DOI:10.7544/issn1000-1239.2015.20140492
52(11):2431-2440,2015

一种层次化的联合识别模型

陈耀东[1,2]　　李仁发[2]
[1](长沙师范学院电子与信息工程系　长沙　410100)
[2](湖南大学信息科学与工程学院　长沙　410082)
(yaodong.chen@gmail.com)

A Hierarchical Model for Joint Object Detection and Pose Estimation

Chen Yaodong[1,2] and Li Renfa[2]
[1](*Department of Electronic and Information Engineering, Changsha Normal University, Changsha* 410100)
[2](*College of Computer Science and Electronic Engineering, Hunan University, Changsha* 410082)

Abstract　Object detection and pose estimation belong to different tasks in computer vision. Viewed from research methods and practical application, there is great complementarity between these two tasks. This paper presents a mixture of hierarchical tree models that consists of three types of nodes, representing the whole object, discriminative parts and components (i.e. semantic parts) respectively. A key point of the model is that the discriminative parts in the middle level characterize not only object features but also mutual information among components. The proposed model can detect articulated objects and estimate their poses in parallel so as to address the error propagation problem that exists in previous joint models. For training the model, we use a latent structured SVM method where the discriminative nodes are viewed as latent variables. A novel learning method is introduced to initialize and optimize the parameters of the discriminative parts automatically. In experiments we design two evaluation scenarios (i.e. multi-task recognition and single-task recognition) to compare the proposed model and obtain the performance with the state-of-the-art joint methods on PASCAL VOC datasets. The results show that the hierarchical model not only outperforms other joint models in both recognition rate, but also has higher time-effectiveness.

Key words　multi-task recognition model; pose estimation; object detection; part-based models; structured SVM; latent variables

摘　要　目标检测与姿态估计在当前视觉研究中分属不同的任务,但两者在研究方法和现实应用上具有较强的互补性.提出了一种混合的层次树模型,该模型包含3类结点,分别描述整体目标、判别部件和组件(即语义部件).中间层的判别部件兼顾承上(目标)与启下(组件)的功能,一方面刻画整体目标的局部特征,另一方面隐含多组件的共现信息.相比当前最新的联合模型,层次树模型能够并行化处理检测与估计,避免串联化联合引发的错误传播.采用基于隐变量的结构化支持向量机训练模型,同时提出了一种新的部件学习方法以自动地初始化和优化判别部件.实验设计了多任务识别和单任务识别2种评估场景,对比了本文模型与当前主流的联合识别模型,实验结果说明层次化模型具有更强的识别性能以及更高的时效性.

关键词　联合识别模型;姿态估计;目标检测;部件模型;结构化支持向量机;隐变量

中图法分类号　TP391.4

收稿日期:2014-06-03;**修回日期**:2014-12-16
基金项目:国家自然科学基金项目(60873047,61173036)

姿态估计研究如何定位非刚性目标的语义组件或者关节点,这些组件所标识的结构称之为姿态.姿态估计是一项复杂而又重要的视觉任务,在识别目标的行为或者意图上具有重要价值,其研究成果广泛应用于安防监控、视频检测、体感游戏等.目标检测是视觉识别中最基本的任务,目的是在给定图像中找出指定目标的所有位置.在当前的视觉研究中,姿态估计与目标检测面向不同的应用场景,分属为2个任务.检测目标不必关心目标内部的组件结构,比如常用的HOG(histogram of oriented gradient)模板检测器[1]将目标描述成一个整体特征模板,直接通过分类器判断目标的出现;而姿态估计一般假定目标位置已确定,仅关心内部组件的定位[2-4].但我们分析认为,这2个任务存在非常强的互补性:1)从研究方法上看,目标的姿态提供了精确的内部特征结构,正确识别姿态能辅助目标检测算法处理因局部形变和多视角变化引起的外观差异[5-6].例如在检测人体时,相比基于左右视角的划分策略[7],根据站立/坐卧/奔跑等多种姿态类别所构建的目标检测器具有更精确的识别效果[8].2)从应用背景的角度看,正确检测目标是进一步分析目标姿态的前提条件,检测性能将直接影响姿态估计的准确率.比如视频监控中首先要检测到行人的大致位置,然后再进一步确认行人的各种姿态.

近几年提出了一些目标检测与姿态估计相结合的识别算法.为解决实用环境下的人体姿态识别,文献[9]先通过一个已训练的躯干检测器定位目标的大致位置,然后再定位人体的各个组件.文献[10]基于关节点标注集自动学习一组检测特征Poselets[11-12],并利用此类特征识别目标的姿态.这些算法一般以模块嵌入的方式进行联合,提高指定任务的识别率.功能模块的添加需要花费额外的训练开销,更为不利的是,这种串行化联合容易引发错误传播.例如若躯干检测器[9]或者检测特征Poselets本身识别错误,则必然导致下一阶段的估计错误.文献[8]提出了一体化检测与估计的方法,以组件结构的得分作为判断目标是否存在的依据,尽管该方法能有效地刻画类内外观差异,但对于姿态估计的性能提升不大,而且直接使用组件描述目标被证明不一定获得最佳的检测性能[13-14].

本文在总结上述问题的基础上提出了一种层次化的联合识别模型(以下简称层次化模型).层次化模型(hierarchical model, HM)借鉴文献[8]的一体化思想,将内部姿态作为检测目标的重要依据.为了能兼顾目标与组件的识别性能,该模型在目标与组件之间添加一组判别部件,判别部件本质上是目标内的局部特征,由训练语料自动获取.文献[13-14]指出自动获取的判别部件相对于目标的固有组件①更适于检测目标;通过限定该部件集的选取与学习方式,层次化模型能建立组件尤其是远距离组件之间的共现信息,作为先验知识辅助组件定位.

图1对比展示了本文的层次化模型与当前最新的联合算法.相比串行化联合模型[9-10],层次化模型整合了目标与组件的得分,实现检测与估计的并行处理.相对于一体化模型[8],层次化模型在目标与组件之间添加一组判别部件,每个部件对于顶层目标结点而言相当于一组判别特征,是分类目标的有效依据;同时,每个部件提供若干组件的共现信息,辅助识别局部姿态.为了实现上述目标,本文应用基于隐变量的结构化支持向量机,该框架以组件为目标参数,将判别部件视作隐变量,在参数寻优的同时实现隐变量的自动学习.相比文献[9-10]的模型训练,本文所有参数的优化处于一个框架,能降低分次训练带来的累积开销.

