括人. 生成提议区域后, 提议区域的特征图将由 Roi-Pooling 分支去进一步执行分类和回归操作, 经过区域建议与 Roi-Pooling 操作后, 提议区域特征图将被输入到后续全连接层判定目标类别并给出相关分数. 最终网络的输出是经过非极大值抑制的, 即对于同一个体周围与自身检测框交并比 (iou) 超过一定阈值的冗余检测框将会被舍弃. 交并比的计算和非极大值抑制过程的公式如下:

$$\mathrm{iou}(b_i, b_j) = \frac{\mathrm{Area}(b_i \cap b_j)}{\mathrm{Area}(b_i \cup b_j)}, \tag{1}$$

$$s_i = \begin{cases} s_i, & \text{iou}(b_i, b_j) < N_t, \\ 0, & \text{iou}(b_i, b_j) \geqslant N_t, \end{cases} \tag{2}$$

其中 b_i, b_j 代表两个不同的检测框, s_i, s_j 代表检测框各自对应的分数, N_t 是非极大值抑制的阈值. 但本工作的输入是人群密集的大场景图像, 使用传统的 NMS 算法将会出现目标检测框被相邻且分数高于自身的检测框抑制的情况, 在人群密集的街道上该问题尤其严重, 为了解决该问题, 第一个尝试是将非极大值抑制的阈值 N_t 从 0.7 降低到 0.3, 虽然结果中遮挡情况下的非冗余检测框被抑制的情况有所下降, 但会造成假阳性结果比例上升的问题, 将影响社交分组网络的分组效果, 最终本工作参考 Bodla 等 [36] 的方法使用 Soft-NMS 算法进行优化. 对于交并比大于阈值的检测框, Soft-NMS 不会像 NMS 一样直接将对比检测框分数置 0 进行抑制, 而是使用相关策略降低其分数, 避免真阳性检测框被错误抑制的情况. 这项改进既增加了密集人群场景检测结果的鲁棒性又减少了最终假阳性结果的数量. 本工作使用线性函数降分策略的 Soft-NMS 算法, 公式如下:

$$s_i = \begin{cases} s_i, & \text{iou}(b_i, b_j) < N_t, \\ s_i \times (1\text{-iou}(b_i, b_j)), & \text{iou}(b_i, b_j) \geqslant N_t, \end{cases} \tag{3}$$

目标检测模块的输出是每个被检测到个体在图像中的位置与该个体属于前景 (人) 的概率, 本工作设定预测概率值超过 0.7 的结果进行输出, 即检测模块认为该检测框范围内的图像是人的概率为 70% 以上的结果作为最终检测模块的输出结果.

3.2 图引导的局部社交对检测

为了降低复杂度同时提高算法的灵活性, 本工作采用图引导的局部社交对检测方式将全局分组简化为局部分组. 具体来说, 首先根据目标检测模块获得的个体位置图进行全局建图, 全局图定义为 $G(\mathcal{V}, E)$, 其中 $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ 表示图中的每个结点 (人), 连接两点 (i, j) 的边的权重为 $E_{i,j}$. 与以往工作不同, 本工作定义了一个全新的动态距离约束函数 $F(\mathcal{V}, \mathcal{B}, G)$, 其中, $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$ 是图中每个结点所对应的检测框位置, G 代表全部个体位置信息的初始全局图, 该函数以动态的值作为每个人自身的范围域半径寻找相邻目标, 增加了距离约束方法的灵活性. 本工作的距离约束过程具体如下: 对于每个结点 v_i, 先计算其自身检测框的宽度, 再以检测框中心为圆心、检测框宽度为检测半径设定圆形范围域, 检测框中心在 v_i 结点范围域内的其余结点被设定为与结点 v_i 相匹配的行人组合, 全部匹配个体构成结点 v_i 的匹配集, 结点 v_i 的匹配集 $\mathcal{P}_i$ 定义如下:

$$\mathcal{P}_i = \{v_i, v_k \mid g \in \mathcal{V}, \text{dis}(c_g, c_i) < w_i\}, \tag{4}$$

其中, c_i 代表个体 i 的检测框中心, w_i 代表个体 i 检测框的宽度, dis() 表示欧氏距离函数. 选用动态值进行距离约束而不是使用固定值进行距离约束的原因是: 对于每张输入图像, 距离相机中心近的目标面积较距离相机中心远的目标面积大, 而面积又决定了检测框的宽度与高度, 同时相对深度差大的两个目标往往不存在社会交互. 最初本工作设置目标检测框宽度的两倍为范围域圆的半径, 该设置导致后续分组网络中存在大量冗余的检测, 经过多次实验, 当范围域圆的半径与检测框宽度相同时达到了最好的分组效果与最快的社交对检测速度.

