JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2015
11
VOL.20
ISSN1006-8961
CN11-3758/TB

动作识别 P1462

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@radi.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

第20卷第11期（总第235期）
2015年11月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@radi.ac.cn
Telephone 010-64807995
Website www.cjig.cn

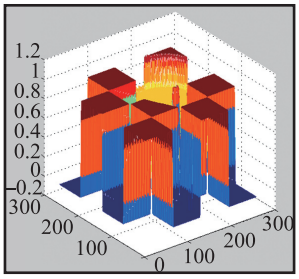
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

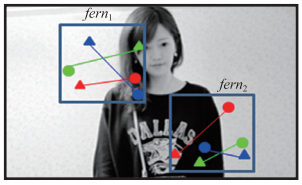
国外发行代号 M1406
国内邮发代号 82-831
国内定价 50.00元

2015年11月 第20卷 第11期（总第235期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

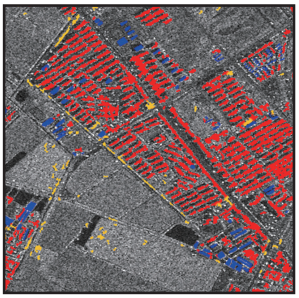
目次



对偶算法在紧框架域TV−L_1去模糊模型中的应用(第1434页)



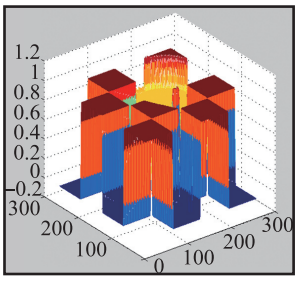
融合检测和跟踪的实时人脸跟踪(第1473页)



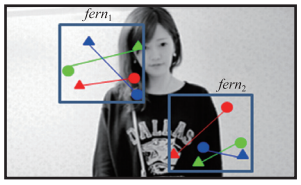
高分辨率SAR影像形态学层级分析的建筑物检测(第1517页)

Volume 20, Number 11
Published November 16, 2015

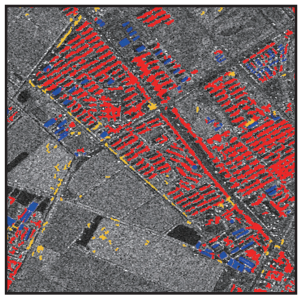
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Application of dual algorithm to TV-L_1 deblurring model of frame domain(第1434页)



Real-time face tracking based on detecting and tracking (第1473页)



Building detection of high-resolution SAR image based on morphological hierarchy analysis(第1517页)

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号：TP391　　**文献标识码**：A　　**文章编号**：1006-8961(2015)11-1462-11

论文引用格式：Luo H L,Feng Y J,Kong F S. Fusing multiple pose estimations for still image action recognition[J]. Journal of Image and Graphics, 2015,20(11):1462-1472.[罗会兰,冯宇杰,孔繁胜. 融合多姿势估计特征的动作识别[J]. 中国图象图形学报,2015,20(11):1462-1472.]
[DOI:10.11834/jig.20151105]

融合多姿势估计特征的动作识别

罗会兰[1]，冯宇杰[1]，孔繁胜[2]

1. 江西理工大学信息工程学院,赣州　341000
2. 浙江大学计算机科学技术学院,杭州　310027

摘　要：**目的**　为了提高静态图像在遮挡等复杂情况下的动作识别效果和鲁棒性,提出融合多种姿势估计得到的特征信息进行动作识别的方法。**方法**　利用已得到的多个动作模型对任意一幅图像进行姿势估计,得到图像的多组姿势特征信息,每组特征信息包括关键点信息和姿势评分。将训练集中各个动作下所有图像的区分性关键点提取出来,并计算每一幅图像中区分性关键点之间的相对距离,一个动作所有图像的特征信息共同构成该动作的模板信息。测试图像在多个动作模型下进行姿势估计,得到多组姿势特征,从每组姿势特征中提取与对应模板一致的特征信息,将提取的多组姿势特征信息分别与对应的模板进行匹配,并通过姿势评分对匹配值优化,根据最终匹配值进行动作分类。**结果**　在两个数据集上,本文方法与5种比较流行的动作识别方法进行比较,获得了较好的平均准确率,在数据集PASCAL VOC 2011-val上较其他一些最新的经典方法平均准确率至少提高近2%。在数据集Stanford 40 actions上,较其他一些最新的经典方法平均准确率至少提高近6%。**结论**　本文方法融合了多个姿势特征,并且能够获取关键部位的遮挡信息,所以能较好应对遮挡等复杂环境情况,具有较高的平均识别准确率。

关键词：动作识别;多姿势估计;模板匹配;遮挡

Fusing multiple pose estimations for still image action recognition

Luo Huilan[1], Feng Yujie[1], Kong Fansheng[2]

1. *School of information Engineering, Jiangxi University of Science and Technology, Ganzhou* 341000, *China*
2. *School of Computer Science and Technology, Zhejiang University, Hangzhou* 310027, *China*

Abstract: **Objective** To adapt to occlusion or other complex situations, an action recognition method is proposed which fuses multiple pose estimation features. **Method** Multiple pose features will be obtained using multiple action models. Each pose feature information includes key point positions and pose scores. Distinguishing key pointsare extracted from all train images and computing relative distances between point pairs. An action template is built using all features of the train images of the action. Multiple feature information consistent with multiple action templates are extracted from each test image from multiple pose features. Multiple features of the test image are-matched with the corresponding action template and then matched values are optimized using pose scores. **Result** The experimental results have shown that the average accuracy of the proposed method is approximately 2% better than some other state-of-the-art methods on VOC 2011-val set, and is ap-

收稿日期:2015-04-16;**修回日期**:2015-06-27

基金项目:国家自然科学基金项目(61105042,61462035);国家重点基础研究发展计划(973)基金项目(2010CB327900);江西省青年科学家(井冈之星)培养对象计划资助。

第一作者简介:罗会兰(1974—),女,教授,2008年于浙江大学获计算机科学与技术专业博士学位,主要研究领域为机器学习、模式识别。E-mail:luohuilan@sina.com

Supported by: National Natural Science Foundation of China(61105042, 61462035); National Basic Research Program of China(2010CB327900)

proximately 6% better than some other state-of-the-art methods on Stanford 40 actions set. **Conclusion** By fusing multiple pose features, the proposed method can adapt to occlusion and other complex situations and improve average recognition accuracy.

