基于骨架特征的人体动作分类研究

庚 晶[1],葛 军[1],郭 林[2]
(1. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
2. 南京邮电大学 物联网学院,江苏 南京 210003)

摘 要:为了能够在丰富复杂的网络信息中快速找到所需图片,提出一种基于骨架特征的人体上半身动作分类方法,以提高相应图片的检索效率。对人体运动图片进行人体运动时上半身姿势识别,得到能够表示人体位置、方向以及大小的“火柴人模型”(即骨架特征),使用矩阵形式对提取到的骨架特征进行描述。为了校正因距离和位置变化造成的尺度差异,对特征矩阵进行归一化处理,然后使用多分类 SVM 方法对提取的骨架特征进行训练,得到可以对不同动作进行分类的分类器。以收集到的人体运动图片作为测试数据库进行实验,实验结果表明,该算法的分类准确率达到 97.36%,能够很好地对人体动作进行分类。同时,在 Buffy 数据库上进行图片检索对比实验,实验结果表明,所提算法的分类准确率更高,更好地提高了图片检索效率。

关键词:动作分类;姿势识别;骨架特征;多分类 SVM
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2017)08-0083-05
doi:10.3969/j.issn.1673-629X.2017.08.017

Investigation on Human Action Classification Based on Skeleton Features

YU Jing[1], GE Jun[1], GUO Lin[2]
(1. School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to find the desired pictures quickly in the abundant and complex network information, a method for human upper-body action classification based on skeleton features is proposed to improve the efficiency of the corresponding pictures. It does the pose estimation for the image of human motion, acquires the “stickman” (skeleton features) representation of the location, orientation, and size of body parts, and describes the skeleton features with matrix form. In order to correct the scale differences caused by distance and position changes, the feature matrix is normalized. Then the multi-classification SVM is used to train the skeleton features and obtain the classifier which can classify different actions. The images of human motion collected are as the test data for experiments which show that its classification accuracy reaches 97.78% and it can do well in human action classification. At the same time, an image retrieval contrast experiment is done on the Buffy database, which show that it has higher classification accuracy and enhance image retrieval efficiency better.

Key words: action classification; pose estimation; skeleton features; multi-class SVM

0 引 言

计算机网络技术、多媒体技术的快速发展,为图像等海量视觉信息的存储和传输创造了便利条件,人们可以从网络上获得大量的图片信息。然而,日益增多的数据量也使得人们寻找自己想要的图片变得越发困难[1]。对网站来说,需要对大量的图片信息进行管理,对图片进行分类,建立索引,从而使用户能够方便地获得所需内容。对广大用户来说,也希望能够快速、有效地找到自己需要的图片信息,减少不必要的时间浪费。因此,对图片进行分类有着重要的实际意义。人体动作行为分类是其中一个重要的组成部分。对人体动作分类进行的深入研究,可提高图片检索效率。

收稿日期:2016-10-10 **修回日期**:2017-01-13 **网络出版时间**:2017-07-05
基金项目:江苏省自然科学基金(BK20130883);南京邮电大学引进人才科研启动基金(NY212016,NY214189)
作者简介:庚 晶(1991-),女,硕士研究生,研究方向为图像处理;葛 军,博士,讲师,硕士生导师,研究方向为图像处理、科学可视化。