本工作的目标是处理现实世界中任意复杂场景下的分组问题, 当场景中包含婴儿和儿童时仅采用上述策略会出现社交对不全的情况. 为了使方法更加鲁棒, 我们在上述策略中添加了面积比较项, 对于结点 v_i 与自身检测框 b_i, 遍历找出至多 K (本工作中 $K = 5$) 个距离个体 i 检测框中心 c_i 欧氏距离

小的对比框 $b_{1\sim k}$ 中面积最小的检测框 b_k 与 b_i 进行面积比较, 如果 b_i 的面积仍小于 b_k 的面积, 那么就认为 b_i 位置处的目标大概率为儿童, 同时将 b_i 所对应范围域圆的半径设定为 b_k 检测框的宽度 w_k.

经过上述操作后图像中每个结点都会得到对于自身的匹配集 $\mathcal{P}_i$, $\mathcal{P}_i$ 中的每个结点都会与自身共同组成分组检测网络的先验社交对.

3.3 社交分组网络

近年来, 深度学习技术已经被广泛应用于处理计算机视觉问题上, 本工作首次尝试使用神经网络去处理大场景图像上的人群分组任务, 提出一个能够面向任意分辨率静态图像进行社交分组任务的深度网络 DSGnet. 本模块将详细介绍深度社交分组网络的组成与相关细节, 主要包括: 网络结构、数据准备两部分内容. DSGnet 的训练细节详见补充材料的 C 和 D 节.

3.3.1 网络结构

本小节将介绍本工作所设计的深度社交分组网络 DSGnet 的具体结构与细节. 在数学领域, 如果在某一变化的过程中, 变量 x 在一定范围内都有一个确定的 y 与其对应且满足 $F(x,y)=0$, 那么该 F 函数就被称为 x 与 y 间映射关系的隐函数. 对于社交对的组间关系判断过程, 意在寻找一种可以将社交对的信息特征与组间关系相对应的映射方法. 从该目的出发, 本工作将 DSGnet 定义为一种特殊的社交对二值隐函数, 其输入为社交对的两种信息图: 掩膜图像 F_M 和相对深度图像 F_D, 输出为社交对存在/不存在组内交互的概率值, 具体表示为

$$\mathrm{DSGnet}(F_M,F_D)=\boldsymbol{r}: F_M\in\mathbb{R}^3,\quad F_D\in\mathbb{R}^2,\quad \boldsymbol{r}\in\mathbb{R}^2. \tag{5}$$

该表示的关键意图是: 对每个社交对, DSGnet 通过输入的两种信息图完成社交对中是否存在组内交互的自动判断. 由此可见, 社交分组网络对输入特征的映射能力将决定分组交互预测性能的表现.

本文将个体间的分组过程局域化为每个匹配社交对是否存在交互的分类问题. 由于掩膜图像 F_M 中包含丰富的语义信息, 复杂场景中掩膜图像内的个体尺度又可能存在明显差异, 因此处理掩膜图像的网络结构应当具有优秀的多尺度融合与特征提取能力. 在现有的方法中, 大多采用对称的网络结构进行多尺度特征的融合操作, 即对卷积后的输入特征图先进行下采样再进行上采样以恢复高分辨率特征图, 并将下采样操作获得的低分辨率特征图与高分辨率特征图进行跳层连接或使用级联金字塔达到特征多尺度融合的效果. 在人体 2D 关节点估计问题中, 上述多尺度融合的方法被广泛使用以满足估计不同尺度下人体关节点位置的需要, 但这种对特征进行先降后升的特征维度变换过程无法还原初始的高分辨率特征, 影响高分辨特征的表征能力, 故本工作参考 HRNet [37] 的设计思路, 设计了一种多尺度、多支路的特征提取网络, 如图 3 所示, 网络中不同分辨率的支路并行连接, 在保证不同支路中特征提取过程相互独立的基础上, 进行多分辨率支路间的特征融合, 经过多次融合后的多尺度特征图将被展开并输入到全连接网络中进行特征分布式表示.

复杂场景图像中存在大量满足距离约束但具有明显视差的假相邻个体, 为了正确预测上述情况的分组结果, 本工作将相对深度图像 F_D 作为社交对的第 2 种先验特征进行输入. 由于 F_D 与 F_M 的通道数量存在差异, 共同输入会使网络更关注通道数量多的 F_M 中的特征信息而忽略个体间的相对深度信息, 导致网络对假相邻个体的预测性能下降. 最终, 本工作添加额外支路进行相对深度特征的提取工作. 为了保证提取特征的全面性, 本支路选择 Resnet 残差网络作为主体结构, 该结构解决了网络加深所造成的退化问题, 在保证特征提取全面性的同时又通过层间的跳跃连接避免了梯度消失问题的出现. 与处理掩膜图像 F_M 的支路相同, 在深度特征提取支路中同样使用三层全连接网络将提取到的特