Key words: action recognition; multiple pose estimations; template matching; occlusion

0 引　言

静态图像中的动作识别是一个非常活跃的领域,它有很多潜在的应用,比如图像搜索,个人相册管理和人机交互等应用。静态图像的动作识别比视频动作识别更具挑战性,在视频中运动线索可以为区分不同的动作提供丰富的信息源,而静态图像的动作识别研究主要根据图像中人动作的姿势特征[1-4]或者结合人体部位和相关物体的空间结构[5-7]进行动作识别。

在根据姿势特征来进行静态图像动作识别的研究方面,文献[1]提出通过获取姿势信息,根据姿势的肢体弯曲程度和形状等潜在信息来判断这种姿势的动作。这种方法虽然将姿势和识别结合作为一个整体,即根据姿势信息直接进行动作分类,不涉及模板匹配等过程,但这种方法所采取的动作模型使得获取的姿势信息较少,在遮挡等复杂环境下识别准确率较低。文献[2]提出利用树形结构空间关系[8]和人体动作关键部位分割的方法[7]训练灵活的人体部位混合模型(FMP)对人的姿势进行估计,这种方法对简单动作姿势的估计具有良好的效果,提取的姿势信息比文献[1]相对较多,也更加准确。文献[3]利用基于 2.5 维图模板匹配的方法对各种动作进行识别。这种方法将测试图像提取的特征信息与模板的特征信息进行匹配,由于模板中包含了动作的各种姿势信息,这种方法对于同一动作不同姿势变化有更好的鲁棒性。文献[4]通过语义金字塔的方法(SP)对人的姿势进行描述,并结合视觉词袋[9-10]的方法对动作进行识别。这种方法结合不同的人体部位检测器[11-14]对人体部位进行检测,并用语义金字塔的方法对人体部位信息进行描述,由于结合了多种不同的部位检测器,这种方法能够获得较准确的姿势信息。文献[15]通过深度信息评估人体关键点的信息,利用相对关键点位置差对人体动作特征进行表达,最后计算测试序列和训练集运动序列 Hausdorff 距离进行相似性匹配,进而进行动作识别。文献[16]通过检测人体动作时变化显著的关键点,并把这些关键点作为区分动作类别的兴趣点,对兴趣点为中心的邻域内捕获的局部形状、外观和运动等特征进行描述,最后构建分类器并识别动作。

最近许多研究者利用人体与图像中的其他相关物体间的空间结构信息进行动作识别。文献[7]提出的视觉短语(VP)方法对静态图像进行动作识别,这种方法通过检测图像中的人和物,并判断他们的交互关系来进行动作识别。但是检测器的训练需要大量的训练数据来获取尽可能多的人与物的交互关系。文献[5]在文献[7]的基础上提出的关系模型(RP)对于姿势信息的获取具有了较大的提高,能获取更加详细姿势信息,也能够在遮挡等复杂情况下获得人的姿势信息。由于这种方法中每一个动作对应一个动作模型,因此要获取任意一幅图像的姿势信息要面临模型的选择问题。文献[6]通过扩展的部位模型(EPM)对图像中目标的外观特征进行描述。这种模型是一个由大量身体部位和相关物体部位组成的集合,通过具有区分性的部位信息对图像中的动作目标进行描述。这种方法对与物体交互的动作具有较高的识别准确率,对单纯的人体动作识别效果相对较差。

根据人的姿势进行动作识别时需要得到图像较准确的姿势信息,由于图像中遮挡情况的存在,造成了图像中人动作姿势的不完整。同时为了匹配测试图像的特征信息与模板信息,找到姿势特征信息的充分表达是关键。为了在遮挡等复杂环境下还能有效进行动作识别,本文提出了融合多种姿势信息和特征的基于姿势估计和样本匹配的动作识别思想。首先,为了得到图像在不同姿势预测下的特征信息,从而融合多种信息进行综合考虑,提出利用多种姿势模型对图像分别进行姿势估计,得到多组特征信息。动作模板的建立是通过将训练集中各个动作下所有图像的区分性关键点提取出来,并计算每一幅图像区分性关键点之间的相对距离,一幅图像的信息为该动作的一条模板信息,所有图像的多条信息

共同构成该动作的模板信息。将测试图像得到的多组特征信息分别与相应模板中得到的特征信息进行匹配。最后利用姿势评分对匹配结果进行优化,将与测试图像最佳匹配的模板的动作类作为动作的分类结果。

1 方法的提出

本文提出融合多个姿势估计得到的特征信息进行动作识别的方法(FMPE)。为了对任意一幅图像进行姿势估计而不需要考虑模型选择问题。例如:跑动作在跑模型下能够得到更加准确的姿势估计信息,但是对于任意一幅不知动作的图像,就无法使用确切的模型对其姿势估计。因此提出融合多姿势估计的方法。即一幅测试图像在多个模型下对其进行姿势估计,得到测试图像的多组姿势特征,这些特征包括每个关键点的坐标、遮挡标记以及对每个姿势的评分。动作模板的建立是提取训练集中各个动作下所有图像中目标的具有区分性关键点的坐标位置、遮挡标记,并计算区分性关键点之间的距离,归一化后作为此类动作模板,每个动作类建立一个模板。最后将测试图像得到的特征信息与所有动作模板进行匹配并通过姿势评分对匹配结果进行优化。本文提出的姿势估计是一幅图像在不同模型下的姿势估计,这样可以得到一幅图像在各个动作模型下的特征信息和评分,即一幅图像可得到多组特征信息。测试图像在不同动作姿势估计下提取的不同特征信息与相应不同动作下的模板信息进行匹配,根据最终匹配值的最小值决定测试图像的分类结果。由于不同的动作姿势估计下提取的关键点个数不同,关键点间的关系也不同,所以在决定最终分类结果前,需要对每种模板匹配下得到的匹配值进行规范化,使得不同动作模板下得到的匹配值具有可比性,以提高动作识别的准确率。

