姿态特征与深度特征在图像动作识别中的混合应用

钱银中 [1] 沈一帆 [2, 3]

摘 要 人体姿态是动作识别的重要语义线索, 而 CNN 能够从图像中提取有很强判别能力的深度特征, 本文从图像局部区域提取姿态特征, 从整体图像中提取深度特征, 探索两者在动作识别中的互补作用. 首先介绍了一种姿态表示方法, 每个肢体部件的姿态由描述该部件姿态的一组 Poselet 检测得分表示. 为了抑制检测错误, 设计了基于部件的模型作为检测上下文. 为了从数量有限的数据集中训练 CNN 网络, 本文使用了预训练和精细调节的方法. 在两个数据集中的实验表明, 本文介绍的姿态特征与深度特征混合使用, 动作识别性能得到了极大提升.

关键词 动作识别, 姿态特征, poselet, 深度特征

引用格式 钱银中, 沈一帆. 姿态特征与深度特征在图像动作识别中的混合应用. 自动化学报, 2019, **45**(3): 626−636

DOI 10.16383/j.aas.2018.c170294

Hybrid of Pose Feature and Depth Feature for Action Recognition in Static Image

QIAN Yin-Zhong[1] SHEN Yi-Fan[2, 3]

Abstract Body pose is an important semantic cue for action recognition, and CNN can extract strong discriminative depth feature. This paper extracts pose feature from local image patches and gets depth feature from holistic image, then exploits their complementary relationship in action recognition. A pose representation is introduced, in which pose of a body part is represented by a collection of poselets which describe its pose variability. To suppress detection ambiguity, part-based model is designed as the context of detection for each poselet. CNN is trained through pre-training and fine tuning on the data set with very limited images. Empirical results demonstrate aggressive performance improvement by concatenating pose feature and depth feature.

Key words Action recognition, pose feature, poselet, depth feature

Citation Qian Yin-Zhong, Shen Yi-Fan. Hybrid of pose feature and depth feature for action recognition in static image. *Acta Automatica Sinica*, 2019, **45**(3): 626−636

动作识别是计算机视觉中的研究热点. 一个特定动作通常由一连串人体的肢体运动组成, 因此长期以来研究者从视频研究动作识别[1−5]. 生活经验告诉我们, 人类具有从图像中识别动作的能力. 例如, 图 6 中的每幅图像都有一个人在执行一个动作, 我们一眼就能看出这个人在干什么. 随着图像分类和物体检测技术的进步, 近十年来出现了很多从静止图像识别动作的研究工作[6].

收稿日期 2017-06-01 录用日期 2018-01-20
Manuscript received June 1, 2017; accepted January 20, 2018
江苏高校品牌专业建设工程资助项目 (PPZY2015A090), 常州信息职业技术学院自然科学项目 (CXZK201803Z) 资助
Supported by Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (PPZY2015A090) and Natural Science Project of Changzhou College of Information Technology (CXZK201803Z)
本文责任编委 赖剑煌
Recommended by Associate Editor LAI Jian-Huang
1. 常州信息职业技术学院软件学院 常州 213164 2. 复旦大学计算机科学技术学院 上海 200433 3. 复旦大学上海市智能信息处理重点实验室 上海 200433
1. School of Software, Changzhou College of Information Technology, Changzhou 213164 2. School of Computer Science, Fudan University, Shanghai 200433 3. Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433

从静止图像中识别动作具有广阔的应用前景. 首先, 可以促进视频识别动作的研究. 视频由一组按照时间顺序排列的帧图像组成, 如果能够设计新的算法, 从少量图像中识别动作, 就可以减少冗余帧, 提高计算效率. 这种算法还可以与视频中的时间维度结合, 提出新的算法, 进一步提高动作识别性能. 其次, 有助于图像自动标注. 由于 Internet 的兴起, 人们能够从网络上获得海量图像数据. 如果计算机能够识别图像中的动作, 就可以把像素信息作为输入, 自动标注这些图像的动作类别. 另外, 还有助于动作图像的检索和管理. 目前搜索引擎检索图像主要还是依赖图像周围的文字, 如果能够从图像中提取动作信息, 无疑将更准确、更方便地检索和收集动作图像.

1 相关研究

仔细分析一下可以发现, 人类通过图像中的线索识别动作, 这些线索包括: 人体的姿态、与人交互的物体、以及背景等. 研究静止图像中的动作识别有两种思路. 一种思路与人类识别动作类似, 先从图像中检测人体姿态或与人交互的物体等高层语义信

息, 然后根据机器学习算法判定动作类别[7−9]. 另一种思路是提取图像的整体特征, 把动作识别当作图像分类问题[10] 处理.

人体姿态是动作识别中最重要的语义线索. 人体很容易被检测到, 检测结果通常用一个矩形窗口表示人体的位置. Ikizler-Cinbis 等[11] 使用 NMF (Non-negative matrix factorization) 方法对训练集中的人体窗口进行姿态聚类, 有相似姿态的图像被划分到同一个聚类中, 然后对每个动作类别训练分类器. 除了矩形窗口外, 还有的方法使用人体轮廓表示人体. Wang 等[12] 用 Canny 边缘检测模板得到的一组边界点作为人体的轮廓, 然后对轮廓特征聚类, 把图像标注到不同的动作类别.

肢体整体难以反映姿态特征的细节, 而人体的各部件, 例如胳膊、腿、躯干等无法变形的原子部件的组合结构有丰富的动作线索. 如果能够识别这些部件以及它们的相对位置, 就能够更准确地识别动作类别. 通过原子部件的组合配置描述人体姿态的经典算法是基于部件的 Pictorial 模型[13−14]. 遗憾的是, 这方面的研究还很不成熟, 通过这些原子部件检测人体姿态只能在一些特定的数据集上做实验, 检测结果无法用于识别动作. 为了避免原子部件在检测中显著性不足的缺点, 研究者提出了 Poselet[15−16]. Poselet 是由训练集中同一肢体部件具有相同关键点配置的图像窗口实例训练而来的检测模板. 一个动作通常由一连串变化的姿态组成, 而一幅图像只能抓住其中一个快照. 即使是同一个姿态, 从不同角度拍摄, 得到的图像也会千差万别. 除此以外, 还要处理人的高矮胖瘦以及不同服饰引起的的外貌特征差异. 因此, 研究者需要定义一组 Poselet 描述同一部件的不同姿态. 由于 Poselet 的粒度可以从最小的原子部件到整个人体, 研究者可以根据检测中的显著性大小、是否包含姿态和动作语义信息等要求定义不同粒度和种类的 Poselet. Yang 等[17] 对左胳膊、右胳膊、肢体上部和腿四个部件分别训练一组反映其姿态变化的 Poselet, 通过一个星形的基于部件的模型把这些部件连接起来. 在此基础上, Wang 等[18] 提出了分层的基于部件的模型, 该模型是由 20 个部件节点组成的有环网状结构.