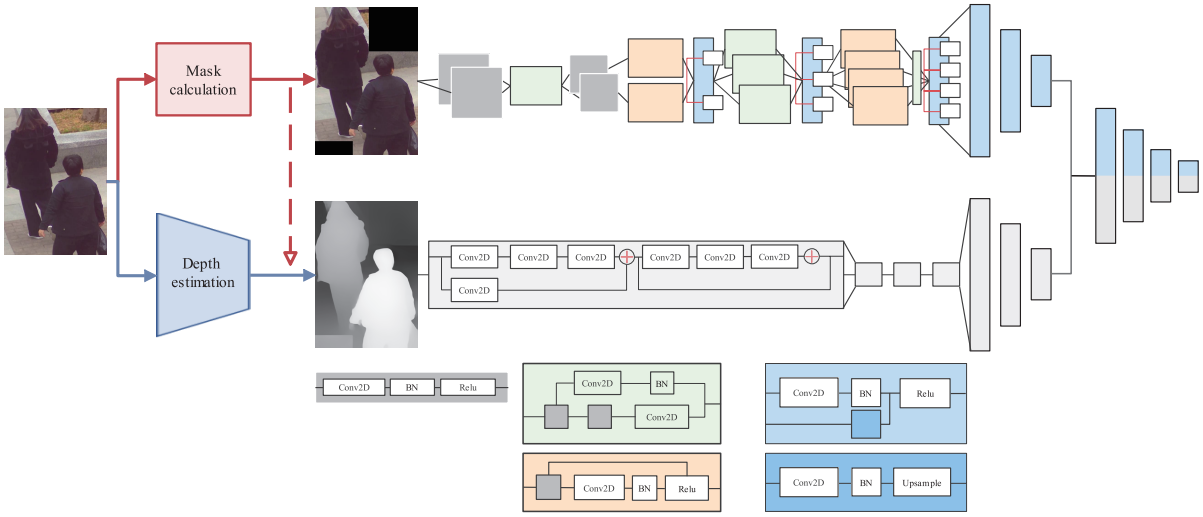

图 3 (网络版彩图) 深度社交分组网络结构图

Figure 3 (Color online) The structure of deep social grouping network (DSGnet)

征图进行分布式表示.

掩膜图像支路与相对深度图像支路的输出是经过各自网络提取后的特征向量, 要使网络自动权衡两种特征对分组结果的影响程度, 需要进一步对包含在两个特征向量内的特征进行选取与整合, 并根据整合后的特征进行组内交互判断. 为了确保特征组合的多样性与有效性, 本工作使用自适应嵌入特征下降模块对两个支路输出的特征向量进行升降维的整合操作. 具体表示为: 将两个支路输出的 128 维向量进行顺序拼接提升维度, 对拼接后的社交对特征向量进行四次特征维度下降的过程, 多次下降过程既避免了一次大幅下降所造成的特征丢失问题, 又提高了网络组合特征的自适应能力. 最终, 社交分组网络的输出是一个二维向量, 向量中的值分别代表该社交对中存在交互/不存在交互的概率值.

3.3.2 数据准备

数据是决定神经网络性能优劣的主要因素之一, 而数据的形式又取决于网络要解决的问题. 本工作尝试使用位置纹理、相对深度两种形式的先验数据对人群分组问题进行探究, 如图 4 所示, 最终采用包含位置纹理特征的掩膜图像与包含相对深度特征的相对深度图像作为网络的输入.

霍夫投票等方法 [8] 解决分组问题的关键是对 F 形中 o- 空间的估算. 组间个体所形成的 p- 空间内根据每个人所处位置又可以组合成多种站位情况. 本工作保留这种空间构成思想并将其表现在数据中, 意在让网络去学习同组个体的不同位置组合形式. 如图 5 所示, 图 5(a) 是 Cristani 等 [8] 提出的 F 形组成, 图 5(b) 首行是 p- 空间中同组个体位置的几种组合形式 [9], 第二行是本工作提出的带有隐含 o- 空间 (椭圆范围) 的掩膜图像数据.

位置纹理特征. 掩膜图像的获取方法如下: 对第 3.2 小节中全局图的每个结点 (人) v_i 和其匹配集 $\mathcal{P}_i$ 进行社交对匹配组合, 在输入原图中截取包含个体 v_i 与 v_j $(v_i \in \mathcal{P}_i)$ 检测框 b_i, b_j 的最小长方形部分, 并以 b_i 与 b_j 作为蒙板进行图像掩膜. 使用 b_i 与 b_j 检测框而不选择边界清晰的实例分割结果作为蒙板的原因是意在完整保留图像中的语义信息, 比如背景位置信息、个体纹理信息等, 语义信息对分组情况的判断能够起到正向的效果. 这种既包含丰富语义信息又能隐含表示社交对 o- 空间的掩膜图像增加了本工作方法的先进性并提高了分组结果的准确性.