部件模型是视觉识别中的一种重要的结构化建模方法,特别是PS(pictorial structure)[2]和DPM(deformable part model)[7],这2个模型分别在目标检测和姿态估计上具有里程碑意义.PS和DPM采用相近的描述方式与识别方法,但因应用背景不同,在定义部件以及度量空间关系上存在差异.一般认为DPM的关键技术有3个:1)基于判别部件的特征描述,部件之间建立"弹性"的空间关系;2)面向多视角的模式混合;3)基于LSVM的参数优化.文献[5]通过实验指出多模式混合对DPM检测率的贡献最大,尤其是针对非刚性目标.换言之,目标检测的最大挑战在于如何根据目标的结构特点建立有效的外观分类,而姿态估计正是分析目标结构的最佳途径.文献[13-14]的实验证明:部件相比组件具有更强的特征描述能力,因而目标检测率更高,这也是DPM未直接使用组件描述目标结构的关键原因.文献[15]通过部件共享降低DPM在多模式学习时的参数数量,提高整体学习效率.

PS作为部件模型与DPM有着相似的结构[2-4],但其内部使用目标的组件或者关节点.文献[3]提出

① 为了区别2类部件,本文将目标固有的语义部件称为组件(components),由算法自动获取且无语义性的判别化部件仍称为部件(parts).

了一个判别化的 PS 模型,将姿态估计描述为一个结构化分类问题,该模型的另一个关键技术是将部件划分为多个子类型,通过子类型的搭配组合可表示数目庞大的姿态形式.出于计算复杂度的考虑,该模型限定部件的关联为树状图结构.为了提高组件的定位率,文献[10]通过训练集自动学习判别化的 Poselets,Poselets 原本应用于目标检测[11-12],这里描述组件间的局部相关性.

本文的层次化模型综合了 DPM 与 PS 的各自特点,例如层次化模型的判别部件与 DPM 功能相似;与 DPM 不同之处在于,我们限定了判别部件的生成过程,使之与给定的组件相关,提供组件共现信息.

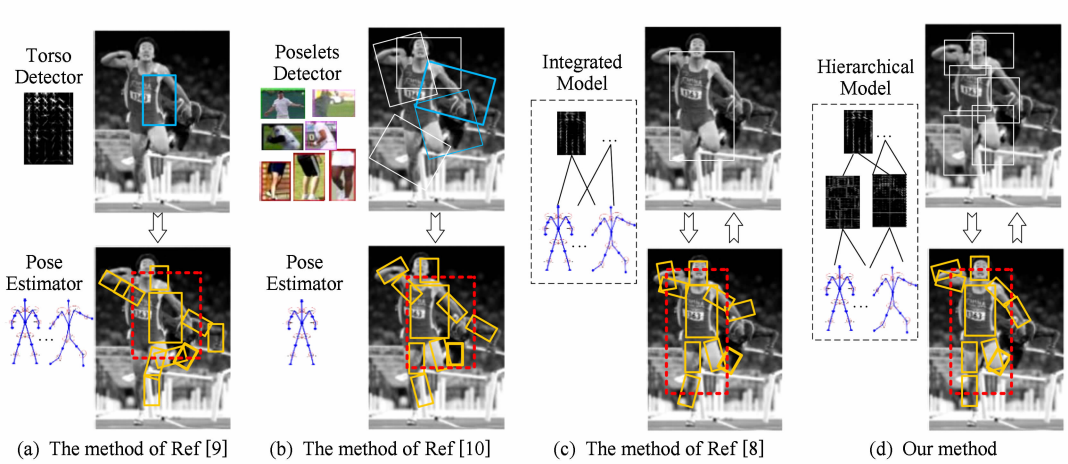


Fig. 1 The comparison between our recognition method and the state-of-the-art methods.

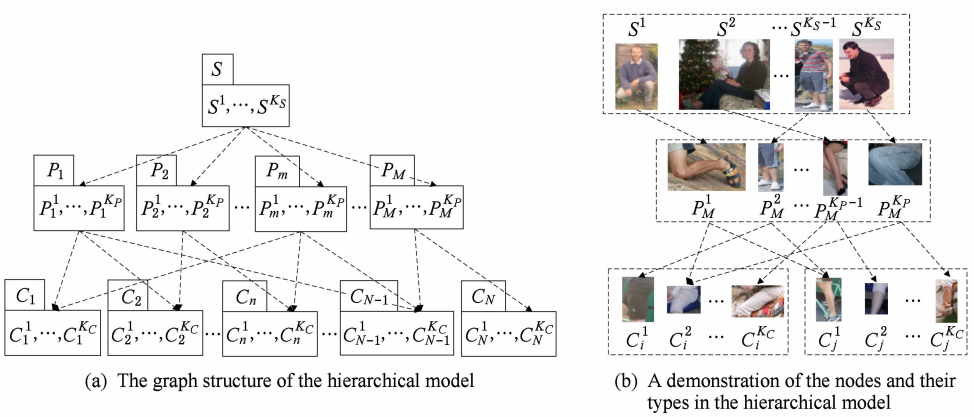
图 1 本文方法与当前联合识别方法的对比

1 层次化模型

1.1 概 述

假设一张图像包含多个非刚性目标(例如人体、动物等),层次化模型的初衷是实现目标和目标内部格局(即所有组件)的同时识别.图 2(a)描述了层次化模型的图状结构,第 1 层代表整体目标(根结点 S);第 2 层为判别部件集 $P_1,P_2,\cdots,P_M$,每个 P_m 代表目标的一组局部特征,判别部件与目标结点基于星状关系,即每个 P_m 相互独立,仅与 S 关联;第 3 层为组件集 $C_1,C_2,\cdots,C_N$,每个 C_n 代表一个关节点或者是目标的语义组件.一个判别部件至少与 2 个或 2 个以上的组件存在关联,组件之间相互独立.为了精确描述结点的外观,将图结构的每个结点划分为若干个子类型,每个子类型表示该结点的某个特定外观.图 2(b)对结点给出直观解释,S 的子类型代表在不同视角下的目标外观,判别结点 P_m 的子类型描述组件 C_i 和 C_j 某种特定的组合外观.例如,对于"右大腿" C_i 和"右小腿" C_j 这 2 个组件,P_m 的子类型实际表示右腿的各种姿态,包括"伸直"、"微曲"、"弯曲"等.