本文用 Poselet 向量空间中的坐标表示肢体部件的姿态, 然后根据这一中间层特征识别动作. 与这一特征类似的有视频动作识别研究中的局部运动部件[19], 这一中间层特征是由稠密轨迹[20] 在局部时空域上聚集而成, 描述诸如胳膊或腿等语义区域内的运动特征. 该方法先从训练集中提取稠密轨迹, 然后在每一视频中根据时间重叠、空间距离以及速度定义的距离对稠密轨迹聚集分组. 如果把单个稠密轨迹看作一个可视化单词, 局部运动部件就是由一组稠密轨迹组成的可视化语句. 与这一方法不同的是, 本文可视化单词是 Poselet, Poselet 是由二维图像窗口训练来的检测模板, 由同一部件的 Poselet 描述该肢体部件的姿态变化, 而稠密轨迹是从三维时空域中提取的运动特征, 局部运动部件由一组相似的稠密轨迹组成, 描述部件的运动特征.

随着大规模训练数据集的出现, CNN 模型逐渐引起研究者的关注. 在 ILSVRC2012 图像分类挑战赛中, Krizhevsky 等[21] 提出的深度网络的性能超过了所有的已有分类算法. 深度特征是从整体图像中依次经过多层架构抽象而来. 尽管深度特征有很强的判别能力, 却无法区分哪些特征各自反映了姿态、交互的物体、属性以及背景等与动作有关的语义线索.

在动作识别领域, 利用两个通道的互补线索提高识别能力有很多研究. 文献 [17−18] 从图像局部区域中提取姿态特征, 从图像全局中提取整体特征, 目标函数由局部特征构成的基于部件的模型加上整体特征组成, 通过模型与图像的匹配判别动作类别. 在视频处理领域, Simonyan 和 Zisserman[22] 根据人类视觉识别的两通道假说[23] 训练两个通道的 CNN 网络, 一个 CNN 从帧图像中提取静态特征, 另一个 CNN 从光流中提取运动特征, 然后合并这两者的深度特征实现动作识别. 这种从两个通道提取互补深度特征的方法在后续研究中有很多变体[24−28]. 例如, Glkioxari 和 Malik[24] 利用这两个通道的深度特征在帧图像中检测动作, 并把每个帧中的检测结果连接为三位时空域中的动作管道. Cheron 等[25] 对帧图像中不同肢体部件提取静态和运动深度特征, 经过合并形成视频的姿态特征, 然后识别动作类别. 受这些研究的启发, 本文通过探索深度特征与手工设计的姿态特征的互补关系. 本文从图像局部区域中提取姿态特征, 从整体图像提取深度特征, 利用这两个通道的互补线索识别静止图像中的动作.

2　提取姿态特征

本文的姿态特征是一组 Poselet 检测结果组成的向量. 本节介绍如何用 Poselet 表示肢体部件的姿态, 以及如何利用上下文检测 Poselet.

2.1　姿态特征的表示方法

控制理论用状态空间表示系统可能出现的状态集合. 状态空间是以状态变量为坐标轴形成的向量空间, 系统的任何一个状态可以用状态空间中的一个坐标表示. 向量空间这种表示可变状态的方法也在计算机视觉中得到了应用. 例如, 在底层信息处理中, 彩色图像中的每个像素由 RGB 三维向量空间中

的坐标表示. 这种表示法还被用来估计人脸和躯体的姿态. Mikolajczyk 等[29] 训练了两个人脸检测模板, 一个是侧面的, 一个是正面的, 测试图像中的人脸朝向由这两个模板的检测结果表示. Maji 等[7] 对训练集中的人脸和肢体朝向做了标注, 对每个朝向训练 Poselet, 在测试图像中检测每个 Poselet 的触发得分组成向量, 脸部或肢体的朝向特征由这个向量表示.

本文把文献 [7] 表示人脸和肢体朝向的方法推广到每个选定的肢体部件. 为了利用训练实例窗口中的交互物体和背景线索提高动作识别能力, 同一 Poselet 的训练实例来自同一动作的训练集. 数据集中人体都标注了关键点, 对于每个肢体部件, 先对同一动作训练集中该部件的图像窗口按照关键点的坐标聚类. 同一聚类组成一个 Poselet 的正实例, 负实例从其他动作的训练集中随机截取. 图 1 显示了部分打高尔夫球动作中胳膊 (包括左胳膊和右胳膊) Poselet 的训练实例.

训练好 Poselet 以后, 就可以从测试图像中检测每个 Poselet 的触发得分. 部件的姿态特征由这组 Poselet 检测得分组成的向量表示.

人体整体具有所有的姿态信息, 但是由于粒度太粗, 难以从中提取姿态线索, 尤其是胳膊和腿包含的与动作有关的姿态语义信息. 另一方面, 胳膊或腿等部件具有含部分姿态. 因此, 本文使用图 2 所示层次部件树中 10 个肢体部件的姿态特征表示人体姿态. 图 2 中, 从根节点开始, 依次根据包含关系产生子节点, 第二层节点包括胳膊、躯干 + 头、腿, 第三层节点有左胳膊、右胳膊、头、躯干、左腿、右腿. 从根到叶, 肢体粒度由粗变细, 可以充分利用这些部件各自在检测中的显著性大小、是否包含姿态和动作语义线索等方面的优点. 更重要的是, 后面将分析, 图 2 的层次部件树结构为检测每个 Poselet 提供了一个精巧的上下文, 可以用来抑制检测错误, 提高动作识别性能.

2.2 Poselet 上下文模型

给定 Poselet 模板, 可以在图像中检测这个 Poselet, 但这样的检测结果模棱两可, 难以辨别真假. 首先, 尽管物体检测水平不断提高, 但仍然是一个未解决的开放问题. 即使有足够的训练实例, 经过充分训练的 Poselet 模板, 也可能在没有人的图像中检测到 Poselet, 而在有这个 Poselet 的图像中却无法检测到. 其次, Poselet 是在训练实例数量少, 质量差的情况训练出来的. 训练 Poselet 需要标注人体的关键点, 这需要耗费大量时间, 因此, 无法为训练 Poselet 提供海量的实例. 同一 Poselet 的训练实例是通过对关键点坐标聚类确定的, 同一聚类中的实例在姿态和外貌上有很大差别, 这种实例集合难以训练出完美的检测模板.