图 4 (网络版彩图) 三组数据样例. 每组从左至右依次为社交对原图、掩膜图像、相对深度图

Figure 4 (Color online) Triple pack data samples. Each group from left to right contains the original image, the mask image, and the relative depth image

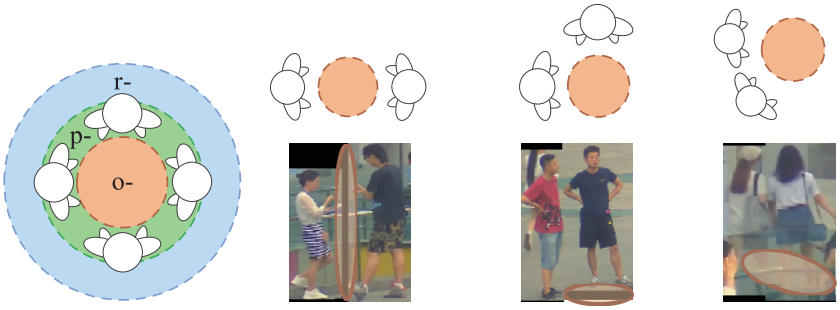

图 5 (网络版彩图) F 形组成结构 [9] (a) 与本文的 "o- 空间" 表示方法 (b)

Figure 5 (Color online) F-formation (a) and our o-space (b)

相对深度特征. 在人群分组任务中, 街道、车站等人群密集的场景图像中往往包含大量空间不相邻但像平面相邻的假相邻个体, 故可通过深度信息对图像中的相邻个体与假相邻个体进行区分, 通过深度去反映个体在场景中的三维位置, 从而利用空间信息对分组情况进行辅助判断, 提高对假相邻个体与复杂人群情况的处理能力. 但直接对大场景图像进行深度估计是困难的. 另一方面, 要创造出在各种情况下都可以有效估计大场景图像深度的学习模型, 需要模型的训练数据具有一致性且数据能够满足现实场景的多样性. 但不同传感器对同一场景采集的深度结果是不同的, 单一数据集往往又不能满足数据的多样性, 故还没有一种对任意大场景都鲁棒的深度估计方法. 为了解决上述问题, 本文将对大场景图像绝对深度的估计与对个体间深度差的计算简化为使用 Lasinger 等 [38] 提出的深度估计方法估计社交对图像的相对深度过程.

本文中, 深度估计网络的输入是包含个体 v_i 与 v_j $(v_i \in \mathcal{P}_i)$ 检测框 b_i, b_j 的最小长方形部分, 对网络输出的结果同样以检测框 b_i 与 b_j 作为蒙板进行图像掩膜, 得到个体 v_i 与 v_j 的相对深度图 F_D. 相对深度信息的添加赋予了网络辨别个体相对空间位置关系的能力, 给分组性能带来了极大的提高.

4 实验结果

图 6 展示了 DSGnet 社交分组网络在 PANDA [16] 数据集上使用真实行人位置条件下的人群分组结果. 其中网络判断出属于同小组的个体间使用红色线段连接, 用深蓝色框标记单独个体, 右侧图像是对左侧黄色框区域的放大显示.

本节下面的部分将对所提出的社交分组网络进行详细的实验验证. 第 4.1 小节将介绍验证时采用的数据集与评价指标, 第 4.2 小节将展示与现有方法进行定性与定量比较的结果, 第 4.3 小节展示社交