1.1 姿势估计

使用在 PASCAL VOC 2011-train 数据集上已经训练得到的多个动作模型[5]对图像中的动作进行姿势估计,以估计得到的关键部位特征作为图像后续动作分类的特征值,从而可以得到同一副图像的多个特征表达。利用这些不同特征表达进行后续的模板匹配,从而旨在得到更优的动作分类效果。实验中采用了在 7 个不同的动作模型下对图像中的动作进行姿势估计,分别是跑、走、跳、骑自行车、骑马、拍照和打电话。

姿势估计是根据动作模型在图像中检测到与模型相匹配的动作,然后对其姿势进行估计。图像姿势估计可以得到关键点的坐标位置,遮挡标记,以及估计评分值。对于姿势估计得到的关键点结构,文献[8]提出了空间结构框架,这种框架的思想是:将一个对象整体分割成各个关键部位,以及部位之间的几何约束,将各关键部位像素的位置和方向进行参数化,最终将整体结构模拟出来。文献[2]提出的模型是在此基础上建立的,设 I 为一图像,通过姿势检测得到 K 个关键部位(也称为关键点),关键部位 i 的坐标为 $\boldsymbol{p}_i=(x_i,y_i)$,$i\in\{1,\cdots,K\}$,$t_i$ 为关键部位 i 的类型,假设总共有 T 种类型,$t_i\in\{1,\cdots,T\}$。为了给局部关键部位结构进行评分,定义评分函数为

$$S_1=\sum_{i\in\boldsymbol{V}}b_i^{t_i}+\sum_{ij\in\boldsymbol{E}}b_{ij}^{t_i,t_j} \tag{1}$$

式中,$b_i^{t_i}$ 表示关键部位 i 为特定类型 t_i 的分配值,分配值大小表示关键部位 i 和特定类型 t_i 相符程度。$b_{ij}^{t_i,t_j}=1-\varepsilon_{ij}$,$b_{ij}^{t_i,t_j}$ 表示关键部位 i 和关键部位 j 共现的分配值,代表了关键性部位之间关联程度,$\varepsilon_{ij}>0$,表示惩罚值,当关键部位 i 和关键部位 j 的类型与当前模型匹配程度较高时,ε_{ij} 越小,反之越大。$b_{ij}^{t_i,t_j}$ 值越大表示关联程度越高。共现是指检测到图像中关键部位 i 和关键部位 j 同时出现。假设 K 个关键点组成的关系图为 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,$\boldsymbol{V}$ 为图中的关键点,$\boldsymbol{E}$ 代表节点间的边。$\boldsymbol{G}$ 是以树形结构将人体各关键部位分为父节点和子节点构成的人体姿势整体结构图,表示关键点之间的约束关系。S_1 用于评价局部结构的好坏。文献[5]在此基础上,对 S_1 进行了少许的修改,去掉了单独的类型分配值,即

$$S_2=\sum_{ij\in\boldsymbol{E}}b_{ij}^{t_i,t_j} \tag{2}$$

由于 Desai 等人引入了与动作相关的物体内容,因此关键点不止人体的各关键部位,还包括物体上的一些关键部位,例如车把、车轮等关键部位。

用于评价检测出的姿势结构与模型的相符程度,或者说检测出的姿势结构好坏程度的评分函数为

$$S_3=S_2+\sum_{i\in\boldsymbol{V}}w_i^{t_i}\cdot\alpha(\boldsymbol{I},\boldsymbol{p}_i)+\sum_{ij\in\boldsymbol{E}}w_{ij}^{t_i,t_j}\cdot\beta(\boldsymbol{p}_i-\boldsymbol{p}_j) \tag{3}$$

式(3)所示的评分函数是在式(2)的基础上产生的,表示人体姿势整体结构以及关键部位之间关系的分配值,值越大表示人体姿势与当前模型越相符。

式中,$\boldsymbol{\alpha}(\boldsymbol{I},\boldsymbol{p}_i)$是从图像 I 的像素位置 $\boldsymbol{p}_i$ 处提取的特征向量,例如方向梯度图(HOG)特征,$\beta(\boldsymbol{p}_i-\boldsymbol{p}_j)=[dx \quad dx^2 dy \quad dy^2]^{\mathrm{T}}$,$dx=x_i-x_j$,$dy=y_i-y_j$,$(x_i,y_i)$表示关键部位 i 的坐标位置,$w_i^{t_i}$ 表示在 $\boldsymbol{p}_i$ 处将关键部位 i 的类型调整为 t_i 的一个编码值,$w_{ij}^{t_i,t_j}$表示关键部位 i 和 j 分别调整为 t_i 和 t_j 的一个编码值。$S_2+\sum_{i\in\boldsymbol{V}} w_i^{t_i}\cdot\alpha(I,\boldsymbol{p}_i)$ 表示外观关系,$\sum_{ij\in\boldsymbol{E}} w_{ij}^{t_i,t_j}\cdot\beta(\boldsymbol{p}_i-\boldsymbol{p}_j)$ 表示空间关系。