值得提出的有 2 点:1)结点以类型混合的方式描述上层结点的各种局部姿态;2)S 是一个单独的



(a) The graph structure of the hierarchical model

(b) A demonstration of the nodes and their types in the hierarchical model

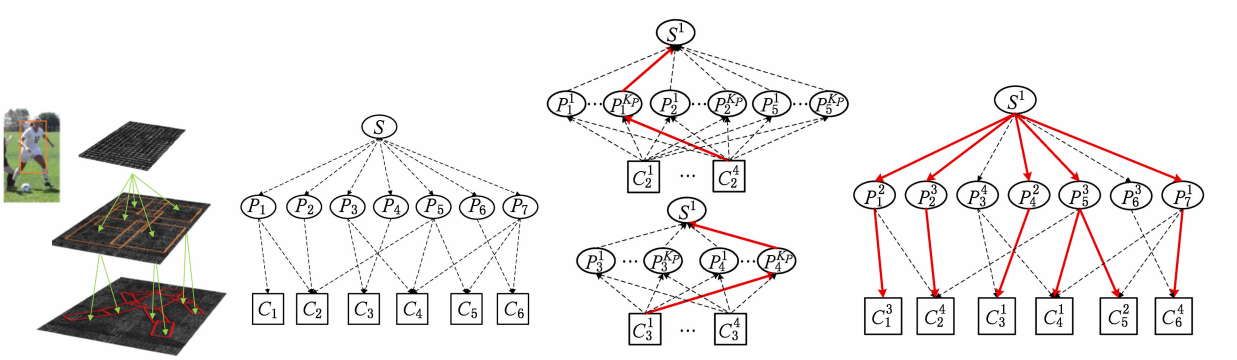
Fig. 2 The structure of the hierarchical model.

图 2 层次化模型的结构

结点,并且与其他结点相同具有独立的外观特征和对应的分类器.图 2(a)中结点子类型数目 K_C,K_P,K_S 的取值不全相同.假设部件数为 m、组件数为 n,上述结构可以描述 $K_S \times m^{K_P} \times n^{K_C}$ 种姿态形式,而 DPM 仅能描述 2～8 种姿态[7].

1.2 检测与估计

给定测试图像,模型建立层次化的特征金字塔(feature pyramid).如图 3(a),特征金字塔共 3 层,中间层设定为图像原始尺寸;上层和下层分别对应中间层像素的低采样和高采样结果,用于描述粗粒度和细粒度的特征值.细粒度特征能更好地描述尺度较小的目标,例如本文的组件;而粗粒度则适于快速检测目标位置.模型在 3 个层次上以滑动窗口策略抽取多个尺度的目标窗口、部件窗口和组件窗口.每个目标窗口在特征金字塔中具有一个最佳的内部格局,即每个结点指派有具体的子类型和空间信息(位置、尺度、方向),最后根据该格局的得分判断该候选窗口是否存在目标.



(a) The feature pyramid (b) An instance of the hierarchical model (c) Optimal path searching (d) An optimal configuration for a given sub-window

Fig. 3 The inference process of the hierarchical model.

图 3 层次化模型的推断过程

我们用 $(\boldsymbol{L},\boldsymbol{T})$ 描述目标窗口的一个候选格局,其中 $\boldsymbol{L}=(\boldsymbol{l}_0,\boldsymbol{l}_1,\cdots,\boldsymbol{l}_M,\boldsymbol{l}_{M+1},\cdots,\boldsymbol{l}_{M+N})$ 表示所有结点的空间信息,$\boldsymbol{T}=(t_0,t_1,\cdots,t_M,t_{M+1},\cdots,t_{M+N})$ 代表结点的具体类型.空间信息 $\boldsymbol{l}_i=(x,y,\sigma,\theta)$ 包括该结点相对窗口原点的坐标 (x,y)、尺度 σ 和方向 θ.θ 取值为 0～360°,实际赋值时为 24 个单元格(bin)(每 15°一个 bin)中的某个索引.注意:目标和判别部件对应结点的方向属性为常数 90°.格局 $(\boldsymbol{L},\boldsymbol{T})$ 的计算即是累积图的结点与边的得分:

$$sc(\boldsymbol{L},\boldsymbol{T})=\sum_{i\in V}sc(X_i^{t_i})+\sum_{i,j\in E}sc(X_i^{t_i},X_j^{t_j}),\quad(1)$$

其中,$sc(X_i^{t_i})=\boldsymbol{w}_i^{t_i}\cdot\boldsymbol{\phi}(\boldsymbol{l}_i)$,

$$sc(X_i^{t_i},X_j^{t_j})=\boldsymbol{w}_i^{t_i,t_j}\cdot\boldsymbol{d}(\boldsymbol{l}_i-\boldsymbol{l}_j).$$

结点得分表示为特征向量 $\boldsymbol{\phi}(\cdot)$ 与权重向量 $\boldsymbol{w}^{t_i}$ 的乘积,t_i 为该结点的具体类型,本文采用 HOG 特征模板[1]抽取位置 $\boldsymbol{l}_i$ 的特征向量.每个结点的得分计算是相互独立的.用 $\boldsymbol{w}^{t_i,t_j}$ 表示特征权重.图边的得分参见式(2),描述为 8 维特征向量,度量 2 个结点的空间相对性.

$$\begin{aligned}\boldsymbol{d}(\boldsymbol{l}_i-\boldsymbol{l}_j)=[&(x_i-x_j),(y_i-y_j),(\sigma_i-\sigma_j),\\&(\theta_i-\theta_j)(x_i-x_j)^2,(y_i-y_j)^2,\\&(\sigma_i-\sigma_j)^2,(\theta_i-\theta_j)^2].\end{aligned}\quad(2)$$

假设图 3(b)为某个目标类的模型结构,该结构本质上是一个受限的有向无环图,该图具有唯一的源点 S 和一组终止点 $C_1,C_2,\cdots,C_N$.图搜索的目标就是要找到所有终止点的最佳路径.给定图像和该图像的特征金字塔,图的搜索是一个从粗到精的递归过程.首先,在低像素层多尺度扫描候选的目标窗口,对每个目标窗口分别计算目标结点 S 所有子类型的得分;然后,对 S 的每个子类型,在对应窗口中计算下一层的部件结点,搜索每个结点的每个子类型以及出现的位置,这一过程直至图像的所有窗口均已得到每个结点的子类型和位置.为了实现计算的可跟踪性,在实际推断中模型格局限定为树形结构.对于每个底层的组件结点,我们采用动态规划策略求取最长路径(路径长度与得分成正比):

$$\begin{aligned}cost(C_n^{k_n},S^{k_s})=&\max_{P_m^{k_m}\in father(C_n^{k_n})}\{sc(C_n^{k_n})+\\&sc(P_m^{k_m},C_n^{k_n})+cost(P_m^{k_m},S^{k_s})\},\end{aligned}\quad(3)$$

其中,$k_m\in\{1,2,\cdots,K_P\}$,$k_n\in\{1,2,\cdots,K_C\}$,$k_s\in\{1,2,\cdots,K_S\}$.