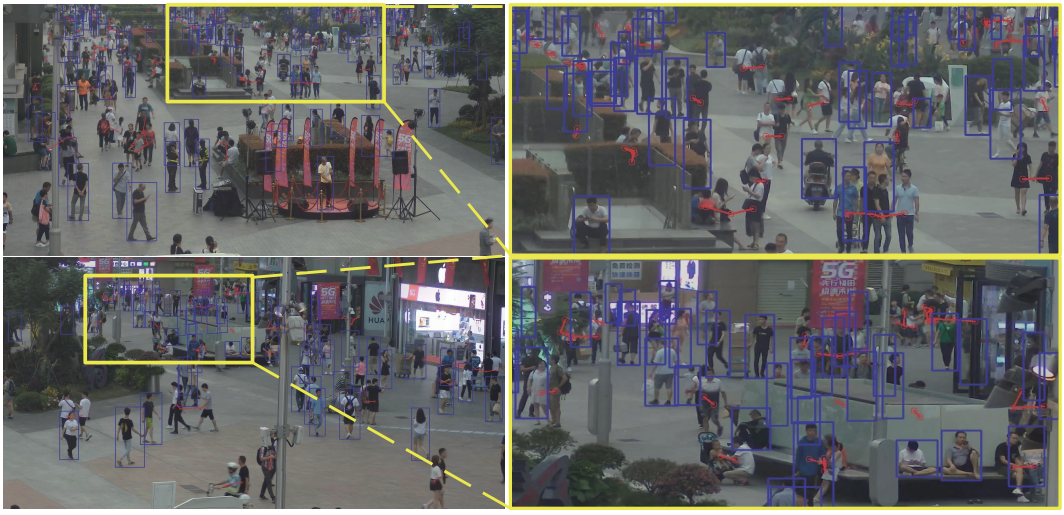

图 6 (网络版彩图) 社交分组网络在密集人群场景下的分组结果

Figure 6 (Color online) The grouping results of deep social grouping network in dense crowd scenes

分组网络不同结构间消融实验的结果, 并在补充材料 F 节中提供了对分组效果的进一步展示与讨论.

4.1 实验设置

数据集. 本工作使用两个公共数据集对社交分组网络的效果进行验证.

(1) PANDA 数据集 [16]. 使用该数据集中 3 个具有大量密集对象的场景 (XiliStreet1, XiliStreet2, HuaQiangBei) 中共计 702 张十亿级像素图像进行测试, 场景中共包含 3364 个人, 共分成 513 个小组.

(2) CoffeeBreak 数据集 [8]. 数据集主要由人们喝咖啡休息的社交场景组成, 共有 119 张图像, 每张图像中最多包含 14 个人, 所组成的小组平均 2 到 3 人一组, 每张图像的分辨率是 1440×1080. 该数据集是 PANDA 数据集出现前最接近真实场景的标准数据集.

评价指标. 社交分组网络的作用是判断社交对中的个体是否存在组内交互, 准确率指标更能直观反映出网络对正例与反例的整体判断能力, 网络对同组个体的正确判断比例又能证明本方法的有效性. 因此本工作使用准确率 (Accuracy) 和召回率 (Recall) 对社交分组网络进行效果验证, 具体的计算公式如下:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

其中, 真阳性 TP 为属于同组且判断正确, 假阴性 FP 为不属于同组但判断错误, FN 为不同组的错误判断, TN 为不同组的正确判断.

4.2 比较分析

本小节将展示本工作提出的大场景图像细粒度人群社交分组框架与现有的领先方法在 CoffeeBreak 数据集上进行定性与定量比较的结果. 比较的方法分别是: 经典的静态分组方法霍夫投票法 HVFF [8] 和具有先进性能的 F 形图切割算法 GCFF [9]. 由于这两种方法需要很多信息作为输入且无法处理十亿级像素图像, 不适用于 PANDA 大场景图像数据集, 因此我们使用以上两种工作的代码在

表 1 CoffeeBreak 数据集上与现有方法定量比较的结果

Table 1 Quantitative comparison with the state-of-the-art methods

Method	CoffeeBreak SEQ1			CoffeeBreak SEQ2		
	Accuracy	Recall	Average	Accuracy	Recall	Average
HVFF [8]	0.6473	0.5574	0.6024	0.6821	0.5363	0.6092
GCFF [9]	0.8170	0.7822	0.7996	0.8804	0.8538	0.8656
Ours	**0.9034**	**0.9510**	**0.9272**	**0.9173**	**0.8630**	**0.8901**

图 7 (网络版彩图) CoffeeBreak 数据集上与现有方法定性比较的结果

Figure 7 (Color online) Qualitative comparison with the state-of-the-art methods

CoffeeBreak 数据集上与本工作的整体框架进行对比验证, 使用具有不同人群密度的两个序列 SEQ1, SEQ2 进行分别验证并展示相关结果.

表 1 展示了本工作提出的大场景图像细粒度人群社交分组框架与霍夫投票法、F 形图切割算法在 CoffeeBreak 数据集上的定量比较结果. CoffeeBreak 数据集中大部分个体间不存在交互, 所分成的组数较少, 因此结果中准确率普遍高于召回率. 总的相比, 本工作提出的分组框架优于 HVFF 与 GCFF, 达到了最优的人群分组性能. 对本文方法从准确率结果进行分析, SEQ1 场景中的人群密度较 SEQ2 大, SEQ1 场景中普遍存在个体间相互遮挡的情况, 由于检测技术的瓶颈或深度混淆带来的误差, 社交分组网络得出的假阳性结果会影响判断的准确性能. 从召回率结果来看, HVFF 与 GCFF 方法对存在交互的同组对象的判断能力较弱, 而本文的方法则可以全面地记忆同组个体间的行为模式从而达到准确的分组判断.