目标检测是利用动作模型通过滑动窗口的形式进行检测,并通过非极大值抑制算法(NMS)产生多个检测目标。检测过程中利用了关键点周围的 HOG 特征信息。由于模型的训练采用了人体部位分割的方法[8]进行训练,因此可得到目标各关键部位的特征信息。针对每一个目标的姿势估计可以根据式(3)得到一个评分,选择评分最高的姿势作为在此模型下检测得到的最佳姿势估计。如图 1 所示,一幅图像在跑模型下对图像进行姿势估计,得到 3 个目标姿势估计,每个姿势估计得到一个评分。图 1(b)中从左到右目标姿势估计评分分别为 -0.929 4、-0.870 9、-0.991 9,选择评分最高的中间目标姿势估计作为在此模型下得到的最佳姿势估计,利用此最佳姿势估计得到图像的一组特征表达,如图 1(c)表示。设一幅图像在 m 模型下检测得到的最佳姿势估计的关键点位置表示为 $\boldsymbol{P}_i^m=(x_i^m,y_i^m)$,$i\in\{1,2,\cdots,K^m\}$,$m$ 表示动作模型,K^m 是在模型 m 下得到最佳姿势估计的关键点个数。最佳姿势的评分表示为 S_m,一幅图像在 m 模型下的最佳姿势估计得到的特征表示为$\{\boldsymbol{P}_1^m,\cdots,\boldsymbol{P}_{K^m}^m,S_m\}$此模型估计下得到的特征维度为 K^m+1。图 2 所示为一幅图像分别在跑、走、跳、骑自行车、骑马、拍照、打电话 7 个模型下得到的 7 组最佳姿势估计,通过这 7 组最佳姿势估计可以得到一幅图像的 7 种特征表达。

(a) 原图像　(b) 多个目标的姿势估计　(c) 最佳姿势估计

图 1　一个动作模型对人动作的姿势估计

Fig. 1　The pose estimation using an action model

(a) 跑模型　(b) 走模型　(c) 跳模型

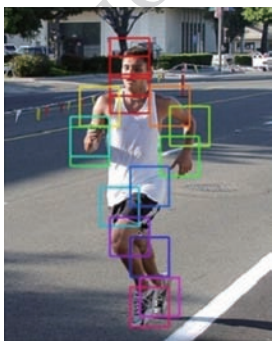
(d) 骑自行车模型　(e) 骑马模型　(f) 拍照模型　(g) 打电话模型

图 2　在多个模型下得到的最佳姿势估计

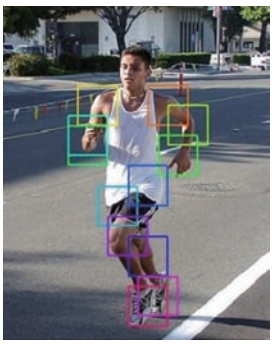
Fig. 2　The pose estimationsusing multiple action models

1.2 模板的建立

为了实现动作识别,需要将测试图像中的特征信息与模板特征信息进行匹配。采用各个动作下所有训练图像中关键点之间的相对距离作为动作模板。首先将每类动作下的训练图像中人动作的具有区分性的关键点提取出来。区分性的关键点表示对动作影响较大的关键点。由于不同的动作姿势,具有不同的关键点空间结构,但有些关键点对动作的区分性不大,例如动作跑和动作走中,脸部关键点对动作区分性没有影响,但是四肢上的关键点对动作的区分性有较大的影响。为了尽可能保证动作之间的区分性,因此选取区分性关键点构建模板。如图3(a)表示跑动作的训练图像所包含的关键点位置,而图3(b)表示跑动作四肢上的12个关键点,即构建跑模板所需要的区分性关键点。然后计算所提取关键点之间的相对距离,最后把同一类动作下的所有训练图像的区分性关键点之间的相对距离作为此类动作的模板。



(a) 训练集中跑动作所包含的关键点



(b) 创建跑模板所需的区分性关键点

图3 训练集图像中选取区分性关键点

Fig. 3 Selecting distinguishing key points from a train image

假设训练集中动作类 A 下的一幅图像 I,它的一个关键点 i 位置表示为 $\boldsymbol{P}_i^A=(x_i^A,y_j^A)$,$i\in\{1,2,\cdots,K^A\}$,所提取的区分性关键点个数为 K^A,图像 I 中区分性关键点表达式为 $\{\boldsymbol{P}_1^A,\cdots,\boldsymbol{P}_{K^A}^A\}$。图像 I 中关键点 i、j 之间归一化后的相对距离为

$$\Delta d_{i,j}=\frac{\sqrt[2]{(x_i-x_j)^2+(y_i-y_j)^2}}{a} \tag{4}$$

式中,a 表示图像 I 中关键点相对距离的平均值,即

$$a=\frac{\sum\limits_{i,j\in K^A}\sqrt[2]{(x_i-x_j)^2+(y_i-y_j)^2}}{K^A(K^A-1)/2}$$

则图像 I 中关键点间相对距离表示为 $\{\Delta d_{1,2},\cdots,\Delta d_{K^A,(K^A-1)}\}$,维数为 $K^A(K^A-1)/2$。把动作 A 中 N 幅训练图像的关键点相对距离存储起来作为动作 A 的模板,即

$$\boldsymbol{T}=\begin{bmatrix}\Delta d_{1,2}^1 & \cdots & \Delta d_{K^A,(K^A-1)}^1\\ \vdots & & \vdots\\ \Delta d_{1,2}^N & \cdots & d_{K^A,(K^A-1)}^N\end{bmatrix} \tag{5}$$

式中,$\Delta d_{i,j}$ 表示关键点 i,j 归一化后的相对距离。$\boldsymbol{T}$ 表示动作 A 的模板特征信息。

1.3 最优匹配

最优匹配是将测试图像通过姿势估计并计算与模板一致的特征信息,将测试图像得到的特征信息与模板进行匹配,将匹配后的结果通过姿势评分进行优化。由于一幅测试图像在多个模型下得到多组关键点信息,因此关键点之间的相对距离需要计算多组。提取与模板一致的具有区分性的关键点信息,关键点个数也为 K^A,每一组为一个 $K^A(K^A-1)/2$ 维的矢量特征,与相应的模板一一对应。一幅测试图像在 m 模型下进行姿势估计可得到关键点坐标位置以及姿势评分 $\{\boldsymbol{P}_1^m,\cdots,\boldsymbol{P}_{K^m}^m,S_m\}$,提取具有区分性的关键点坐标位置以及评分 $\{\boldsymbol{P}_1^m,\cdots,\boldsymbol{P}_{K^A}^m,S_m\}$,根据式(4)可得到测试图像关键点之间相对距离的特征表达 $\boldsymbol{t}=\{\Delta d_{1,2},\cdots,\Delta d_{K^A,(K^A-1)}\}$,为 $K^A(K^A-1)/2$ 维,在多个模型下可得到多组这样的关键点特征表达。每组与式(5)所示的相应模板进行匹配,测试图像的特征信息与模板信息的差异值为

$$D_m^{\text{old}}=\min_{i=1\cdots N}\left\{\frac{\sum\limits_{j=1}^{K^A}{}^{(K^A-1)/2}(t_j-T_{i,j})}{K^A(K^A-1)/2}\right\} \tag{6}$$