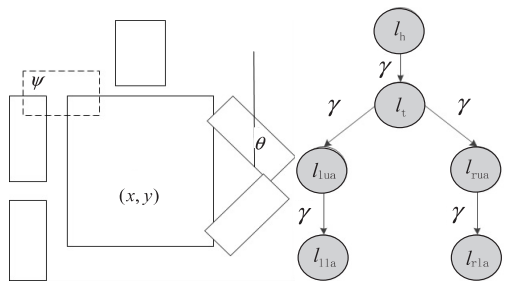
对人体动作进行分类,首先需要对人体结构进行分析,建立相应的人体结构模型,然后在此基础上进行动作姿势识别,提取动作特征,从而实现对相应动作的分类。Leung M K 等[2]使用二维带状模型来表示人体在体操动作中的每一个姿态,并通过对姿态外轮廓的单独测算得出人体的动作结构。M. Eichner 等[3-4]基于对 Ramanan 图形结构模型的扩展,通过预处理减小背景干扰,利用图像的边缘信息和区域信息对人体上半身姿势进行识别,准确地对人体运动姿势进行描述。Kellokumpu 等[5]利用从人体轮廓得到的仿射不变傅里叶描述子来实现姿势分类,该方法能够对基本动作进行正确识别,但并没有对动作分类产生真正的意义。Liu Hong 等[6]提出一种连续词袋方法,通过将一个动作分割成多个子动作来捕捉时间连续结构,最终用这些子动作分别进行分类并投票得出统一结果。Hao Yan 采用 3D Zernike 矩阵来计算人体动作的全局特征,然后使用基于 AdaBoost 的贝叶斯分类器对图像序列进行分类[7]。Sun Qianru 等[8]提出将视觉词之间的时空共生关系添加到视觉词袋中,以更加丰富地表达人体动作特征,更好地进行动作分类。文献[9]利用视觉捕捉技术,通过对视觉数据的处理来判断用户的动作。基于视觉捕捉技术,在特征表达方面,起初是采用人体轮廓作为姿势特征表达,但是轮廓特征从整体角度描述姿势,忽略了身体各部位的细节,不能精确地表示丰富多彩的人体姿势。与传统的统计理论相比,统计学习理论[10-11]基本上不涉及概率测度的定义及大数定律。它避免了人工神经网络等方法的网络结构选择、过学习和欠学习以及局部极小等问题。基于该理论发展的支持向量机(Support Vector Machine)逐渐成熟并已在模式识别、函数估计等人工智能领域得到较好的应用。因此,使用 SVM[12]对特征数据进行分类。人体动作中,上半身动作往往能够代表此时人的行为状态。为此,提出了基于人体骨架特征的人体上半身动作分类方法。该方法基于文献[4]所采用的姿势识别方法,通过对人体上半身动作进行姿势识别,从而实现表示人体位置、方向以及大小的‘火柴人模型’(骨架特征)简化特征提取,使之生动且准确地表示当前人体的动作特征。在此基础上,应用多分类 SVM[13-14]方法对提取到的人体上半身动作的骨架特征进行学习,可以对 8 种人体上半身动作进行分类、学习。

1 人体运动姿势识别

通常情况下,人的一个姿势或者一系列姿势代表了人的态度及行为,因此获得人的动作姿势特征具有重要的意义。利用图形结构估计人体外观模型,然后对得到的人体结构模型进行姿势识别。具体实现步骤包括检测人体位置、前景突出及图像解析,最终得到表示人体骨架结构的‘火柴人模型’。

1.1 图形结构模型

图形结构模型是根据一系列部件以及部件间的位置关系来表示目标,每个部件描述目标的一个局部属性(即代表一个身体部位),通过部件间的连接表示模型配置。Ramanan 模型如图 1(a)所示,其中的矩形表示各个身体部位 $l_i(x,y,\theta)$,(x,y) 表示位置信息,θ 表示方向。人体通过坐标 (x,y) 和方向 θ 参数化,通过位置先验 ψ 连接。所使用的 Eichner 的图形结构模型是基于 Ramanan 图形结构模型并利用位置先验进行扩展得到的,模型包括人身体躯干 l_t,左上手臂 l_{lua}、右上手臂 l_{rua}、左下手臂 l_{lla}、右下手臂 l_{rla},以及头部 l_h 六部分,图形结构模型如图 1(b)所示。人体上半身的六个身体部位通过二元约束项 $\psi(l_i,l_j)$ 连接在一个树状结构 E 中,即中一个节点表示一个身体部位。