$cost(C_n^{k_n},S^{k_s})$ 定义了从组件类型 $C_n^{k_n}$ 到源点类型 S^{k_s} 的一条路径开销.开销由 3 部分组成:当前组件的外观得分 $sc(C_n^{k_n})$、组件与某个父结点的边得分 $sc(P_m^{k_m},C_n^{k_n})$ 以及该父结点到源点的开销 $cost(P_m^{k_m},S^{k_s})$.$father(C_n^{k_n})$ 代表该组件的父结点集,如图 3

(c)中 $father(C_3^1)=\{P_3^1,P_3^2,\cdots,P_3^{K_P},P_4^1,P_4^2,\cdots,P_4^{K_P}\}$.基于式(3),层次化模型能够计算出该组件到某一目标类型 S^{k_s} 的最佳匹配类型 C_n^* 以及对应的最佳判别部件 P_m^*:

$$C_n^*=\arg\max_{k_n} cost(C_n^{k_n},S^{k_s})=\arg\max_{k_n,k_m}\{sc(C_n^{k_n})+sc(P_m^{k_m},C_n^{k_n})+cost(P_m^{k_m},S^{k_s})\}. \quad (4)$$

图 3(c)展示了 2 个组件 C_2 和 C_3 的搜索过程,首先计算该组件所有子类型到相应部件类型的路径开销,然后获取部件类型到所有源点类型的开销.式(5)定义了指定目标窗口下每个源点类型的开销,描述为该类型的外观得分加上到所有终止点(组件)的最佳路径开销.最后通过式(6)得到指定窗口的最优格局.图 3(d)展示了一个最优格局 $(\boldsymbol{L},\boldsymbol{T})^*$,该格局表现为一棵树状图.

$$cost(S^{k_s})=sc(S^{k_s})+\sum_{C_n} cost(C_n^*,S^{k_s}), \quad (5)$$

$$S^*=\arg\max_{k_s} cost(S^{k_s}). \quad (6)$$

2 模型训练

2.1 基于隐变量的结构化 SVM

观察式(1),层次化模型的参数包括:1)结点的外观特征权重.每个结点的外观权重可写为串行向量 $\boldsymbol{w}_i=\{\boldsymbol{w}_i^1,\boldsymbol{w}_i^2,\cdots,\boldsymbol{w}_i^k\}$,每个元素为相应子类型的权重向量.2)结点的空间特征权重.$\boldsymbol{w}_{i,j}=\{\boldsymbol{w}_{i,j}^{1,1},\boldsymbol{w}_{i,j}^{1,2},\cdots,\boldsymbol{w}_{i,j}^{k-1,k},\boldsymbol{w}_{i,j}^{k,k}\}$,每个元素为具体 2 个子类型的空间权重(参见式(2)).对于一棵生成的树状图,每个结点赋值一个具体的类型(假设为 k_i),则 $\boldsymbol{w}_i=\{\boldsymbol{0},\cdots,\boldsymbol{w}_i^{k_i},\cdots,\boldsymbol{0}\}$.相似地,每条边赋值一对具体的结点类型(假设为 k_i,k_j),则 $\boldsymbol{w}_{i,j}=\{\boldsymbol{0},\cdots,\boldsymbol{w}_{i,j}^{k_i,k_j},\cdots,\boldsymbol{0}\}$.在此基础上,式(1)可另记为如下线性形式:

$$sc(\boldsymbol{L},\boldsymbol{T})=(\boldsymbol{w}_0,\boldsymbol{w}_1,\cdots,\boldsymbol{w}_M,\cdots,\boldsymbol{w}_N,\cdots,\boldsymbol{w}_{i,j},\cdots)^{\mathrm{T}}\cdot(\boldsymbol{\phi}(\boldsymbol{l}_0),\cdots,\boldsymbol{\phi}(\boldsymbol{l}_N),\cdots,\boldsymbol{d}(\boldsymbol{l}_i-\boldsymbol{l}_j),\cdots)=\boldsymbol{\beta}\cdot\boldsymbol{\Phi}(\boldsymbol{L}). \quad (7)$$

当前,用于姿态估计的训练集给出了所有组件的标注窗口,通过组件标注很容易得到整体目标窗口.层次化模型的训练目标就是根据组件标注集学习式(7)的参数 $\boldsymbol{\beta}$.若不考虑判别部件,层次化模型可采用结构化 SVM 以监督方式优化参数.由于判别部件未在训练集标注也无法预先定义,只能从训练集自动获取,因此本文将判别部件的空间属性视为隐变量,引入基于隐变量的结构化 SVM[16] 解决式(7)的参数优化问题.该优化问题描述为:给定训练集 $S=\{(x^i,y^i)\}$,其中 x^i 为第 i 个训练图像,$y^i=\{C_1^i,C_2^i,\cdots,C_N^i\}$ 为标注的组件窗口.设定 Ω 为训练样本中判别部件可能出现的位置空间,$\boldsymbol{z}=(\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_M)$ 为该空间内的一组赋值,则有:

$$Loss_{\boldsymbol{\beta},\boldsymbol{z}}=\min\frac{1}{2}\|\boldsymbol{\beta}\|^2+C\cdot\sum_{i=1}^{N}\xi_{\boldsymbol{\beta},\boldsymbol{z}}^i,\quad \text{s.t. } \forall\boldsymbol{z}\in\Omega, \quad (8)$$

$$\xi_{\boldsymbol{\beta},\boldsymbol{z}}^i=\max_{y,\boldsymbol{z}}[\boldsymbol{\beta}\cdot\boldsymbol{\Phi}(x^i,y,\boldsymbol{z})+\Delta(y^i,y,\boldsymbol{z})-\max_{\boldsymbol{z}'}\boldsymbol{\beta}\cdot\boldsymbol{\Phi}(x^i,y^i,\boldsymbol{z}')], \quad (9)$$

其中,

$$\Delta(y^i,y,\boldsymbol{z})=\frac{1}{N}\sum_{n=1}^{N}[1-overlap(C_n^i,C_n)]. \quad (10)$$

式(8)搜索组件赋值 y^i 和隐变量 $\boldsymbol{z}$ 以获得具有最小惩罚度的 ξ 值.式(9)的 3 个分项分别代表从当前样本推断出的组件得分、标注格局与推断格局的差别度 Δ 以及标注格局得分,注意前 2 项的隐变量 $\boldsymbol{z}$ 与第 3 项的隐变量 $\boldsymbol{z}'$ 的取值并不相同.差别度 Δ 定义为 2 个标注组件集 y^i 与推断出的组件集 y 之间覆盖度的平均值.函数 $overlap(C_n^i,C_n)$ 度量标注组件 C_n^i 与推断组件 C_n 的窗口覆盖率.式(8)采用 CCCP(concave-convex procedure)[16] 算法求解,具体过程参见图 4(a).