根据式(6),一幅测试图像的多组特征表达分别与相应模板进行匹配,从而可得到多个匹配值。D_m^{old} 越小,表示测试图像与模板越匹配。假

设一幅测试图像在 m 模型下进行姿势估计可得到姿势评分 S_m,匹配的初步结果可表示为$\{D_m^{\text{old}}, S_m\}$。

一幅图像的动作识别需要提取多组特征信息与相应的模板进行匹配,由于不同模板或姿势估计所得到的特征间有较大差别,从而得到的匹配值也有很大的差别,如果直接进行比较,会出现图像在有的动作模型下得到的姿势评分 S_m 很高,即对测试图像的估计符合相应的动作模型,但是匹配结果较差。所以为了更好地比较不同姿势估计下的匹配结果,提高动作识别准确率,需要调整参数对结果进行优化。根据每种姿势估计下得到的姿势评分 S_m 进行匹配值的调整,提出在姿势估计中,当姿势评分 $S_m > -0.9$时,表示有较好的姿势估计,使用一个小于 1 的参数 p_m^1 进行匹配值的调整,使 D_m^{old} 变小。当姿势评分 $-0.95 < S_m < 0.9$,姿势估计的姿势一般,使用一个大于 1 的参数 p_m^2 进行匹配值调整,使 D_m^{old} 变大。$S_m < -0.95$ 表示较差的姿势估计,使用一个大于 p_m^2 的参数 p_m^3 进行匹配值的调整,使 D_m^{old} 变得更大。D_m^{old} 表示模型 m 下得到的估计信息与相应模板匹配得到初始的匹配值,D_m^{new} 表示经过参数调整处理后的匹配值。*result* 为 D_m^{new} 中的最小值,最小匹配值对应的动作类作为测试图像的动作分类结果,匹配值调整过程如下所示:

$$\text{If } S_m > -0.9$$
$$D_m^{\text{new}} = D_m^{\text{old}} p_m^1$$
$$\text{If } S_m > -0.95 \;\&\&\; \text{If } S_m \leqslant -0.9$$
$$D_m^{\text{new}} = D_m^{\text{old}} p_m^2$$
$$\text{If } S_m \leqslant -0.95$$
$$D_m^{\text{new}} = D_m^{\text{old}} p_m^3$$
$$result = \min(D_m^{\text{new}})$$

1.4　算法描述及流程图

本文提出的融合多姿势估计特征的动作识别方法 FMPE 的主要思想是:结合 RP[5] 姿势估计模型,得到图像的多组姿势估计特征。将训练集中多个动作下所有图像中区分性关键点之间的相对距离特征信息作为模板。从测试图像估计得到的特征中提取与模板一致的区分性关键点,并计算区分性关键点之间的相对距离,然后与模板进行匹配,根据匹配值进行动作分类。本文方法 FMPE 的动作模板的创建流程图如图 4 所示,选取某动作下 N 幅图像,每一幅图像提取该动作下的区分性关键点,并计算区分性关键点之间的相对距离。将 N 幅图像得到的区分性关键点之间的相对距离作为模板保存;选择 M 个动作,最终得到 M 个模板。测试图像动作识别流程图如图 5 所示,输入任意一幅测试图像 I,分别在 M 个模型下得到测试图像的 M 组关键点信息以及姿势评分。动作模型是利用文献[5]已训练好的模型。一幅测试图像在一个动作模型下可得到图像中