(a)Ramanan模型 (b)所用图形结构模型

图 1 图形结构模型

给定图像 I,身体各部位组合为 L,则表示人体上半身姿势的公式即为:

$$P(L/I) \propto \exp\Big(\sum_{(i,j)\in E}\psi(l_i,l_j) + \sum_i \Phi(l_i) + \gamma(l_h) + \gamma(l_t)\Big) \quad (1)$$

其中,Φ 为一元势函数;$\Phi(l_i)$ 表示身体部位 l_i 处的局部图像特征;二元约束项 $\psi(l_i,l_j)$ 表示身体部位 i 和身体部位 j 的位置先验;$\gamma(\cdot)$ 设定接近垂直的一些 θ 值为均匀概率,设定其他方向的值为零概率,这样能够减少躯干和头部的搜索空间,从而提高它们被正确识别的机率;$\gamma(l_h)$ 表示需要身体躯干方向接近垂直的先验;$\gamma(l_t)$ 表示需要头部方向接近垂直的先验。这样能够提高正确识别的概率,也有利于对手臂的姿势识别,因为身体躯干通过位置先验 ψ 对它们的位置进行了控制。

1.2 前景突出

对图像进行人体上半身姿势识别时,由于图像中存在干扰因素,会使得姿势识别结果受到影响。因此首先需要对图像进行预处理,以消除背景因素的影响。通过输入检测框 $[p,t,w,h]$(p 和 t 分别表示包含人体的方框的左上角的横纵坐标值,w 和 h 分别为方框

的宽和高)框出图片中的人体位置,则姿势估计就在该检测框中进行,以提高搜索效率。根据输入的检测框产生一个扩大的矩形框。在得到的矩形框内对图像进行初始化 Grabcut 分割[15-16],分割出前/背景,并细化矩形框内的人体所在的范围,这样消除了大部分背景杂波。这里的前景即为人体各个身体部位。

1.3　图像解析

Ramanan 提出一个迭代的图像解析过程[17]。此阶段要解析的区域部分为前景突出输出的区域。利用式(1),结合迭代计算过程就能够有效地估计人体姿势。具体方法是利用图像边缘特征进行第一次推断,得到图像中人体各个身体部位的概率分布 $P_i(x,y)$;根据 $P_i(x,y)$ 为每个身体部位分别建立前景和背景的颜色直方图,即可得到每个身体部位的前景直方图和背景直方图,这即是一次迭代的过程,通过多次迭代即可得到一个较为准确的值来获取人体姿势。

根据以上几个步骤,就可对一幅图像中的人进行上半身动作姿势识别,得到其骨架模型,生动且准确地表示当前人体的动作特征。具体实现流程见图 2。

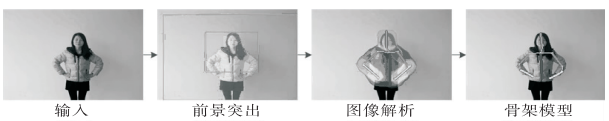


图 2　姿势识别实现效果流程图

2　基于多分类 SVM 的动作分类

SVM 基本模型定义为特征空间上的间隔最大的线性分类器,即其学习策略便是间隔最大化,最终可转化为一个凸二次规划问题的求解。SVM 方法的核心是支持向量,分类超平面由支持向量完全决定,分类函数表达式为:

$$f(x)=\mathrm{sgn}[(\omega\cdot x_i)+b]=\mathrm{sgn}\Big[\sum_{i=1}^{L}(\alpha_i-\alpha_i^*)(x\cdot x_i)+b\Big] \tag{2}$$

SVM 算法最初是为二值分类问题设计的,当处理多类问题时,就需要构造合适的多类分类器。方法主要有两类:一类是直接法,另一类是间接法。间接法主要是通过组合多个二分类器来实现多分类器的构造,常见的方法有一对多法和一对一法,使用第二种多分类 SVM 方法来实现对数据库中人体不同动作的分类。