训练框架包含 2 个部分——参数的初始化和优化.由于层次化模型各结点的外观计算是独立的,而且训练集 S 提供了目标结点和组件结点的标注,因此我们可以在模型结构确立(即判别部件自动获取)之前完成上述结点外观参数的初始化.在具体实施时,本文参考文献[11]的做法对结点相关的标注集执行 k-means 聚类(默认 $k=4$)以获取结点的子类型,相似度定义为 2 个样本窗口对应 HOG 特征向量的欧几里德距离:

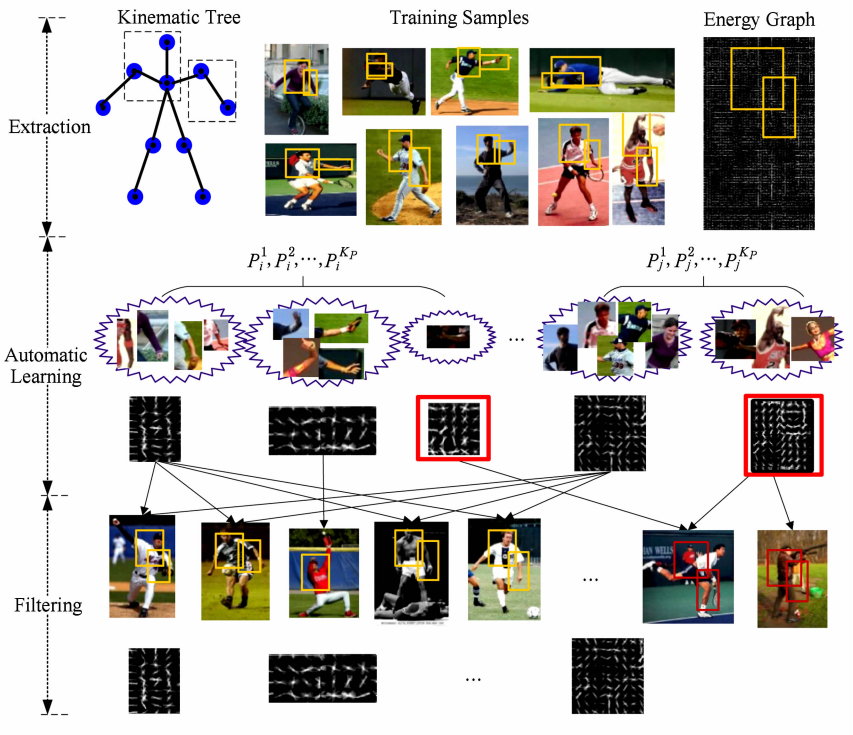
$$D_{\text{appearance}}(\boldsymbol{h}_1,\boldsymbol{h}_2)=\sqrt{\sum_{i=1}^{l}[h_1(i)-h_2(i)]^2}, \quad (11)$$

其中,$\boldsymbol{h}=\{h(1),h(2),\cdots h(l)\}$,$l$ 为向量长度.

将标注窗口划分成若干组后每组提供对应子类型的正样本,负样本随机取自不含目标的图像.根据正负样本集,我们采用线性 SVM 优化可分别确定组件外观参数 $\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_M\}$ 和目标相关的参数 $\boldsymbol{w}_s$ 的初值.初始化过程更为重要的是确定判别部件的形式,我们将在下一节详述其学习过程 *initialize*.假定判别部件已自动获取并已赋有初值,这意味着整体模型结构已经确立.我们给模型的每条关联边赋给相同的初始值.

Input: training set S, iteration number τ, a tree structure of components PT;
Output: parameter $\boldsymbol{\beta}$.
1) initialization of $\boldsymbol{\beta}$:
1.1 $\boldsymbol{w}_s \leftarrow svm(S)$, $\boldsymbol{w}_n \leftarrow svm(S)$
$\forall n \in \{1,2,\cdots,N\}$;
1.2 $\{\boldsymbol{w}_1 \cdots \boldsymbol{w}_M\} \leftarrow initialize(S, PT)$;
1.3 $\boldsymbol{w}_{i,j} \leftarrow \{0,0,0,0,1,1,1,1\}$,
$\forall (i,j) \in PT$;
2) running of CCCP, for $t := 1:\tau$:
2.1 $\boldsymbol{z}_i^* = \arg\max_{\boldsymbol{z}} \boldsymbol{\beta} \cdot \boldsymbol{\Phi}(i, y_i, \boldsymbol{z})$,
$\forall i \in S$;
2.2 $\boldsymbol{\beta}_t \leftarrow struc\_svm(\boldsymbol{\beta}_{t-1}, \boldsymbol{z}^*)$;

(a) The training framework of the model parameters (pseudo-code)



(b) An automatic learning of discriminative parts

Fig. 4 The training process of the hierarchical model.

图 4 层次化模型的训练过程

参数初始化完成后我们重新扫描样本集,执行 CCCP 在含有隐变量的条件下优化模型的各参数,这一过程简单描述为:根据当前参数搜索每个训练样本中判别部件的位置,相当于确定隐变量 $\boldsymbol{z}$ 的取值;固定样本的 $\boldsymbol{z}$ 值,执行常规的结构化 SVM 优化模型参数. 迭代 τ 次(默认 $\tau=8$)后训练完毕,关于 CCCP 的详细步骤参见文献[16].

2.2 部件学习

判别部件是层次化模型的关键所在,本节提出一种限定的部件学习方法. "限定"是相对于 LSVM[7] 的无限制部件搜索策略而言的. LSVM 在全局目标区域任意搜索高能量区作为描述目标的判别部件,而本文则将搜索限制于某些局部区域,这些区域在训练时定义为若干关联组件的合并窗口. 以此方法获取的判别部件既能有效地描述目标的局部外观,又提供了组件共现的先验知识. 图 4(b)展示了部件学习过程,共分为 3 步:

1) 提取局部搜索窗口. 算法首先依据常用的 Kinematic Tree 确定关联组件. 为了尽可能涵盖组件的共现信息,我们分别提取双关联组件和三关联组件,后者对发现组件的远距离共现更有帮助. 假定训练集设定组件数为 10(下面均以 10 组件为基准),总共可提取 19(三关联)+14(双关联)种关联组合. 每种组合对应一组样本窗口将作为判别部件的搜索区域.

2) 学习判别部件. 此步自动搜索与学习判别部件,包括如下子步骤:

① 构建特征能量图. 以均值长宽比正规化目标样本集,使用已训练的目标分类器和组件分类器扫描样本窗口,得到每个样本的 HOG 特征能量图,这里能量定义为各分类器计算出的 HOG 值之和.

② 搜索判别部件. 对于第 1)步得到的关联组合,每种组合实际对应一个候选的判别部件(共有 19+14 个候选部件). 基于每种组合对应的样本窗口集,我们搜索每个样本的局部高能量区. 所有高能量区形成一个判别部件指定的正样本集.