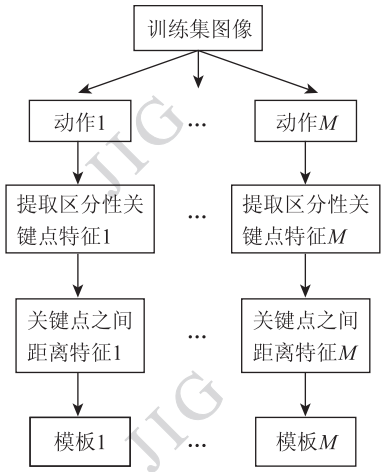


图 4　模板建立流程图

Fig. 4　The flow chart for building action templates

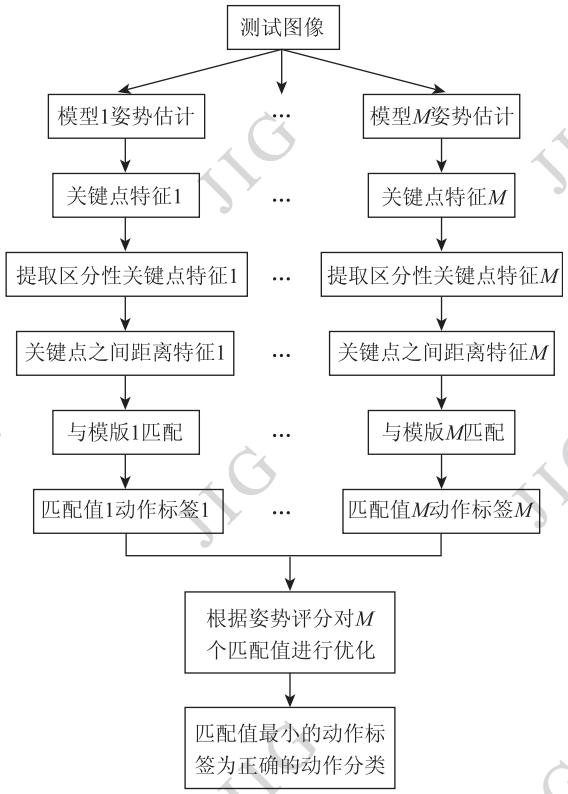


图 5　FMPE 算法流程图

Fig. 5　The flow chart of FMPE

人体动作的一个姿势估计结果,估计结果包括关键点位置、遮挡标记和姿势评分。在 M 个模型下分别对测试图像进行姿势估计,可得到测试图像的 M 个姿势特征,即可得到测试图像的 M 组关键点信息,分别表示为 $\{f_1,S_1\}\cdots\{f_M,S_M\}$。选取每组中与模板对应的区分性关键点,并计算区分性关键点之间的相对距 $\boldsymbol{d}$,$\boldsymbol{d}$ 为多维向量。测试图像的 M 组关键点信息最终得到 M 组相对离特征 $\boldsymbol{d}_1,\cdots,\boldsymbol{d}_M$,将 M 组区分性关键点相对距离特征分别与 M 个已构建好的模板进行匹配,选取每组匹配的最小值作为初始匹配值,分别为 $D_1^{\text{old}},\cdots,D_M^{\text{old}}$,对应的动作标签分别为 $A_1,\cdots,A_M$,通过姿势评分 $S_1,\cdots,S_M$ 对 M 个初始匹配值调整优化,将得到 M 个新的匹配值 $D_1^{\text{new}},\cdots,D_M^{\text{new}}$,选取其中最小匹配值的动作标签作为正确的动作分类。

2 实验结果与分析

为了验证动作识别的准确性,实验采用目前较常用的两种数据集:PASCAL VOC 2011-val set 和 Stanford 40 actions。在这些数据上验证本文方法 FMPE 动作识别准确率。

2.1 实验数据集及参数设置

PASCAL VOC 2011 数据集包括训练集和测试集,分别为 PASCAL VOC 2011-train 和 PASCAL VOC 2011-val set,各包含 10 个动作类,在实验中选择其中 7 个动作类,各个动作类图像情况如表 1 所示。该数据集特点是类内差异较大,图像中可能包括多个动作类目标,存在目标部分被遮挡的情况。模板的构建是在训练集 PASCAL VOC 2011-train 上进行的。本文 FMPE 方法与 3 种比较流行的动作识别方法在 PASCAL VOC 2011-val set 集上进行动作识别准确率的比较。3 种参与比较的动作识别方法分别是灵活的人体部位混合模型(FMP)[2]、视觉短语(VP)[7]、关系模型(RP)[5]。

Stanford40 actions 数据集包括 40 个动作类,在实验中只考虑其中的测试集,在测试集中选择与本实验中一致的跑、骑马、跳、骑自行车、拍照和打电话 6 个动作类进行实验,各个动作类图像如表 2 所示。选择两种最新的动作识别方法在此数据上与本文提出的方法 FMPE 进行动作识别准确率的比较,这两种方法分别是 SP[4] 和 EPM[6]。

由于本文采用的模型对骑马等和物体交互的动作的姿势估计具有更好的表现,更容易得到较高的姿势评分,但是这两个动作的关键点相对距离与模板的匹配达不到理想的效果。因此当 7 组姿势评分中,如果骑马姿势评分较高,通过更小的参数对骑马和骑自行车的匹配值进行调整,以提高姿势评分对骑马和骑自行车动作识别的影响,这样能够提高骑马和骑自行车的动作识别准确率。实验中当骑马的姿势评分 $S_{\text{rindinghorse}}>-0.9$ 时,匹配结果 $D_{\text{rindinghorse}}^{\text{old}}\times 0.6$,而对于 $S_{\text{walking}}>-0.9$ 时,$D_{\text{walking}}^{\text{old}}\times 0.8$。参数设置的目的是尽可能保证姿势评分对骑马和骑自行车的影响度。

区分性关键点是根据动作属性进行设定,在本实验中跑动作、走动作和跳动作的区分性关键点为四肢上的总共 12 个关键点。打电话和拍照的区分性关键点为头部以及两个胳膊上的 6 个关键点,共 7 个区分性关键点。骑马的区分性关键点为左右肩、左右手、左右臀、左右脚、马头、马尾和马脖子 11 个关键点。骑自行车的区分性关键点为左右肩、左右手、左右脚、左右臀、车把和前后轮 11 个区分性关键点。区分性关键点的设置是为了提高类与类之间的区分性。

表 1 PASCAL VOC 2011 动作数据集

Table 1 PASCAL VOC 2011 action set

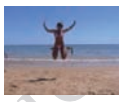
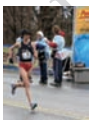
	跑	走	跳	骑自行车	骑马	拍照	打电话
训练集图像数量	157	164	122	153	141	116	113
测试集图像数量	158	165	122	153	142	117	114
示例图像							

表 2　Stanford 40 actions 测试集

Table 2　Stanford 40 actions test set

	跑	跳	骑自行车	骑马	拍照	打电话
测试集图像数量	151	195	193	196	97	159
示例图像						

2.2　动作识别及结果分析

本文算法 FMPE 能够适应存在遮挡等复杂情况下的动作识别。图 6 和图 7 分别是从 PASCAL VOC 2011-val set 中的骑马动作和跑动作中选择的环境较为复杂的两幅图像。两幅图像都出现了多个动作目标和遮挡的情况。图 6 是骑马动作在 7 个模型下的姿势估计结果,图像中有多个目标,骑马动作相对其他动作更加明显。通过多个模型对图像中的动作进行姿势估计,可以看出图 6(f) 中的姿势估计更加准确。图 6(f) 中目标的左腿被马遮挡,但是仍然能够将臀部、膝盖和脚 3 处的关键点估计出来。将图 6(b)—(h) 的特征信息分别与相应不同动作模板匹配得到的匹配值分别为 0.016 7、0.017 6、0.037 2、0.023 0、0.007 7、0.129 1 和 0.043 8,图 6(f) 对应的值最小,即识别为骑马动作。图 7 中出现了多个目标,图 7(b) 中膝盖被遮挡,但是仍然能够将被遮挡的膝盖标记出来。图 7(b)—(h) 特征信息分别与 7 种不同模板匹配得到的匹配值分别为 0.011 4、