图 7 展示了不同方法在 CoffeeBreak 数据集上的可视化分组结果, 其中前两列为 HVFF 与 GCFF 的可视化的结果, 最后一列为本工作的输出结果. 为了明显地进行分组效果区分, 此组图中只对存在小组关系的个体进行连接, 没有被连接的个体将被视为单独个体.

4.3 消融分析

本小节将验证社交分组网络中不同支路的有效性, 下文将分组网络中处理掩膜图像的支路称为掩

表 2　社交分组网络定量消融实验结果

Table 2　Quantitative ablation study of social grouping network

	XiliStreet1			XiliStreet2			HuaQiangBei		
Structure	Accuracy	Recall	Average	Accuracy	Recall	Average	Accuracy	Recall	Average
Mask branch	**0.7060**	0.4566	0.5813	0.6315	0.6368	0.6342	**0.7086**	0.5625	0.6356
Depth branch	0.6272	**0.6770**	**0.6521**	0.6399	0.7176	0.6788	0.6013	**0.7296**	0.6655
Integrated network	0.6916	0.5895	0.6401	**0.6571**	**0.7424**	**0.6998**	0.6652	0.6950	**0.6801**

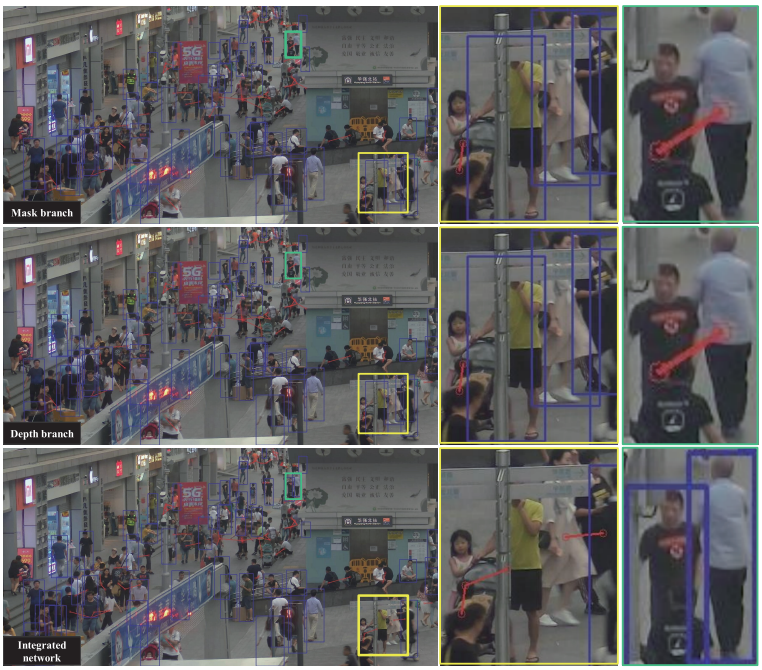

图 8　(网络版彩图) 不同深度社交分组网络变体的定性结果

Figure 8　(Color online) Visual results of different variants of deep social grouping network

膜支路, 将处理相对深度图像的支路称为深度支路. 以每张图像与个体真实位置作为输入, 将两个支路分别作为网络的整体模型进行训练与验证.

掩膜支路. 当只根据掩膜图像 F_M 进行分组判断时, 将掩膜支路的特征提取结果直接作为社交分组特征融合模块的输入, 同时将融合支路的输入维度由 256 维降到 128 维, 输出维度保持不变.

深度支路. 当只根据相对深度图像 F_D 进行分组判断时, 同样将深度支路的特征提取结果直接作为社交分组特征融合模块的输入, 对特征融合模块的维度改变与掩膜支路操作相同, 进而维护消融实验的一致性原则.

表 2 记录了不同测试场景下 3 种网络的准确率与召回率, 展示了平面信息与空间信息对社交分组网络性能的影响. 本文方法的掩膜支路意在根据社交对中个体是否形成 o- 空间进行分组情况的判断, 不属于同组的社交对图像中往往不存在隐含 o- 空间, 掩膜支路对该类图像的预测性能较好, 但对存在 o- 空间的假相邻社交对情况判断能力较弱, 因此掩膜支路的准确率高于召回率. 方法中深度支路意在使用空间位置信息进行分组情况的判断, 若存在小组关系的个体间相对深度差小, 深度支路可以对其进行准确判断, 根据深度信息又能在一定程度上区分假相邻个体, 但密集场景中存在属于同组但

具有较大深度差的特殊情况, 深度支路对该类情况的判断性能较弱. 由于存在组内交互的社交对呈现出的相对深度信息具有一定的相似性, 网络能够根据学习到的深度相似性先验对存在交互的个体对进行准确的判断, 因此深度支路的召回率高于准确率. 从整体社交分组网络的平均指标结果来看, 掩膜支路与深度支路结合后的社交分组网络综合了两支路各自的优势并达到了一定的组内交互判断水平. 相关的定性比较结果与掩膜图像的作用性分析实验分别在补充材料的 B 和 E 节进行展示.