在物体识别领域, 物体所在的上下文被用来提高检测能力. 检测 Poselet 与检测物体本质上是一致的, Poselet 的检测也可以利用上下文提高检测能力. 例如, 一个圆形的轮廓可能是人脸, 如果在它的

图 1 打高尔夫球动作中部分胳膊 Poselet 训练实例

Fig. 1 Instances for some arm poselets in playing golf

下部检测到躯体, 则人脸的可能性就更大. Bourdev 等[16] 把 Poselet 作为相互检测的上下文, 为每个 Poselet 训练一个星型的上下文模型. 这种上下文模型类似一个两层的前向网络, 输入层是每个 Poselet 的检测结果, 在输出层, 每个 Poselet 利用上下文改善检测结果, 获得上下文环境中的检测得分. 这个过程中, Poselet 数量不变, 但每个 Poselet 都利用了与其他 Poselet 的同时出现关系组成的上下文优化了检测结果.

本文的上下文模型也利用每个 Poselet 的检测结果, 不同的是, 本文利用了肢体部件之间相对位置和同时出现的制约关系构成的 Pictorial 结构, 为检测其中每个 Poselet 构建了一个共有的上下文, 见图 2.

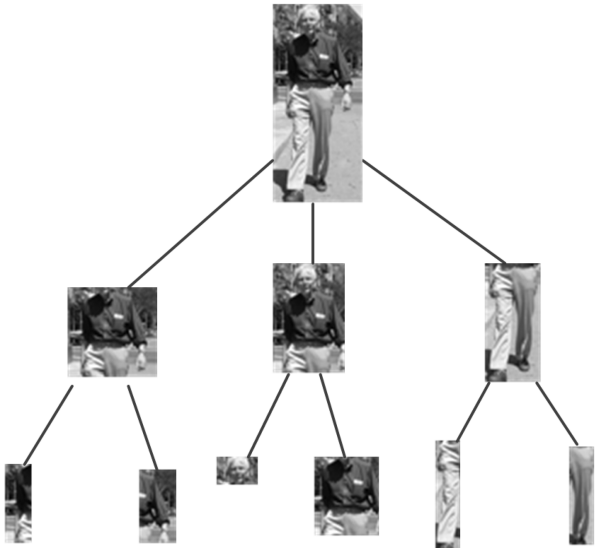

图 2　层次部件树

Fig. 2　Hierarchical part tree

图 2 中的每个部件由一组 Poselet 表示. 对于图像 I, $a \in \{a_1, a_2, \cdots, a_m\} = \Lambda$ 为动作类别, m 是类别数量. $L = (l_1, l_2, \cdots, l_{10})$ 是层次部件树上 10 个肢体部件的位置, $l_i = (p_i, t_i)$, 其中 $p_i = (x_i, y_i)$ 是部件 i 在图像中的坐标, $t_i \in T_i$ 是该图像中部件 i 的 Poselet 类别代号. 用 $G = (V, E)$ 表示层次部件树, $i \in V$ 表示部件 i, $(i, j) \in E$ 表示连接 i 和 j 的边, 层次部件树构成了一个 Pictorial 可变形模型[13]:

$$\sum_{i \in V} \Phi_a(I, l_i) + \sum_{(i,j) \in E} \Psi_a(l_i, l_j) \tag{1}$$

其中, $\Phi_a(I, l_i)$ 是部件 i 的外貌潜在函数, 描述动作 a 中部件 i 的 Poselet 模板 $\alpha_{t_i,a}$ 在图像坐标 p_i 处的匹配得分, 按下式计算:

$$\Phi_a(I, l_i) = \alpha_{t_i,a} \cdot \varphi(I, p_i) \tag{2}$$

$\alpha_{t_i,a}$ 是 Poselet t_i 的模板, $\varphi(I, p_i)$ 是图像 p_i 处的 HOG 特征. $\Psi_a(l_i, l_j)$ 是部件 i 和 j 的对偶潜在函数, 表示给定动作 a 时, 两者的变形代价和同时出现代价, 按下式计算:

$$\Psi_a(l_i, l_j) = \beta_{ij,a} \cdot \psi(\mathrm{d}p_{ij}) + b_{ij,a}^{t_i t_j} \tag{3}$$

$\psi(\mathrm{d}p_{ij}) = [\mathrm{d}x_{ij}, (\mathrm{d}x_{ij})^2, \mathrm{d}y_{ij}, (\mathrm{d}y_{ij})^2]^{\mathrm{T}}$, $\mathrm{d}x_{ij} = x_i - x_j - x_0$, $\mathrm{d}y_{ij} = y_i - y_j - y_0$, (x_0, y_0) 是部件 i 相对于它的父节点 j 的锚接位置, 即训练数据集中 i 相对于 j 的平均坐标差. 与文献 [17−18] 的可变形模型相比, 式 (1) 中的对偶潜在函数 (3) 中除了类似弹簧的变形代价, 还增加了同时出现代价, 描述 Poselet t_i 与 t_j 彼此相容还是互斥及其程度. 不同的动作图像中, 有不同的同时出现关系. 例如, 在行走图像中, 胳膊 Poselet 以及腿 Poselet 或者都是背面的, 或者都是侧面的, 不会出现背面的胳膊 Poselet 与侧面的腿 Poselet 在同一图像中. 而在舞蹈图像中, 这种同时出现关系就经常发生. $\Psi_a(l_i, l_j)$ 中引入同时出现代价能够进一步利用 Poselet 外貌之间的关系, 更准确地描述 Poselet 之间的约束关系. 在本文两个数据集上的对比实验结果表明, 引入同时出现代价后, 动作识别性能分别提高了 2.7 % 和 2.1 %.

图 3 是上下文模型的示意图, 每个部件节点由一组 Poselet 表示, 这些部件节点的 Poselet 之间形成了紧密的制约关系. 图中第二层的第一个节点列举了三个姿态各异的 Poselet, 每个 Poselet 由四个训练实例表示, 从身体朝向可以看出, 第一个向前, 第二个向后, 第三个偏向右. 如果在检测第一个 Poselet 时, 它的上层节点以及下层节点的锚节点附近出现了很强的相容 Poselet 信号, 上下文环境就支持这个 Poselet, 因此要提高它的检测得分. 沿着树的边直到根节点以及其他节点, 可以从树中每个节点获得检测这个 Poselet 的支持或扣分. 可见, 这一模型为检测 Poselet 提供了一个结构严密的上下文.