③ 训练部件分类器. 对每个候选部件,通过 CP (clique-partitioning)算法聚类第②步的正样本集将得到该部件的子类型. 相比 k-means,CP 无须指定类别数,而是通过设定相似阈值控制类别的凝聚度,因而 CP 每个类别的元素个数差异较大.

$$D(P_1,P_2) = \lambda \lg D_{\text{spatial}}(P_1,P_2) + (1-\lambda)\lg D_{\text{appearance}}(P_1,P_2), \quad (12)$$

$$D_{\text{spatial}}(P_1,P_2) = \sum_c \| P_1(c) - P_2(c) \|^2 (1 + e_{1,2}(c)). \quad (13)$$

我们定义相似度式(12),其中 P_1 和 P_2 代表部件的 2 个窗口, $D_{\text{appearance}}$ 表示 2 个窗口的外观相似度,由式(11)计算. D_{spatial} 表示窗口所含组件的空间相似性,定义为 2 个部件对应组件的空间距离($P(c)=[x,y]$),e 设为惩罚值,若组件 c 同时存在于 2 个窗口,则 $e_{1,2}(c)=0$,否则赋值为某个指定的常数. 考虑到部件的主要作用在于刻画多个组件的共现信息,因而设定权重调节因子 $\lambda=0.7$. CP 的相似度阈值设为 0.1. 聚类完毕后每个部件形成 2～8 个不等的子类别,我们去除所含元素较少的类别,对剩下的每个类别采用常规 SVM 算法获取指定部件的子类型分类器,通过以上操作大约可得到 100～150 个子类型分类器,每个判别部件约有 2～5 个子类型.

3) 过滤子类型分类器. 本步骤目的是去除可信度低的部件分类器. 过滤方法类似于封闭式训练. 我们在训练样本上运行所有的部件分类器,对检测结果与步骤 1)提取的候选窗口进行比较,保留平均覆盖度大于 30%的部件. 接下来根据检测到的窗口集重新训练部件分类器,并重新扫描训练样本. 我们对检测结果进行统计,每个训练样本按检测得分排序候选部件,然后以贪婪方式选取覆盖样本数最多的前 n 个(默认 $n=30$)作为最后的输出,经过过滤和优化后的部件分类器具有足够的判别能力以检测相关组件.

3 实验与讨论

层次化模型的特点在于不仅能处理目标检测和姿态估计的联合识别,而且从单个任务的角度而言,能通过辅助信息增强识别率. 本节实验以当前主流的检测与估计模型为参照,评估层次化模型在联合任务和单任务上的识别性能.

3.1 联合识别性能

本节实验基于开放数据集 IP (iterator image parsing)①,IP 包含 305 张人体 Pose,每个 Pose 由 14 个关节点标注. 该数据集在单张图像中可能包含多个人体实例,因而也可用于目标检测的评测. 我们共选取 4 种联合模型,即 Torso-PS[9],Poselet-PS[10],IM[8] 和本文的层次化模型 HM. 联合识别场景并未指定目标位置,需要解决目标检测与各组件的识别双重任务,为此实验以 recall 和 FPPI(false positive per image)[17] 作为评估各组件的识别指标,FPPI 表示每个测试图像中容许的错误检测数,实验设定 FPPI 为 4. 正确检测组件的标准为 PCP(percentage of correct parts)大于 50%,正确检测目标的标准为候选检测窗口与标注窗口的覆盖率大于 0.5.

图 5 展示了各模型对组件与目标的检测结果. 各组件标识的名称解释如下:Torso—躯干、Head—头部、L-U-arms—左上臂、L-F-arms—左前臂、R-U-arms—右上臂、R-F-arms—右前臂、L-U-legs—左小腿、L-L-legs—左大腿、R-U-legs—右小腿、R-L-legs—右大腿、Part-Avg—组件平均准确率、Obj—目标检测率.

从图 5 可知,本文的层次化模型 HM 在大部分组件以及整体目标的检测率均优于其他 3 种联合模型,HM 在目标检测率上超过 Torso-PS 模型约 10%. Torso-PS 的串行识别方式致使躯干检测率限定了各组件识别性能的上限,换言之,躯干检测失败

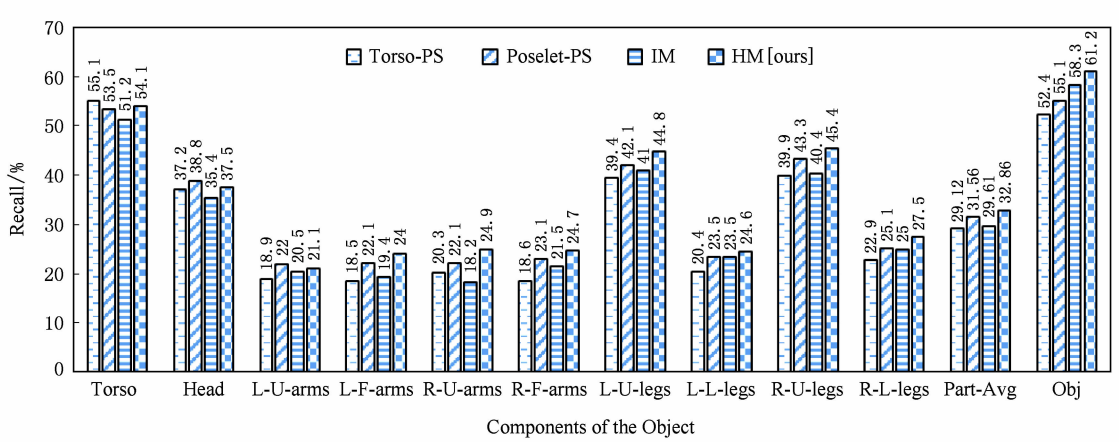


Fig. 5 The multi-task recognition performance of the evaluated methods when FPPI is 4.

图 5 当 FPPI 为 4 时联合识别方法的性能对比

① 参见 http://www.ics.uci.edu/~dramanan/papers/parse/index.html.