图 8 展示了相关的定性结果, 其中首行为只利用掩膜支路的分组结果, 第 2 行为只利用深度支路的分组结果, 尾行为完整社交分组网络的分组结果. 使用绿色框与黄色框标记出完整社交分组网络与只利用掩膜支路、深度支路的不同之处. 对结果的不同之处进行分析, 发现完整的社交分组网络可以更有效地利用社交对图像的平面信息与空间信息, 达到两种社交分组支路网络变体所达不到的分组性能.

5 总结

本文针对如何对大场景中数量不定与关系复杂的行人进行分组这一问题, 提出了一种具有全局到局部模式的大场景图像细粒度人群社交分组框架, 并设计了一个可以探究社交对组内关系的端到端网络. 框架中使用成熟的目标检测技术准确地定位个体, 采用全局到局部的匹配映射降低群体关系的复杂程度, 进而设计出一种空域与深度联合引导的端到端多支路网络完成社交分组关系的判断. 总体来说, 本框架以灵活划分人群为目标、以学习交互特征为主体, 进而完成了多尺度目标检测、细粒度人群划分、空域特征与深度特征融合的交互判断, 成为一种新颖且可靠的静态图像人群分组方法. 各种实验表明, 本工作提出的基于深度学习的静态图像人群分组框架实现了大场景数据集上的细粒度人群分组, 且在公共小场景数据集上的性能优于现有方法. 未来我们将引入个体的面部朝向与个体间的交互动作, 改进人群分组问题的解决方法, 以达到更佳的分组性能.

补充材料 A∼E. 本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1 Zhang B, Zhu J, Su H. Toward the third generation of artificial intelligence. Sci Sin Inform, 2020, 50: 1281–1302 [张钹, 朱军, 苏航. 迈向第三代人工智能. 中国科学: 信息科学, 2020, 50: 1281–1302]

2 Guo W, You S S, Gao J Y, et al. Deep relative metric learning for visual tracking. Sci Sin Inform, 2018, 48: 60–78 [郭文, 游思思, 高君宇, 等. 深度相对度量学习的视觉跟踪. 中国科学: 信息科学, 2018, 48: 60–78]

3 Jacques J C S, Braun A, Soldera J, et al. Understanding people motion in video sequences using Voronoi diagrams. Pattern Anal Appl, 2007, 10: 321–332

4 Yu T, Lim S N, Patwardhan K, et al. Monitoring, recognizing and discovering social networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 1462–1469

5 Robertson N M, Reid I D. Automatic reasoning about causal events in surveillance video. EURASIP J Image Video Process, 2011, 2011: 1–19

6 Tran K N, Bedagkar-Gala A, Kakadiaris I A, et al. Social cues in group formation and local interactions for collective activity analysis. In: Proceedings of International Conference on Computer Vision Theory and Applications, Barcelona, 2013. 539–548

7 Vinciarelli A, Pantic M, Bourlard H. Social signal processing: survey of an emerging domain. Image Vision Comput, 2009, 27: 1743–1759

8 Cristani M, Bazzani L, Paggetti G, et al. Social interaction discovery by statistical analysis of F-formations. In: Proceedings of British Machine Vision Conference, 2011. 2: 4

9 Setti F, Russell C, Bassetti C, et al. F-formation detection: individuating free-standing conversational groups in images. Plos One, 2015, 10: e0123783

10 Lerner A, Chrysanthou Y, Lischinski D. Crowds by example. Comput Graph Forum, 2007, 26: 655–664

11 Pellegrini S, Ess A, Schindler K, et al. You'll never walk alone: modeling social behavior for multi-target tracking. In: Proceedings of IEEE International Conference on Computer Vision, Kyoto, 2009. 261–268

12 Ferryman J, Shahrokni A. Pets2009: dataset and challenge. In: Proceedings of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, 2009. 1–6

13 Yuan X, Fang L, Dai Q, et al. Multiscale gigapixel video: a cross resolution image matching and warping approach. In: Proceedings of IEEE International Conference on Computational Photography, Stanford, 2017. 1–9

14 Zhang J, Zhu T, Zhang A, et al. Multiscale-VR: multiscale gigapixel 3D panoramic videography for virtual reality. In: Proceedings of IEEE International Conference on Computational Photography, Saint Louis, 2020. 1–12

15 Yuan X, Ji M, Wu J, et al. A modular hierarchical array camera. Light Sci Appl, 2021, 10: 37

16 Wang X, Zhang X, Zhu Y, et al. PANDA: a gigapixel-level human-centric video dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, 2020. 3268–3278