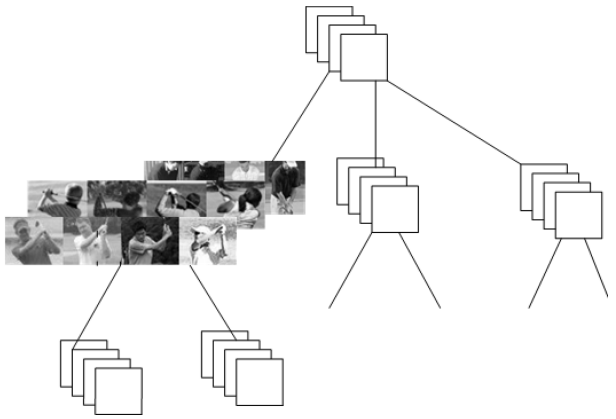

图 3　Poselet 上下文模型

Fig. 3　Poselet context

一种简单而幼稚的做法是为每个 Poselet 训练一个上下文模型. 这样, 把待检测 Poselet 所在的节点作为根节点, 利用 Pictorial 结构[13] 的推理, 置信度传播到根节点时, 就可以得到这个 Poselet 在上下文空间 L 中的最优得分. 但这样做, 需要训练大量的模型. 本文使用同一个模型, 为层次部件树上每个 Poselet 提供一个共有的上下文环境.

图 4 是该上下文模型的示意图. 模型参数训练好以后, 利用这个模型作为上下文环境, 在图像中逐一检测其中的每个 Poselet. 图中显示了检测到第二层节点的第三个 Poselet 时的情况. 此时, 该节点 Poselet 限制为一个, 但检测模板可以在图像空间中移动, 其他节点各自调整 Poselet 类别以及位置, 使得目标函数取得最大值. 式 (1) 匹配的目标函数为:

$$S_a(I,L)=\max_{L}(\sum_{i\in V}\Phi_a(I,l_i)+\sum_{(i,j)\in E}\Psi_a(l_i,l_j)) \tag{4}$$

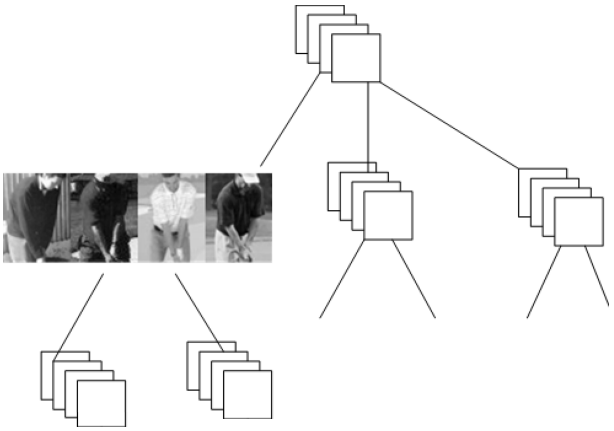

图 4 在上下文环境中检测 Poselet

Fig. 4 Detecting Poselet in context

利用上下文模型检测节点 v 的 Poselet τ 时, 该节点的 Poselet 类别限制为 τ, 目标函数为:

$$\begin{aligned}S_a(I,(p_v,\tau))=&\\ &\max_{L}(\Phi_a(I,(p_v,\tau))+\sum_{i\in V\setminus v}\Phi_a(I,l_i)+\\ &\Psi_a(l_v^{\tau},l_{\mathrm{parent}(v)})+\sum_{i\in\mathrm{children}(v)}\Psi_a(l_i,l_v^{\tau})+\\ &\sum_{(i,j)\in E\wedge i\neq v\wedge j\neq v}\Psi_a(l_i,l_j))\end{aligned} \tag{5}$$

等式右边的第一项是 τ 在坐标 $p_v=(x_v,y_v)$ 的检测得分, 第二项是除 v 节点外的其他节点在各自 l_i 处的检测得分, 第三项 $\Psi_a(l_v^{\tau},l_{\mathrm{parent}(v)})$ 是 τ 与节点 v 的父节点 $\mathrm{parent}(v)$ 之间的对偶潜在函数, 如果 τ 是根节点, 这一项为 0. 第四项是 v 及其所有孩子节点之间的对偶潜在函数之和, 如果 v 是叶子节点, 这一项为 0. 最后一项是层次部件树中与 v 节点无关的对偶潜在函数之和.

在动作图像中检测节点 v 的 Poselet τ 时, 目标函数 (5) 衡量 τ 及其他部件节点的 Poselet 与上下文模型的匹配程度. 匹配是一个置信度传播的过程, 这一过程可以通过 Pictorial 结构[13] 的距离转换算法高效地实现. 置信度传播从 HPT 的叶节点开始, 到根节点结束. 用 $S_a(I,l_i)$ 表示节点 i 收到每个孩子传来的置信度后的得分, 则有:

$$S_a(I,l_i)=\Phi_a(I,l_i)+\sum_{k\in C_i}m_k(l_i) \tag{6}$$

其中, $C_i=\{c_1,c_2,\cdots\}$ 是 i 的孩子节点集合, 当 i 为叶节点时, $m_k(l_i)$ 为 0. $\Phi_a(I,l_i)$ 是三维矩阵, 保存部件 i 的一组 Poselet 在图像二维空间中的检测结果. 当 Poselet τ 位于 i 节点时, 只保留 τ 的二维矩阵, 其他 Poselet 的检测结果均设置为 0. 节点 i 向它的父节点 j 传播的置信度为:

$$\begin{aligned}m_i(l_j)=&\max_{l_i}(S_a(I,l_i)+\Psi_a(l_i,l_j))=\\ &\max_{l_i}(S_a(I,l_i)+\beta_{ij,a}^{\mathrm{T}}\cdot\psi(\mathrm{d}p_{ij}))+\max_{t_i}b_{ij,a}^{t_it_j}\end{aligned} \tag{7}$$

距离转换以后执行与 $b_{ij,a}^{t_it_j}$ 的加运算, 因此没有增加时间复杂度. 式 (7) 中 $b_{ij,a}^{t_it_j}$ 是一个标量, 表示 Poselet t_i 和 t_j 的同时出现代价.

2.3 上下文模型的训练

式 (6) 匹配以后可以返回每个节点的坐标和 Poselet 类别, 因此模型训练是一个结构支持向量机的训练问题. 首先使用初始模型根据式 (6) 在训练图像上匹配, 获得返回结果形成的特征向量, 据此训练模型参数. 本文训练过程使用随机坐标下降法[30], 下面介绍模型参数的训练过程.

本文使用上下文模型的目的不是这个返回结果, 而是抑制检测错误. 在没有 Poselet τ 的图像中式 (6) 得分应该低, 而在有 τ 的图像中, 如果返回的 Poselet 类别和坐标与实际值吻合, 得分应该高.