(a) 原图像　(b) 跑模型　(c) 走模型　(d) 跳模型

(e) 骑自行车模型　(f) 骑马模型　(g) 拍照模型　(h) 打电话模型

图 6　骑马动作在不同模型下姿势估计结果

Fig. 6　The pose estimation results of a ridinghorse imageusing different models

(a) 原图像　(b) 跑模型　(c) 走模型　(d) 跳模型

(e) 骑自行车模型　(f) 骑马模型　(g) 拍照模型　(h) 打电话模型

图 7　跑动作在不同模型下姿势估计结果

Fig. 7　The pose estimationresults of a running image using different models

0.022 8、0.034 5、0.018 5、0.019 7、0.101 9 和 0.042 3，图 7(b)对应的值最小，即识别为跑动作。本文算法由于综合考虑了不同模型的姿势估计信息，能够更好地适应存在遮挡等复杂情况下的动作识别。

表 3 是本文方法 FMPE 与 VP[7]、FMP[2] 和 RP[5]3 种方法在 PASCAL VOC 2011-val set 数据集上的动作识别准确率比较结果。VP、FMP 和 RP 的动作识别准确率来自文献[5]。本文方法 FMPE 的平均识别准确率为 57.1%，比 VP 的平均准确率提高了近 14 个百分点，比 RP 方法平均准确率提高了近两个百分点。本文方法 FMPE 在跑、走和打电话 3 个动作的识别准确率是四种方法中最高的，与其他方法的识别准确率相比，跑动作至少提升了 5 个百分点，走动作至少提升了 10 个百分点，打电话动作提升了至少 15 个百分点，骑马动作基本与 RP 的识别准确率持平，动作骑自行车、拍照的准确率较 RP 方法次好。

表 4 是本文方法 FMPE 与 SP[7]，EPM[2] 两种方法在 Stanford 40 actions 数据集上的动作识别准确率比较结果，EPM、SP 的平均识别准确率来自文献[4]。本文方法 FMPE 的平均识别准确率为 59.1%，比 EPM 的平均准确率提高了近 17 个百分点。比 SP 的平均准确率高了近 6 个百分点。

表 3　PASCAL VOC 2011-val set 数据集上动作识别准确率对比

Table 3　Action recognition accuracy comparisons on PASCAL VOC 2011-val set

/%

动作	DPM/VP	FMP	RP	FMPE
running	**63.2**	**62.5**	**69.0**	**74.1**
walking	32.1	29.0	40.3	**49.7**
jumping	**28.8**	**44.2**	**49.6**	43.4
ridingbike	**66.4**	**66.9**	**81.7**	**71.2**
ridinghorse	79.7	**84.7**	**90.3**	**90.1**
takingphoto	12.1	11.7	**24.3**	**23.9**
phoning	21.2	21.3	32.9	**47.3**
平均准确率	43.3	45.8	**55.4**	57.1

注：粗体数字代表识别准确率最高，带下划线数字代表识别准确率次高。

表 4　Stanford 40 actions 数据集上动作识别准确率对比

Table 4 Action recognition accuracy comparisons on Stanford 40 actions

方法	EPM	SP	FMPE
平均识别准确率/%	**42.2**	53.0	**59.1**

注：粗体数字代表识别准确率最高。

图 8 为本文方法在 PASCAL VOC 2011-val set 中 7 类动作识别结果，从图 8(a)(b)可以看出，跑动作的识别误差很大一部分是因为误识为走，而走也有一部分误识为跑，这两类动作仅凭单幅图像确实有时很难区分。图 8(f)中拍照动作一部分识别为打电话。虽然动作之间的相似性影响了对动作识别的准确率，但是本文方法 FMPE 在跑动作、走动作、打电话动作上的平均识别准确率较其他方法有了一定的提升。

3　结　论

基于姿势估计和样本匹配，提出了一种利用多个模型进行姿势检测，从而得到多个特征表达的动作识别方法。在公开数据集 VOC 2011 中提取信息并将其作为模板。在特征提取阶段，根据姿势估计得到的信息，将各个关键点的坐标以及姿势评分提取出来，并计算特定关键点之间的相对距离，进行归