将直接影响目标(该模型实际是通过躯干位置指定目标位置)与其他组件的检测结果. Poselet-PS 同样基于串行识别方式,但通过 Poselets 增强了目标和组件的识别效果,因而在组件平均识别率(Part-Avg)和目标检测率方面均高于 Torso-PS. 由文献[11]可知,Poselets 的构建主要面向目标检测任务,因而 Poselet-PS 模型所使用的 Poselets 并不指向某个具体的组件,当定位单个组件时需要依据所有 Poselets 的打分. 若某些 Poselets 与指定组件无关,例如从腿部区域获取的 Poseles 对上臂的定位,这些 Poselets 对组件的识别不仅没有参考作用,甚至会干扰定位的正确性. 相比而言, IM 与 HM 模型整合组件信息作为目标检测的依据,因而在目标检测率方面均优于串行化模型. 但注意到 IM 并未提供对组件识别更有效的特征信息,因而其组件识别率与 Torso-PS 和 Poselet-PS 相仿;而 HM 借助中间层的判别部件融合组件共现特征,提高了组件的局部定位能力. 进一步分析认为,这些判别部件与 Poselets 的功能相似,但判别部件与组件为双向对应,即定位某个组件时 HM 仅考虑少量直接相关的部件,而不是所有判别部件.

串行化联合模型 Torso-PS 和 Poselet-PS 在第 1 阶段分别以线性时间找到候选的 Torso 和所有 Poselets,在第 2 阶段仍基于动态规划搜索最佳的组件格局,因而 2 个模型的整体复杂度为 $O(N_p^2 \times N_c^2)$,其中 N_p 表示部件数,N_c 为每个候选部件在图像中可能存在的位置数量. 并行化联合模型 IM 与 HM 的计算方式相同,只是 HM 的搜索结点数多于 IM,根据式(3)~(6),整体复杂度亦为 $O(N_p^2 \times N_c^2)$. 4 个评估模型尽管复杂度的量级相同,但实时环境中串行化模型的推断时间均高于并行化模型. 我们以 HP BL460 服务器(Intel 2.4 GHz 4 核,16 GB 内存)和 Matlab 2012a 作为运行环境,设定 10 级 scales 和 24 个 orientation,在 640×480 的图像中执行联合识别,4 个联合模型的平均时间① 分别为 Torso-PS:110.5 s,Poselet-PS:142.6 s,IM:65.2 s,HM:87.1 s. HM 由于待搜索结点多于 IM,因而运行时间相比较长一些;其次,在模型训练过程我们发现 Poselet-PS 的训练时间(达到 38.9 h)和储存开销均高于其他 3 个模型,主要原因在于 Poselets 的聚类构建过程花费巨大;HM 运行 CCCP 迭代计算,训练时间为 18.5 h,略高于 IM 的 15.3 h 训练时间.

3.2 单任务识别性能

本节分别在姿态估计和目标检测的传统评测场景下观察层次化模型的识别性能.

1) 姿态估计

我们使用 LSP(leeds sport pose)② 作为姿态估计的评测集. LSP 是当前专门针对姿态估计研究的大规模标注集,共包含 2 000 张图像. 每个人体实例标注 14 个关键点(同 3.1 节实验的 IP 数据集). 我们仍然采用 PCP 大于 50% 作为各组件识别正确的标准. 在给定目标窗口的条件下,Torso-PS 等价于经典的 PS 模型[3]. 对于 Poselet-PS,实验选取与目标标注窗口覆盖率大于 50% 的候选 Poselets 参与第 2 阶段的姿态估计. 图 6 展示 4 种联合模型的评估结果. 从平均定位率上,Poselet-PS 和本文 HM 模型虽优于其他 2 个模型,但并不显著(约 1%~2%),这说明开发组件间的共现信息,特别是远距离共现,对定位率具有积极作用. 由实验结果可知,诸模型在不同组件上表现各不相同,例如 Torso-PS 在 Torso 上表现最好,而 IM 在 L-U-arms 识别率最高;其次,各组件的检测率高低差异较大,例如 Torso 为 77%,R-F-arms 为 33%. 以上 2 点说明,对于单纯的姿态估计,目前的估计算法在特征开发与计算方法上均达到了相当的水平. 如何提高某些难处理组件(例如 L-F-arms,R-F-arms 等)的识别率是今后研究的一个挑战.

2) 目标检测

实验最后在传统的目标检测场景下评估了 2 个并行化识别模型 IM 和 HM. 我们使用带部件标注的 PASCAL VOC[9] 作为测试集,该数据集对 PASCAL VOC 2007/2010 中的 6 种非刚性目标类(horse,cow,cat,bird,dog,sheep)进行了部件层标注,每个目标类别含 6~9 个语义部件. 为更好地观察联合模型的检测性能,实验另选取了当前 2 个 state-of-the-art 检测算法 DPM[7] 和 S-DPM[14],这 2 个算法均采用部件化描述目标. DPM 的部件来自于全局能量搜索. S-DPM 是 DPM 的一个改进版本,通过标注的语义部件层:1)提供更精确的格局分类(即目标姿态分类);2)改进 DPM 粗糙的部件初始化(详情参见文献[14]). 实验评估指标为 AP(average precision),表 1 展示了各模型的目标检测率,表 1 中黑体数据代表对应列的最大值.

① 模型实际运行过程执行了广义距离转换(generalized distance transforms)[18],以降低部件格局的搜索空间.

② 参见 http://www.comp.leeds.ac.uk/mat4saj/lsp.html.

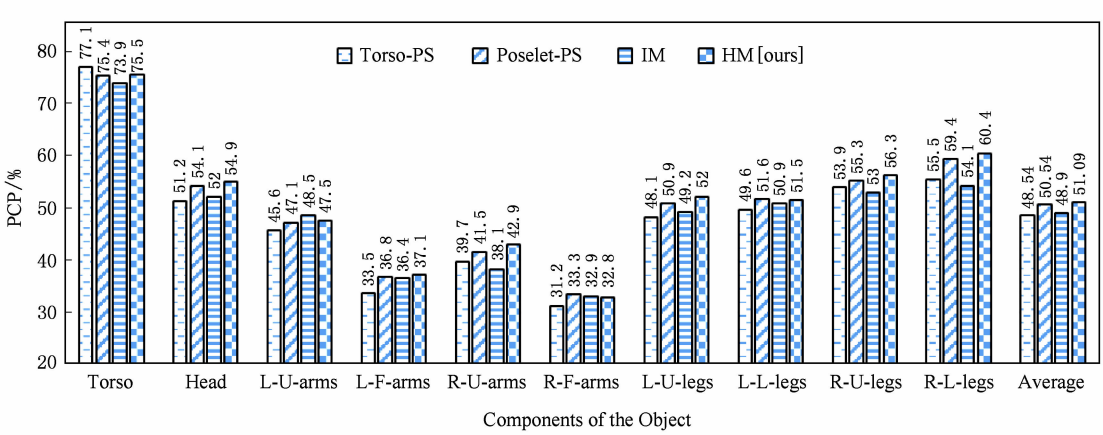


Fig. 6 The single-task recognition performance (i.e. pose estimation) of the evaluated methods.