给定 N 个训练图像 $\{I_i,L_i,A_i\}_{i=1}^{N}$, I_i 表示第 i 个图像, L_i 是图像中第 i 个部件的坐标和 Poselet 类别的标注, Poselet 类别是训练实例聚类以后分配的聚类编号, 坐标是该部件内所有关键点的中心坐标, $A_i\in\{a_1,a_2,\cdots,a_m\}=\Lambda$ 是图像的动作类别, 对动作 a 训练上下文模型 $\theta_a=(\alpha_a,\beta_a,b_a)$ 的目标

函数是:

$$\underset{\theta_a,\xi_i>0}{\operatorname{argmin}}(\parallel \theta_a \parallel^2 + C\sum_i \xi_i) \tag{8}$$

$$\text{s.t. } \forall i \in pos, \quad \theta_a^{\mathrm{T}} \cdot \Omega_i(\hat{L}_i) \geq \Delta_i - \xi_i \tag{9}$$

$$\forall i \in neg, \forall L, \quad \theta_a^{\mathrm{T}} \cdot \Omega_i(\hat{L}_i) \leq -1 + \xi_i \tag{10}$$

约束 (9) 中, $\Omega_i(\hat{L}_i)$ 是在训练图像上执行置信度传播后的返回的特征向量, Δ_i 根据返回 Poselet 的类别与标注类别的吻合程度取值, 如果超过半数的部件节点返回了正确的 Poselet 类别, 则 $\Delta_i = 1$, 否则 $\Delta_i = 0.5$. 约束 (9) 说明, 对于正实例, 匹配得分应该大于分界线 Δ_i, 式 (10) 说明对于负实例, 匹配得分应该小于分界线 -1, ξ_i 是松弛因子. 正实例是该动作类别的训练图像, 负实例是其他所有动作的训练图像. 对式 (10) 两端同乘 -1, 合并两个约束条件, 目标函数变为:

$$\underset{\theta_a,\xi_i>0}{\operatorname{argmin}}(\parallel \theta_a \parallel^2 + C\sum_i \xi_i) \tag{11}$$

$$\text{s.t. } \forall i, \quad \theta_a^{\mathrm{T}} \cdot \Omega_i \geq \Delta_i - \xi_i \tag{12}$$

约束 (12) 中, 对于负实例, $\Delta_i = 1$. 训练过程中, 为了确保目标函数是开口向下的凸函数, 当变形代价参数 β_a 中的二次变形权重为负时, 将其调整为 0.01.

3　提取深度特征

Alex 的 CNN 网络[21] 由五层卷积网络和三层全连通网络组成. 第一层卷积网络的输入是经过预处理的三维图像, 最后一层全连通网络的输出是对应 1 000 种物体类别的向量, 据此判定输入图像的类别, 中间每层网络的输出是下一层网络的输入. 这一架构的底层从图像提取诸如方向、梯度、颜色、频率等底层特征, 经过连续多次的卷积计算, 形成丰富的中间层特征, 以及更加抽象且有判别能力的高层特征[31].

CNN 是一个庞大的多层结构, Alex 的 8 层 CNN 网络有 6 千万个参数和 65 万个神经元. 从 CNN 提取深度特征, 首先需要百万数量级的图像数据集训练 CNN 网络. 否则, 训练过程中产生过度拟合, 而在测试数据上则无法得到理想的结果. 本文的难题是训练数据集只有数千个训练图像, 无法提供足够的训练数据, 对此有两种解决方法.

第一种方法是预训练和精细调节. 先在有足够训练实例的数据集上训练预备模型, 然后以预备模型的参数初始化 CNN 网络, 在新的数据集上以更小的速度训练最终模型. Alex 网络出现以来, 很多训练实例数量受限的研究选择此方法先在 ILSVRC 2012 数据集训练预备模型, 然后在各自的数据集上精细调节模型参数[32]. 第二种方法是利用 CNN 网络良好的领域适应性, 不做精细调节, 直接把第一种方法中的预备模型作为一个黑盒的特征提取器, 对新数据集中的图像提取深度特征. 使用这种方法, Donahue 等[31] 在 ILSVRC 2012 数据集训练 CNN 模型, 用来对鸟类数据集以及场景数据集提取深度特征.

本文以第一种方法提取深度特征. 预训练使用 Alex 的 8 层网络结构, 在 ILSVRC 2012 训练集上训练预备模型, 参见图 5 上部, 其中 C 表示卷积层,

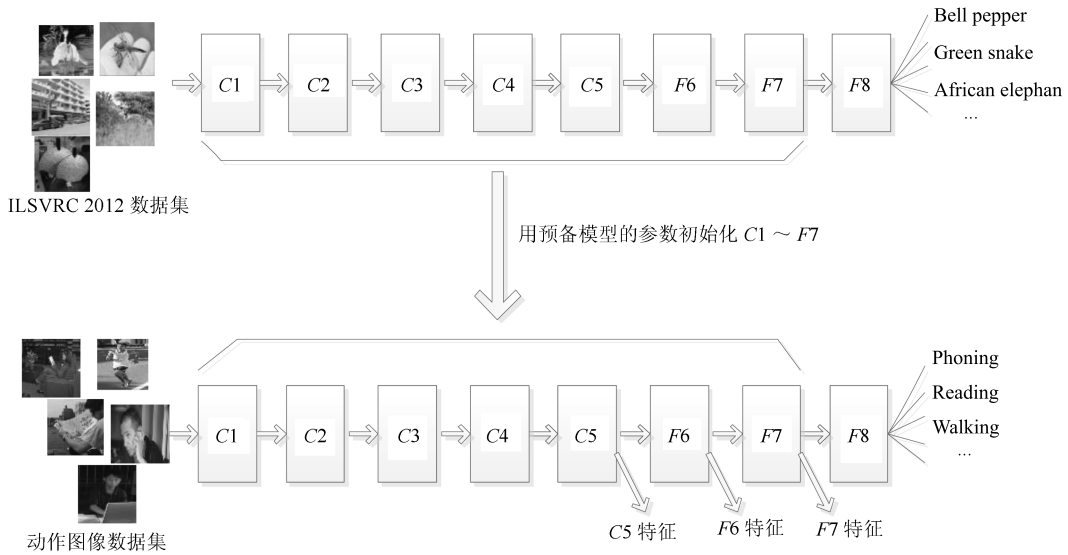

图 5　提取深度特征

Fig. 5　Extract deep features

F 表示全连通层. 用预备模型 $C1 \sim F7$ 层的参数初始化新模型的 $C1 \sim F7$ 层以后重新开始精细训练, 新模型 $F8$ 层的初始参数随机生成, 输出类别数量设置为数据集动作类别的数量 5. 精细调节训练的初始速度设置为 0.0001, 即训练预备模型初始速度的十分之一.

4 实验

下面分别介绍本文使用的数据集, 姿态特征单独使用时的动作识别性能, CNN 网络的动作识别性能, 以及姿态特征与深度特征混合使用后的性能.

4.1 数据集

在动作识别中, PASCAL VOC 2012 动作数据集被广泛用来检验识别算法, 并且与其他算法做性能比较. 但这个数据集很多图像中的人体只有部分可见. 例如很多阅读或者玩电脑的图像只有上半身, 打电话和摄影的很多图像只有头部和胳膊的一部分. 本文的姿态特征从肢体各部件中依次提取, 要求肢体的所有部分出现在图像中, 因此不使用这个数据集.