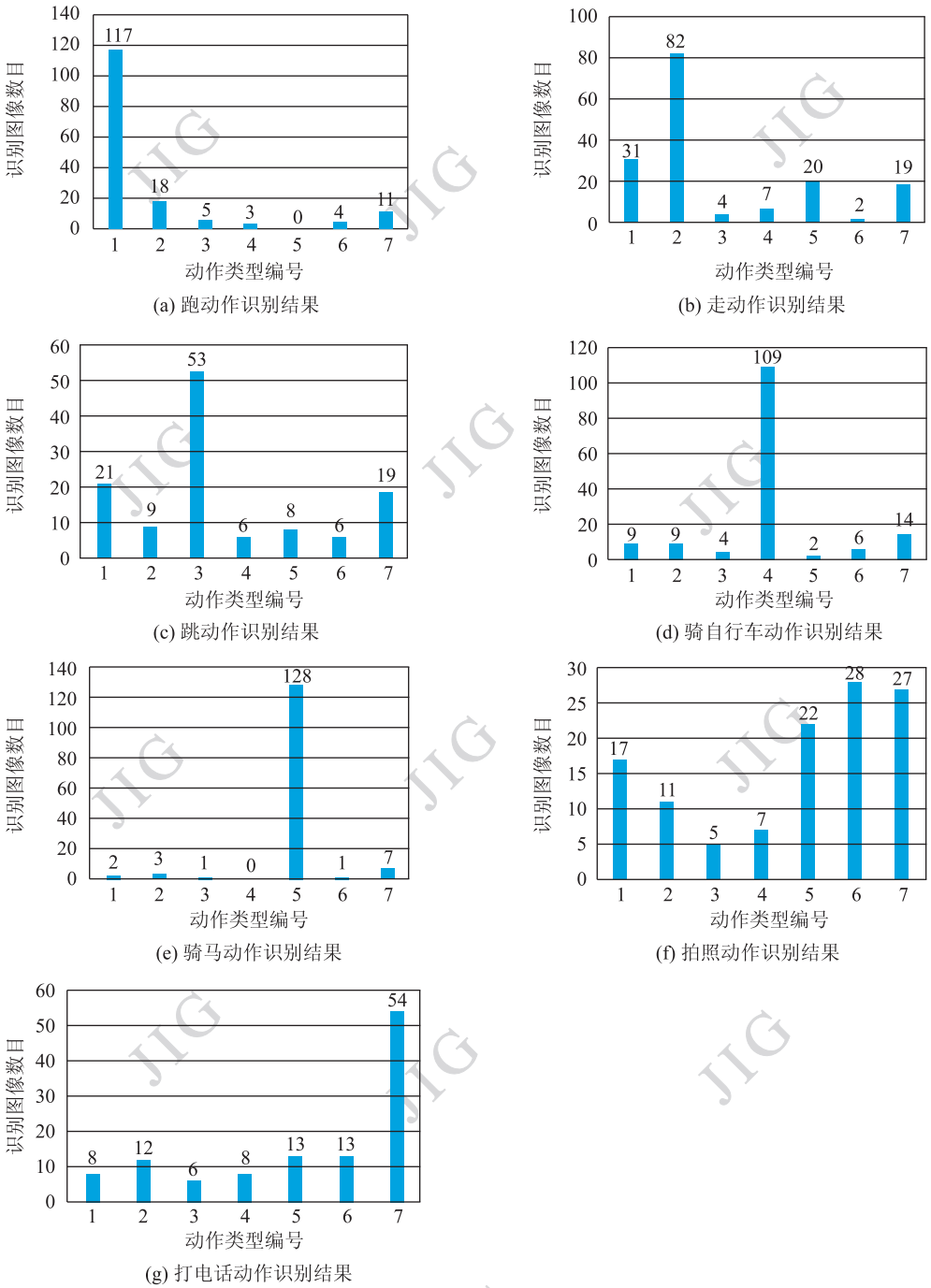


图 8　PASCAL VOC 2011-val set 中 7 类动作识别结果

Fig. 8　The recognition results of seven actions on PASCAL VOC 2011-val set

一化后与模板进行匹配。通过得到的姿势评分对结果进行优化,以提高动作识别的准确率。实验结果表明,本文方法在识别难点如遮挡、环境复杂等情况下表现良好。本文采用的模型不是统一的模型,而是每一个动作针对一个模型,这样姿势估计会更加准确,能得到更准确、更多的特征信息,从而提高动作识别的准确率。

由于本文的动作识别准确率与姿势估计的准确率直接相关,另外由于本文提出的融合多姿势估计的方法需要对图像进行多次姿势估计,随着动作类型的增加,会减低运行效率,下一步将对其进行改进,进一步提高动作识别的准确率和运行效率。

参考文献(References)

[1]　Yang W, Wang Y, Mori G. Recognizing human actions from still

images with latent poses[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA: IEEE, 2010:2030-2037.

[2] Yang Y, Ramanan D. Articulated pose estimation with flexiblemixtures-of-parts[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO: IEEE, 2011:1385-1392.

[3] Yao B, Fei-Fei L. Action recognition with exemplar based 2.5 D graph matching[C]//Proceedings of the 12th European Conference on Computer Vision. Berlin: Springer, 2012: 173-186.

[4] Khan F S, Weijer vandeJ, Rao M A, et al. Semantic pyramids for gender and action recognition[J]. IEEE Transactions on Image Processing, 2014, 23(8):3633-3645.

[5] Desai C, Ramanan D. Detecting actions, poses, and objects with relational phraselets [C]//Proceedings of the 12th European Conference on Computer Vision. Berlin: Springer, 2012: 158-172.

[6] Sharma G, Jurie F, Schmid C. Expanded parts model for human attribute and action recognition in still images[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR:IEEE, 2013: 652-659.

[7] Sadeghi M A, Farhadi A. Recognition using visual phrases[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO:IEEE, 2011:1745-1752.

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

[9] Khan F S, Anwer R M, Weijer van de J, et al. Coloring action recognition in still images[J]. International Journal of Computer Vision, 2013, 105(3):205-221.

[10] Delaitre V, Laptev I, Sivic J. Recognizing human actions in still images: a study of bag-of-features and part-based representations[C]//Proceedings of the 21st British Machine Vision Conference. Aberystwyth:BMVA, 2010:97.1-97.11.

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

[12] Eichner M, Marin-Jimenez M, Zisserman A, et al. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images[J]. International Journal of Computer Vision, 2012, 99(2):190-214.

[13] Walk S, Majer N, Schindler K, et al. New features and insights for pedestrian detection[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA: IEEE,2010:1030-1037.

[14] Yao B, Jiang X, Khosla A, et al. Human action recognition by learning bases of action attributes and parts[C]//Proceedings of IEEE International Conference on Computer Vision. Barcelona, Spain:IEEE, 2011:1331-1338.

[15] Wang X, Wo B H, Guan Q, et al. Human action recognition based on manifold learning[J]. Journal of Image and Graphics, 2014, 19(6):914-923. [王鑫, 沃波海, 管秋, 等. 基于流形学习的人体动作识别[J]. 中国图象图形学报, 2014, 19(6):914-923.][DOI:10.11834/jig.20140612]

[16] Wang Y Y, Li Y B, Ji X F. Human action recognition based on super-interst points features[J]. Journalof Image and Graphics, 2013, 18(7): 805-812. [王扬扬, 李一波, 姬晓飞. 人体动作的超兴趣点特征表述及识别[J]. 中国图象图形学报, 2013, 18(7): 805-812.][DOI:10.11834/jig.20130710]