通过对一幅图像进行人体动作识别后,得到骨架特征,六条线段分别表示身体躯干、头部、上手臂以及下手臂(见图 2)。得到的人体骨架特征由 4×6 的矩阵表示,图 2 中的骨架特征矩阵如下所示。其中,矩阵列数据表示骨架模型中的六条线段,行数据表示每条线段上下两个终点的横纵坐标值。

身体躯干	左上手臂	右上手臂	左下手臂	右下手臂	头部
348.304 3	241.869 6	439.534 2	302.689 4	382.515 5	348.304 3
367.050 4	287.285 7	298.680 7	355.655 5	363.252 1	196.126 1
340.701 9	287.484 5	393.919 3	249.472 0	435.732 9	340.701 9
230.310 9	234.109 2	230.310 9	321.470 6	329.067 2	127.756 3

为了校正因距离和位置变化造成的尺度差异,对上面输出的矩阵数据进行图像归一化处理,以消除影响。图片中心点为坐标(0,0),图片左上角坐标为(-1,-1),图片右下角坐标为(1,1),对得到的矩阵中的数据进行归一化处理,使所有数据在(-1,1)之间,归一化表达式如式(3)所示。

$$\begin{cases} m' = (m - w')/w' \\ n' = (n - h')/h' \end{cases} \tag{3}$$

其中,m 和 n 分别为线段终点的横坐标值和纵坐标值;w' 为输入图片宽度的一半;h' 为输入图片高度的一半;m' 和 n' 为经过归一化后的数值。

归一化后矩阵如下所示:

身体躯干	左上手臂	右上手臂	左下手臂	右下手臂	头部
0.088 5	-0.244 2	0.373 5	-0.054 1	0.195 4	0.088 5
0.529 4	0.197 0	0.244 5	0.481 9	0.513 6	-0.182 8
0.064 7	-0.101 6	0.231 0	-0.220 4	0.361 7	0.064 7
-0.040 4	-0.024 5	-0.040 4	0.339 5	0.371 1	-0.467 7

用多分类 SVM 对得到的特征集进行处理时,为便于数据处理,将 4×6 的矩阵转换为 1×24 的矩阵,即依次为六条线段 12 个端点的横纵坐标值,则输入 N 幅图像的特征集表示为 $N\times24$ 的矩阵,动作标签种类根据处理动作的种类数 m 依次标记为 1 到 m。使用多分类 SVM 对训练集训练后可得到一个分类器,然后使用分类器对测试集图片进行分类,得到每幅图像的动作分类结果。算法实现流程如图 3 所示。

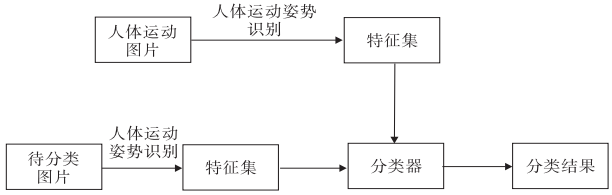


图 3　算法实现流程示意图

3 实验及结果分析

提出方法涉及的数据库是对不同人拍摄得到的。包含 8 个人,每个人 8 个动作(叉腰[1]、双臂举起[2]、站立[3]、右手臂与身体垂直[4]、左手臂竖直向上[5]、左手臂与身体垂直[6]、右手臂抬起[7]以及走路[8],人体可正面可背面,其中双臂举起可以是任意高度),每个人同一个动作的图片为不同角度拍摄,且运动者可调整距离相机的远近及偏移距离。每个动作 7 ~ 12 幅图片,共计 608 幅,图片像素为 640×480。实验所用计算机硬件配置为 Intel(R) Core(TM) i5-2450@ 2.5 GHz,4.00 GB 内存,软件环境为 Windows7 操作系统,使用 MATLAB 2011a 编程实现。具体实现步骤如下:

3.1 对数据库中的图片进行姿势识别

在姿势识别中,人体被划分为 6 个部位:身体躯干,头部,左右、上下手臂,通过这些身体部位的动作描述人的行为状态。首先输入检测框 $[p,t,w,h]$ (p 和 t 分别表示包含人体的方框的左上角的横纵坐标值, w 和 h 分别为方框的宽和高)框出图片中人体位置,经过姿势识别后,得到4 段线段衔接起来的人体骨架‘火柴人模型’,如图 4 所示。

图 4 人体上半身姿势识别结果

3.2 多分类 SVM 训练并预测

对所有图片经过姿势识别后,将得到的骨架特征数据分为训练集和测试集。选取其中 6 个人的动作骨架特征用作训练集,另外 2 个人的动作骨架特征用来测试分类器的分类准确率,训练集包含 456 幅图片,测试集包含 152 幅图片。使用多分类 SVM 算法对训练集数据进行训练,得到可以对不同动作进行分类的分类器,并对测试集进行预测。经过训练得到的分类器对训练集的分类正确率为 100%,然后对测试集中 8 个不同动作进行分类的准确率为 97.36%。8 种动作分类准确率的实验数据如表 1 所示。

表 1 8 种动作的分类准确率

动作	准确率/%	动作	准确率/%
1	100	5	100
2	100	6	100
3	100	7	100
4	100	8	80.95

从实验结果可以看出,对于 1 ~ 7 种动作,提出方法都具有较高的分类准确率,第 8 种动作的分类准确率相对较低。分析原因得知,由于人体动作图片为各个角度拍摄得到,使得动作 1 立正休息与动作 8 走路在某些角度下的姿势识别得到的骨架模型较为相似,因此将动作 8 误分类为动作 1,使得其准确率(16/21)有所下降。实验结果表明,总的分类准确率达到 97.36%,能够有效应对由于相机位置、角度的变化、不同人体高度和肢体差异等因素带来的影响,具有较高的动作分类准确率。

3.3 Buffy 数据库上的分类对比实验

对特定动作进行分类,文献[4]中使用提出的三种描述子分别表示人体结构模型。描述子 A:部位位置;描述子 B:部位方向、相对位置及相对方向;描述子 C:部位软分割,然后使用线性 SVM 分类出数据集中相应的动作。其中描述子 B 具有最好的分类效果。提出方法与其在 Buffy 数据库上进行动作分类对比实验。数据集中图片像素为 720×405,选取三种动作(包括 Hips、Rest 以及 Folded)进行动作分类效果对比,三种动作的图片共计 60 幅。

将得到的 Buffy 数据库上图片中人体的骨架特征,使用提出的方法经过矩阵归一化、转化为 N ×24 的矩阵等操作后,用 SVM 多分类方法训练特征集并对三种动作进行分类,计算分类准确率;同时使用文献[4]中 B 描述子所述方法对 Buffy 数据库上图片相应动作建立模型,实现对三种动作的分类。两种算法对这三种动作的分类效果如表 2 所示。

表 2 两种算法对三种动作的分类 %

动作	文献[4]	文中方法
Hips	30.9	100
Rest	54.6	100
Folded	11.1	58

从实验数据能够看出,Folded 动作的分类正确率相对较低,分析得知是由于数据库中这个动作的人体姿势角度变化较多(包括偏左侧立、正面站立以及偏右侧立)和双臂折叠长短的原因,使得在某些动作下得到的‘火柴人模型’类似于 Hips 动作,从而使得分类出错。通过实验可以看到,文中方法的分类准确率较高,能够提高图片检索效率。

4 结束语

在前人工作的基础上,有效地实现对人体运动动作的分类。通过对选取的 8 个常见动作进行姿势识别,得到能够有效表示人体运动状态的骨架模型,使用多分类 SVM 的方法对得到的骨架特征进行训练,从而

得到分类器,最终实现对人体动作的分类。同时使用文中方法与文献[4]中的算法在 Buffy 数据库上对特定动作进行分类对比。实验结果表明,文中方法具有很高的分类准确率,能够实现对人体运动动作的有效分类。

参考文献:

[1] Wang J Z, Gemen D, Luo J, et al. Real-world image annotation and retrieal: an introduction to the special section[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1873-1876.