实验使用一个静止图像数据集和一个视频截图数据集, 本文在静止图像数据集的训练集上训练模型, 然后用同一模型分别评价在静止图像数据集以及视频截图数据集上的性能.

静止图像数据集是 Ikizler-Cinbis 等[11] 从互联网收集的静止图像. Cinbis 等以动作名称为关键字用 Google 以及 Yahoo 搜索引擎检索图像, 然后通过迭代执行检测人体和去除无关图像的无监督方式对搜索得到的数据集进行清洗. 数据集共有 2 458 个图像, 分别属于 5 个动作, 舞蹈、打高尔夫球、跑步、坐和行走, 部分图像参见图 6. 这些图像在拍摄角度, 人体外形以及人体姿态上有非常大的差别. 同行算法[17−18] 对该数据集中的关键点做过标注, 但是没有公开标注结果. 本文首先对其中每个人体标注 17 个关键点, 包括左右眼睛、左右耳朵、鼻子、左右手关节、左右肘关节、左右肩关节、左右臀关节、左右膝关节、左右脚关节, 参见图 7. 根据标注结果从原始图像中截取 128×64 大小的人体图像, 把其中 1/3 的图像作为训练集, 1/3 作为检验集, 剩下 1/3 作为测试集.

另一个数据集是 Niebles 等收集的低分辨率视频数据集 YouTube[33]. Ikizler-Cinbis 等[11] 用人体检测模板先对这个数据集的帧图像做人体检测, 然后对检测到的正确结果用矩形框做了标注, 并标注了动作类别. 与文献 [11, 17] 一样, 本文从标注的矩形框中截取人体窗口, 作为视频截图数据集. 该数据集的动作名称和数量与前一个数据集一致. 由于人体窗口是按照检测结果截取的, 存在人体不在图像中心位置, 人体尺度大小不一的问题, 图 8 显示了其中的部分图像.

图 6 静止图像数据集中的部分图像

Fig. 6 Some images in static image data set

图 7 标注了关键点的图像

Fig. 7 Some images with annotated key points

4.2 姿态特征的动作识别性能

上下文模型训练好以后, 对于给定的图像, 就可以根据 (5) 计算每个 Poselet 在层次部件树环境中的检测得分, 然后组合成人体的姿态特征. 层次部件树有 10 个节点, 每个节点由 12 个 Poselet 表示, 共 5 个动作, 姿态特征的维度为 $10 \times 12 \times 5 = 600$. 本文在静止图像数据集的训练集上训练线性 SVM, 然后分别在静止图像数据集和视频截图数据集上评价动作识别性能.

由于没有公开静止图像数据集的标注, 本文实验以及同行算法[17−18] 各自标注关键点, 训练集和测试集的划分也不一致. 这种情况下, 仅仅比较动作识别精度是不够的. 由于同行算法[11, 17−18] 都把对

整体图像 HOG 特征的多类别 SVM 作为基准算法, 本实验沿用这一基准, 下列性能比较中把相对于基准算法的性能提升作为一个重要的性能评价指标.

图 8　视频截图数据集中的部分图像

Fig. 8　Some images in video data set

表 1 显示了静止图像数据集上本文的姿态特征与同行算法的动作识别精度比较, 其中平均性能是 5 个动作识别精度的平均值. 表 1 中 CNN 黑盒是利用文献 [31] 训练的 CNN 网络作为特征提取器, 从动作图像中提取深度特征, 然后使用线性 SVM 分类的结果. Poselets[7] 是本文的姿态特征在没有应用上下文模型的分类性能. 从中可以看到, 利用上下文模型后, 本文的姿态特征性能提高了 5.07 %. 与其他平面特征相比, 本文的姿态特征具有最好的平均精度, 且识别精度比基准算法提高了 9.99 %, 这两个指标非常接近 CNN 黑盒[31] 提取的深度特征.

表 1　静止图像数据集上的动作识别精度 (%)

Table 1　Precision on static image data set (%)

方法	姿态特征平均精度	基准平均精度	与基准比较
四节点星型[17]	61.07	56.45	+ 4.62
20 节点模型[18]	65.15	62.8	+ 2.35
POSELETS[7]	61.33	56.41	+ 4.92
CNN 黑盒特征[31]	67.20	56.41	+ 10.79
本文的姿态特征	66.40	56.41	+ 9.99

表 2 是视频截图数据集上的动作识别性能. 这个数据集是根据人体检测结果标注的, 图像中人体尺度大小差别很大, 而且很多图像中人体没有居中. 同行算法 multiLR_NMF[11] 采用抖动技术使得人体大小适中且居中布置, 降低了识别难度, 动作识别性能及基准算法的识别性能有很大提高. 从表 2 可以看到, 本文的姿态特征在没有对截图图像做预处理的情况下, 取得了与同行算法[11] 非常接近的平均识别精度, 与基准算法比较提高了 14.16 %, 而算法[11] 只比基准算法提高了 4.26 %. 在该数据集上, 本文姿态特征的性能与 CNN 黑盒[31] 特征几乎持平.

表 2　视频截图数据集上的动作识别精度 (%)

Table 2　Precision on image form video data set (%)

方法	姿态特征平均精度	基准平均精度	与基准比较
四节点星型[17]	50.58	46.98	+ 3.6
multiLR_NMF[11]	63.61	59.35	+ 4.26
POSELETS[7]	54.84	49.42	+ 5.42
CNN 黑盒特征[31]	63.84	49.42	+ 14.42
本文的姿态特征	63.58	49.42	+ 14.16

4.3　CNN 的动作识别性能

本文通过预训练和精细调节方法训练 CNN 网络, 先在 ILSVRC 2012 训练集上训练预备模型, 然后在静止图像数据集上精细调节模型参数. 为了展示使用预备模型的效果, 本文对使用预训练模型前后在静止图像数据集上做了对比实验. 前者对 $C1 \sim F8$ 层均使用随机的初始数据, 在静止图像数据集的训练集上训练 CNN 网络. 图 9 比较了使用预备模型前后 CNN 训练中出错率的变化过程. 从图中可以看到, 没有使用预备模型的情况下, 训练结束时的出错率为 37 %, 经过预训练和精细调节出错率降低到 28 %, 下降了 9 %. 可见, 通过精细调节, 很大程度上缓解了过度拟合问题.

4.4　姿态特征与深度特征混合后的性能

CNN 的第一层卷积网络检测图像中的频率、方向、颜色等底层特征, 有判别能力的特征在高层[21]. 本文从卷积网络的最后一层 $C5$ 开始, 依次提取图像的 $C5$、$F6$、$F7$ 层输出分别组成深度特征向量, 参见图 5 下部, 通过实验验证深度特征与本文提出的姿态特征在动作识别中的互补作用.