图 6 单任务识别(即姿态估计)的性能对比

Table 1 Detection Performance of the Models

表 1 目标检测的性能对比

Model	PASCAL VOC 2007/PASCAL VOC 2010						
	Horse	Cow	Cat	Bird	Dog	Sheep	Average
DPM	57.0/41.1	25.4/21.0	25.4/21.0	9.9/9.5	11.1/10.3	17.9/26.6	23.43/21.83
S-DPM	63.1/48.7	34.9/26.0	**26.5/26.1**	12.4/11.3	**18.8/23.5**	**22.9**/27.9	**29.77/27.25**
HM	**65.8/50.5**	**35.2/25.8**	22.9/24.0	**13.7/11.6**	18.2/22.5	22.5/**28.2**	29.72/27.10
IM	58.9/48.2	27.7/23.6	23.0/21.6	11.0/10.8	13.1/22.9	22.4/30.4	26.02/26.25

表 1 展示的数据说明联合模型 HM 和 IM 相比经典的 DPM 均有较大改进,性能提升约为 3%和 6%.分析认为,联合模型能以组件混合搭配的方式描述大量而且精确的部件格局,这些部件格局能够反映目标内部的特征分布,非常适于处理因视角或局部形状变化引起的外观差异;其次,我们看到 HM 的检测性能比 IM 更高,IM 尽管能刻画目标的姿态,但目标直接由语义组件描述(参见图 1(c)).文献[13]的结论表明对单个组件有效的局部特征不一定对分类整体目标同样有效,因而语义组件并不适于直接描述目标的局部外观.HM 采用类似 DPM 的部件学习方式自动获得判别部件(参见 2.2 节),因而具有更好的检测效果.注意尽管 HM 与 S-DPM 具有相似的目标描述方式,但 S-DPM 稍优于 HM,我们认为其中一个重要原因在于部件学习方式的差异.S-DPM 在训练样本中搜索部件时增加了语义约束(即限定搜索的判别部件与标注的语义部件之间的覆盖率必须大于 0.3),此约束克服了 DPM 搜索部件时的盲目性,但仍是针对目标全局的搜索;而 HM 限定部件的搜索必须在关联组件的联合覆盖区域,这是为了权衡判别部件能同时对目标和组件均具有判别性,但却忽略了某些跨组件的目标特征,因而 HM 的判别部件在目标描述的能力方面弱于 S-DPM 的部件.

参考文献[7,14],DPM 与 S-DPM 均为树结构推断,因而 4 种检测模型的计算复杂度属于同一量级.但实时环境中 IM 和 HM 的结点数以及每个结点的子类型数均大于 DPM 与 S-DPM,例如 DPM 判别部件的子类型数为 2,而 HM 设定的子类型数为 4,因此层次化模型的状态搜索空间大于部件模型.以 3.1 节实验相同的测试环境为例,DPM 在 640×480 的图像中检测目标的平均时间为 35.1 s,而 HM 为 62.6 s.

4 总 结

当前目标检测与姿态估计分属不同的研究任务,本文阐述了联合 2 个任务的必要性与可行性,分析了当前一些联合算法的缺陷.在此基础上,本文对现有的部件模型进行改进,提出一种层次化的联合检测模型.该模型的主要优势在于并行化处理检测与估计,而且能平衡目标与组件在特征描述上的判别性,同时提升两者的识别性能.当前许多视觉任务均采用部件化建模方式,例如事件检测、基于属性的

目标识别等.我们下一步将扩大联合模型的应用范围,进一步研究其他任务之间的相关性,设计更有效的多任务处理算法.就姿态估计而言,目前各模型在总体检测率上已达到相当的水平,但某些组件识别率过低的问题仍未得到解决,我们将继续深化对估计模型的研究,在组件关联、特征描述、遮挡等方面提出更好的解决方法.

参 考 文 献

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] //Proc of the 18th IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 886-893

[2] Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition [J]. International Journal of Computer Vision, 2005, 61(1): 55-79

[3] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts [C] //Proc of the 24th IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 1385-1392

[4] Eichner M, Ferrari V. Appearance sharing for collective human pose estimation [C] //Proc of the 11th Asian Conf on Computer Vision. Berlin: Springer, 2013: 138-151

[5] Divvala S K, Efros A A, Hebert M. How important are "Deformable Parts" in the deformable parts model [C] //Proc of the 12th European Conf on Computer Vision, Workshops and Demonstrations. Berlin: Springer, 2012: 836-849

[6] Gu C, Ren X. Discriminative mixture-of-templates for viewpoint classification [C] //Proc of the 11th European Conf on Computer Vision. Berlin: Springer, 2010: 408-421

[7] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645

[8] Sun M, Savarese S. Articulated part-based model for joint object detection and pose estimation [C] //Proc of the 13th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2011: 723-730

[9] Pishchulin L, Jain A, Andriluka M, et al. Articulated people detection and pose estimation: Reshaping the future [C] // Proc of the 25th IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3178-3185

[10] Pishchulin L, Andriluka M, Gehler P, et al. Poselet conditioned pictorial structures [C] //Proc of the 26th IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 588-595

[11] Bourdev L, Maji S, Brox T, et al. Detecting people using mutually consistent poselet activations [C] //Proc of the 11th European Conf on Computer Vision. Berlin: Springer, 2010: 168-181

[12] Bourdev L, Malik J. Poselets: Body part detectors trained using 3D human pose annotations [C] //Proc of the 12th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2009: 1365-1372

[13] Chen Yaodong, Li Renfa. Learning parts for object detection [J]. Journal of Computer Research and Development, 2013, 50(9): 1902-1913 (in Chinese)
(陈耀东, 李仁发. 一种面向目标检测的部件学习方法[J]. 计算机研究与发展, 2013, 50(9): 1902-1913)

[14] Azizpour H, Laptev I. Object detection using strongly-supervised deformable part models [C] //Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2012: 836-849

[15] Ott P, Everingham M. Shared parts for deformable part-based models [C] //Proc of the 24th IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 1513-1520

[16] Yu C N J, Joachims T. Learning structural SVMs with latent variables [C] //Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 1169-1176

[17] Sun M. Technical report of articulated part-based model [OL]. 2011 [2014-11-28]. http://cvgl.stanford.edu/papers/sun_iccv11_iccv_supp_camera.pdf

[18] Felzenszwalb P F, Huttenlocher D. Distance transforms of sampled functions, CIS-TR-2004-1963 [R]. Ithaca, NY: Cornell University, 2004

Chen Yaodong, born in 1978. Received his PhD degree from Hunan University in 2014. Student member of China Computer Federation. His main research interests include computer vision, machine learning.

Li Renfa, born in 1957. Professor and PhD supervisor of Hunan University. Senior member of China Computer Federation. His main research interests include embedded systems, wireless sensor network and CPS (lirenfa@vip.sina.com).