17 Hongeng S, Nevatia R. Large-scale event detection using semi-hidden Markov models. In: Proceedings of IEEE International Conference on Computer Vision, Nice, 2003. 3: 1455–1462

18 Cheng Z, Qin L, Huang Q, et al. Group activity recognition by Gaussian processes estimation. In: Proceedings of International Conference on Pattern Recognition, Turkey, 2010. 3228–3231

19 Ni B, Yan S, Kassim A. Recognizing human group activities with localized causalities. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 1470–1477

20 Amer M R, Lei P, Todorovic S. HIRF: hierarchical random field for collective activity recognition in videos. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 572–585

21 Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition. In: Proceedings of European Conference on Computer Vision, Florence, 2012. 215–230

22 Choi W, Savarese S. Understanding collective activities of people from videos. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 1242–1257

23 Choi W, Shahid K, Savarese S. Learning context for collective activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, 2011. 3273–3280

24 Lan T, Sigal L, Mori G. Social roles in hierarchical models for human activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 1354–1361

25 Lan T, Wang Y, Yang W L, et al. Discriminative latent models for recognizing contextual group activities. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 1549–1562

26 Shu T, Xie D, Rothrock B, et al. Joint inference of groups, events and human roles in aerial videos. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 4576–4584

27 Amer M R, Xie D, Zhao M, et al. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: Proceedings of European Conference on Computer Vision, Florence, 2012. 187–200

28 Ibrahim M S, Muralidharan S, Deng Z, et al. A hierarchical deep temporal model for group activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1971–1980

29 Wu J, Wang L, Wang L, et al. Learning actor relation graphs for group activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 9964–9974

30 Bazzani L, Cristani M, Tosato D, et al. Social interactions by visual focus of attention in a three-dimensional environment. Expert Syst, 2013, 30: 115–127

31 Hung H, Krose B. Detecting F-formations as dominant sets. In: Proceedings of International Conference on Multimodal Interfaces, New York, 2011. 231–238

32 Setti F, Hung H, Cristani M. Group detection in still images by F-formation modeling: a comparative study. In: Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services, Paris, 2013. 1–4

33 Vascon S, Mequanint E Z, Cristani M, et al. A game-theoretic probabilistic approach for detecting conversational groups. In: Proceedings of Asian Conference on Computer Vision, Singapore, 2014. 658–675

34 Setti F, Lanz O, Ferrario R, et al. Multi-scale F-formation discovery for group detection. In: Proceedings of IEEE International Conference on Image Processing, Melbourne, 2013. 3547–3551
35 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149
36 Bodla N, Singh B, Chellappa R, et al. Improving object detection with one line of code. 2017. ArXiv:1704.04503
37 Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 5693–5703
38 Lasinger K, Ranftl R, Schindler K, et al. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. 2019. ArXiv:1907.01341

Deep social grouping network for large scenes with multiple subjects

Kun LI[1], Wanpeng LI[1], Xiaokun SUN[1] & Lu FANG[2*]

1. *College of Intelligence and Computing, Tianjin University, Tianjin* 300350*, China*;
2. *Department of Electronic Engineering, Tsinghua University, Beijing* 100084*, China*
* Corresponding author. E-mail: fanglu@tsinghua.edu.cn

Abstract In computer vision, more attention has been paid to group analysis, and the group detection in images becomes a key technology of human analysis on groups. The existing social grouping methods only focus on small scenes with fixed number of persons and cannot deal with large scene images in the real world. This paper proposes the first fine-grained social grouping framework for gigapixel large scene images based on deep learning, which consists of a graph-guided global-to-local partition strategy and a deep grouping network that learns an implicit respresentation for social pairs. The framework has achieved accurate grouping on large scene images. Our method is also applicable to small scene images, and has outperformed the existing methods. The relevant code and the training dataset will be released soon.

Keywords group, large-scene image, deep learning, social grouping, graph-guided

Kun LI was born in 1983. She received her B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and her master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She visited École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. She is currently an associate professor at the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests include dynamic scene 3D reconstruction and image/video processing.

Wanpeng LI was born in 1997. He received his B.S. degree in computer science and technology from Northeast Agricultural University, Harbin, China in 2019. Currently, he is a master candidate in Tianjin University. His research interest lies in computer vision.

Xiaokun SUN was born in 1999. He received his B.S. degree in communication engineering from Hefei University of Technology in 2021. Currently, he is a master candidate in Tianjin University. His research interest lies in computer vision.

Lu FANG was born in 1986. She received her Ph.D. degree from the Hong Kong University of Science and Technology in 2011, and B.E. from University of Science and Technology of China in 2007. She is currently an associate professor in the Department of Electronic Engineering, Tsinghua University. Her research interests include computational photography and visual intelligence.