图 5 中, $C5$ 是 Alex 网络的最后一个卷积层, 对 $C4$ 层输出的卷积运算结果执行 Relu 非线性运算后, 再经过 max 合并, $C5$ 层的输出是一个 $6 \times 6 \times 256 = 9\,216$ 矩阵. $F6$ 层通过 $6 \times 6 \times 256 \times 4\,096$ 卷积模板对 $C5$ 输出执行全连通卷积运算, 然后执行 Relu 变换, 输出是一个 4 096 维向量. $F7$ 层经过 $4\,096 \times 4\,096$ 全连通卷积, 输出特征仍然是 4 096 维向量.

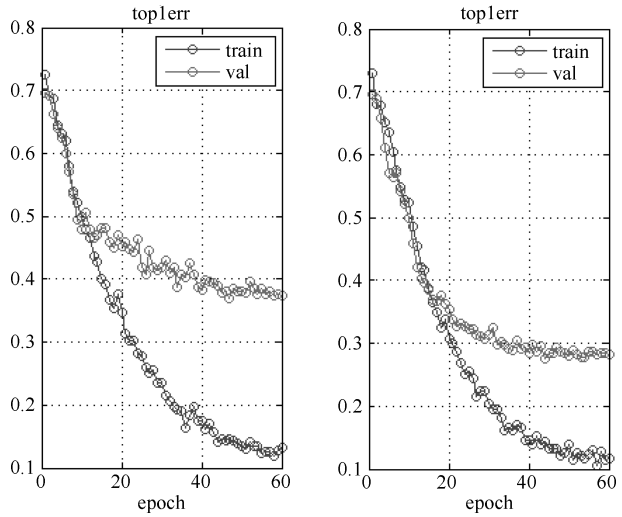

图 9 使用预备模型前后 CNN 训练过程 top1 错误率比较

Fig. 9 Comparison of top1 error between whether using pre trained model

表 3 是静止图像数据集上本文提出的姿态特征, CNN, 以及姿态特征与不同深度特征混合后的性能比较. 其中 CNN 是通过预训练和精细调节的 CNN 深度网络分类的结果, 姿态特征和深度特征的混合是把两者连接后使用线性 SVM 的分类结果. 从表中可以看出, CNN 以 71.7 % 精度高于姿态特征的识别精度. 每种姿态特征与不同深度特征混合以后, 性能均超过 CNN 以及姿态特征单独使用时的性能, 且三种深度特征有不同的表现. 姿态特征与 $F6$ 和 $C5$ 混合后的性能比较接近, 高出与 $F7$ 混合后的性能 1.4 个百分点.

表 3 静止图像数据集姿态特征、CNN 及混合后性能比较 (%)

Table 3 Precision comparison on static image data set (%)

方法	舞蹈	打高尔夫球	跑步	坐	行走	平均
姿态特征	69	65	74	65	59	66.4
CNN	72.4	76.4	70.2	73.9	65.6	71.7
姿态特征 + $C5$	79.4	68.8	79.9	77.4	72.0	75.5
姿态特征 + $F6$	78.5	70.2	77.9	78.3	74.5	75.8
姿态特征 + $F7$	75.3	68.9	76.2	79.3	72.4	74.4

表 4 是视频截图数据集上本文提出的姿态特征、CNN、以及姿态特征与不同深度特征混合后的性能比较. 从表 4 可以看到与表 3 类似的结果, 即姿态特征与深度特征的混合使用远远大于两者单独使用的性能.

从表 3 和表 4 可以看出, 对不同的动作类别, 姿态特征和 CNN 深度特征各有优势. 对于跑步动作, 两个数据集上姿态特征的性能均高于 CNN 深度特征, 因为跑步中的人体姿态明显不同于其他四个动作, 尤其是胳膊和腿的姿态. 打高尔夫球也有独特的人体姿态, 视频截图数据集上的结果显示姿态特征明显优于深度特征. 而舞蹈动作中的很多姿态与行走以及坐相似, 这三种动作的姿态容易混淆. 表 3 和表 4 中显示, 对这三种动作, 从整体图像提取的深度特征性能高于姿态特征.

表 4 视频截图数据集姿态特征、CNN 及混合后精度比较 (%)

Table 4 Precision comparision on video data set (%)

方法	舞蹈	打高尔夫球	跑步	坐	行走	平均
姿态特征	52.1	91.5	83.6	39.4	51.4	63.6
CNN	62.2	58.9	76.2	63.9	58.9	64.0
姿态特征 + $C5$	63.4	65.7	82.3	61.5	65.5	67.6
姿态特征 + $F6$	69.8	64.6	84.5	64.3	66.3	69.9
姿态特征 + $F7$	67.5	64.1	82.5	63.7	65.8	68.7

为了进一步认识姿态特征和深度特征在动作识别中的互补作用, 图 10 显示了部分姿态特征识别正确而深度特征识别错误的图像, 图 11 显示了部分深度特征识别正确而姿态特征识别错误的图像, 两图左侧是静止图像数据集中的图像, 右侧是视频截图数据集中的图像. 图 10 中的每个图像都有典型的动

图 10 姿态特征识别正确而深度特征识别错误的图像

Fig. 10 Some images recognized accurately by pose feature but falsely by deep feature

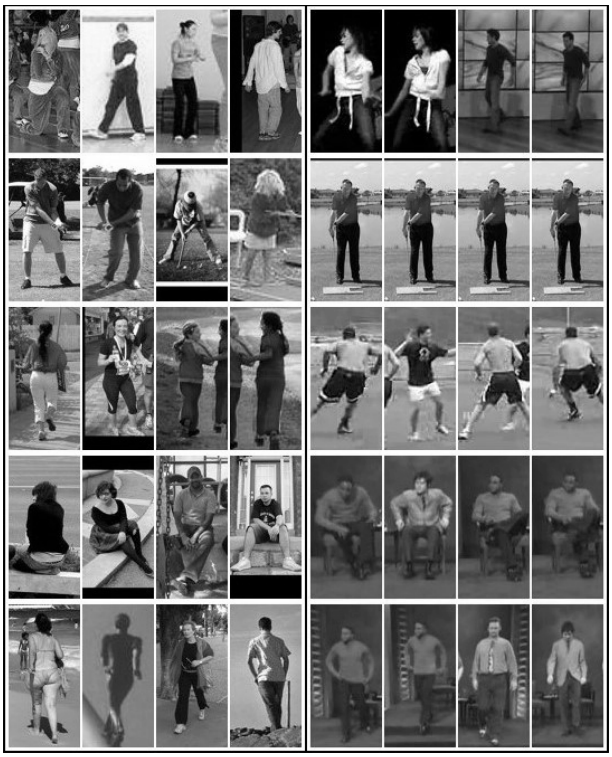

图 11　深度特征识别正确而姿态特征识别错误的图像

Fig. 11　Some images recognized accurately by deep feature but falsely by pose feature

作姿态, 但往往包含了干扰动作识别的背景, 甚至有些是无背景的卡通图像, 可见, 这类图像中深度特征不如姿态特征. 图 11 中的图像有些没有明显的动作姿态, 有些尺度过大或过小, 但图像中包含该动作的背景或物体, 例如凳子、球场等等, 这类图像中, 深度特征的识别能力超过姿态特征.