[2] Leung M K, Yang Y H. First sight: a human body outline labeling system[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(4): 359-377.

[3] Eichner M, Ferrari V. Better appearance models for pictorial structures[C]//British machine vision conference. [s. l.]: [s. n.], 2009.

[4] 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.

[5] Kellokumpu V, Pietikäinen M, Heikkilä J. Human activity recognition using sequences of postures[C]//LAPR conference on machine vision applications. Japan: [s. n.], 2005: 570-573.

[6] Liu Hong, Zhang Qiaoduo. Human action classification based on sequential bag-of-words model[C]//International conference on robotics and biomimetics. [s. l.]: IEEE, 2014: 2280-2285.

[7] 胡　琼,秦　磊,黄庆明. 基于视觉的人体动作识别综述[J]. 计算机学报,2013,36(12):2512-2524.

[8] Sun Qianru, Liu Hong. Learning spatio-temporal co-occurrence correlograms for efficient human action classification[C]//2013 IEEE international conference on image processing. Melbourne: IEEE, 2013: 3220-3224.

[9] Xiong Ziyou, Radhakrishnan R. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework[C]//International conference on acoustics, speech, and signal processing. Hong Kong: Institute of Electrical and Electronics Engineers Inc, 2003: 632-636.

[10] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-42.

[11] Vapnik V N. Estimation of dependences based on empirical data[M]. New York: Springer-Verlag, 1982.

[12] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[13] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[14] 相　洁,陈俊杰. 基于 SVM 的 FMRI 数据分类:一种解码思维的方法[J]. 计算机研究与发展,2010,47(2):286-291.

[15] Rother C, Minka T. Grabcut: interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics, 2004, 23(3): 307-312.

[16] Hernándezvela A, Reyes M, Ponce V, et al. GrabCut-based human segmentation in video sequences[J]. Sensors, 2012, 12(11): 15376-15393.

[17] Schölkopf B, Platt J, Hofmann T. Learning to parse images of articulated bodies[C]//Conference on advances in neural information processing systems. [s. l.]: [s. n.], 2007: 1129-1136.

(上接第 82 页)

actions on Audio, Speech, and Language Processing, 2007, 15(4): 1435-1447.

[7] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing. [s. l.]: IEEE, 2013: 6645-6649.

[8] Chorowski J K, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition[C]//Advances in neural information processing systems. [s. l.]: Neural Information Processing Systems Foundation, 2015: 577-585.

[9] Hu Y, Wu D, Nucci A. Fuzzy-clustering-based decision tree approach for large population speaker identification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(4): 762-774.

[10] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 660-674.

[11] 孙吉贵,刘　杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.

[12] de Cheveigné A, Kawahara H. YIN, a fundamental frequency estimator for speech and music[J]. Journal of the Acoustical Society of America, 2002, 111(4): 1917-1930.

[13] Musicus B R. Levinson and fast Choleski algorithms for Toeplitz and almost Toeplitz matrices[D]. [s. l.]: Massachusetts Institute of Technology, 1988.

[14] Sim K S, Lim M S, Yeap Z X. Performance of signal-to-noise ratio estimation for scanning electron microscope using autocorrelation Levinson - Durbin recursion model[J]. Journal of Microscopy, 2016, 263(1): 64-77.

[15] Selvaperumal S K, Nataraj C, Thiruchelvam V, et al. Speech to text synthesis from video automated subtitling using levinson durbin method of linear predictive coding[J]. International Journal of Applied Engineering Research, 2016, 11(4): 2388-2395.

[16] Lloyd S P. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28(2): 129-137.