5　结论

静止图像中的动作可以通过图像局部中的姿态线索识别, 也可以对整体图像训练 CNN 网络识别. 姿态特征描述人体的姿态语义信息, 深度网络中的高层特征是经过深度学习得到的抽象特征, 本文通过实验验证了两者在动作识别中的互补关系. 为了在训练实例数量有限的数据集上提取深度特征, 通过预训练和精细调节的方法训练深度网络, 然后提取 $C5$、$F6$、$F7$ 三层的输出作为深度特征. 实验结果表明, 这三层特征与姿态特征混合使用, 性能均得到很大提升, 且远远超过了单独使用 CNN 的性能. 两个数据集上的实验结果表明, F6 层深度特征与姿态特征混合使用性能最好, 且最稳定.

References

1 Aggarwal J K, Ryoo M S. Human activity analysis: a review. *ACM Computing Surveys*, 2011, **43**(3): Article No. 16

2 Jiang Y G, Wu Z X, Wang J, Xue X Y, Chang S F. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **40**(2): 352−364

3 Wu Z X, Jiang Y G, Wang X, Ye H, Xue X Y. Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the 2016 ACM on Multimedia Conference. Amsterdam, The Netherlands: ACM, 2016. 791−800

4 Zhu Yu, Zhao Jiang-Kun, Wang Yi-Ning, Zheng Bing-Bing. A review of human action recognition based on deep learning. *Acta Automatica Sinica*, 2016, **42**(6): 848−857
(朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述. 自动化学报, 2016, **42**(6): 848−857)

5 Guan Qiu-Ju, Luo Xiao-Mu, Guo Xue-Mei, Wang Guo-Li. Compressive infrared classification of human motion using HMM. *Acta Automatica Sinica*, 2017, **43**(3): 398−406
(关秋菊, 罗晓牧, 郭雪梅, 王国利. 基于隐马尔科夫模型的人体动作压缩红外分类. 自动化学报, 2017, **43**(3): 398−406)

6 Guo G D, Lai A. A survey on still image based human action recognition. *Pattern Recognition*, 2014, **47**(10): 3343−3361

7 Maji S, Bourdev L, Malik J. Action recognition from a distributed representation of pose and appearance. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE, 2011. 3177−3184

8 Yao B P, Li F F. Action recognition with exemplar based 2.5D graph matching. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 173−186

9 Yao B P, Jiang X Y, Khosla A, Lin A L, Guibas L, Li F F. Human action recognition by learning bases of action attributes and parts. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 1331−1338

10 Delaitre V, Laptev I, Sivic J. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: Proceedings of the 21st British Machine Vision Conference. Aberystwyth, UK: BMVC Press, 2010. 1−11

11 Ikizler-Cinbis N, Cinbis R G, Sclaroff S. Learning actions from the Web. In: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 995−1002

12 Wang Y, Jiang H, Drew M S, Li Z N, Mori G. Unsupervised discovery of action classes. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006. 1654−1661

13 Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005, **61**(1): 55−79

14 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE, 2011. 1385−1392

15 Bourdev L, Malik J. Poselets: body part detectors trained using 3D human pose annotations. In: Proceedings of IEEE the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 1365−1372

16 Bourdev L, Maji S, Brox T, Malik J. Detecting people using mutually consistent poselet activations. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010. 168−181

17 Yang W L, Wang Y, Mori G. Recognizing human actions from still images with latent poses. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2030−2037

18 Wang Y, Tran D, Liao Z C, Forsyth D. Discriminative hierarchical part-based models for human parsing and action recognition. *Journal of Machine Learning Research*, 2012, **13**(1): 3075−3102

19 Ni B B, Moulin P, Yang X K, Yan S C. Motion part regularization: improving action recognition via trajectory group selection. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3698−3706

20 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 3551−3558

21 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: ACM, 2012. 84−90

22 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 2014 Advances in Neural Information Processing Systems. Montrál, Canada: NIPS, 2014. 1−8

23 Goodale M A, Milner A D. Separate visual pathways for perception and action. *Trends in Neurosciences*, 1992, **15**(1): 20−25

24 Gkioxari G, Malik J. Finding action tubes. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 759−768

25 Cheron G, Laptev I, Schmid C. P-CNN: pose-based CNN features for action recognition. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 3218−3226

26 Chen J W, Wu J, Konrad J, Ishwar P. Semi-coupled two-stream fusion ConvNets for action recognition at extremely low resolutions. In: Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision. Santa Rosa, CA, USA: IEEE, 2017. 139−147

27 Shi Y M, Tian Y H, Wang Y W, Huang T J. Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimedia*, 2017, **19**(7): 1510−1520

28 Tu Z G, Cao J, Li Y K, Li B X. MSR-CNN: applying motion salient region based descriptors for action recognition. In: Proceedings of the 23rd International Conference on Pattern Recognition. Cancun, Mexico: IEEE, 2016. 3524−3529

29 Mikolajczyk K, Choudhury R, Schmid C. Face detection in a video sequence-a temporal approach. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA: IEEE, 2001. II-96−II-101

30 Ramanan D. Dual coordinate solvers for large-scale structural SVMs. USA: UC Irvine, 2013

31 Donahue J, Jia Y Q, Vinyals O, Hoffman J, Zhang N, Tzeng E, et al. DeCAF: a deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: ICML, 2014. 647−655

32 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2014. 580−587

33 Niebles J C, Han B, Ferencz A, Li F F. Extracting moving people from internet videos. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 527−540

钱银中 复旦大学计算机科学技术学院博士研究生, 常州信息职业技术学院副教授. 主要研究方向为计算机视觉和机器学习. 本文通信作者.
E-mail: yinzhongqian10@fudan.edu.cn
(**QIAN Yin-Zhong** Ph. D. candidate at the School of Computer Science, Fudan University. He is also an associate professor in Changzhou College of Information Technology. His research interest covers computer vision and machine learning. Corresponding author of this paper.)

沈一帆 复旦大学计算机科学技术学院教授. 研究方向为计算机图形学和科学计算的可视化. E-mail: yfshen@fudan.edu.cn
(**SHEN Yi-Fan** Professor at the School of Computer Science, Fudan University. His research interest covers computer graphics and scientific computing visualization.)