ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, Supplement, December 2010, pp.205−213
© by *Institute of Software, the Chinese Academy of Sciences*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于异构特征组效应的图像人物和动作标注方法*

邵 健+, 赵师聪

(浙江大学 计算机科学与技术学院,浙江 杭州 310027)

An Approach for Human and Motion Word Annotation with the Grouping Effect of Heterogeneous Features

SHAO Jian+, ZHAO Shi-Cong

(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: E-mail: jshao@zju.edu.cn

Shao J, Zhao SC. An approach for human and motion word annotation with the grouping effect of heterogeneous features. ***Journal of Software*, 2010,21(Suppl.):205−213.** http://www.jos.org.cn/1000-9825/10022.htm

Abstract: It is very important to select the most suitable motion words from surrounding text to describe the persons' motion expressed in images during semantic understanding. Traditional approaches often learn a generative model to denote the occurrence probability between visual objects & motion and their corresponding annotated tags, and the learned model is then utilized to recognize persons' actions in a new image outside training samples. However, all of existing approaches neglect the grouping effect of high-dimensional heterogeneous features inherent in images. In fact, different kinds of heterogeneous features have different intrinsic discriminative power for image understanding. For instance, the features extracted from arms are most discriminative to human waving motion. The selection of groups of discriminative features for motion recognition is hence crucial. In this paper, we propose an approach to select discriminative subgroup visual features from high-dimensional pose features by Group LASSO during the learning of generative model in order to boost the motion recognition. Experiments show that the proposed approach in this paper can obtain better performance for the recognition of motions with large pose change.

Key words: Group LASSO; generative model; group effect; motion word annotation

摘 要: 从图像伴随文本中选择合适动词去描述图像中人物动作对于理解图像语义具有重要意义.现有方法通常学习得到表示图像人物和运动与其标注名词-动词之间概率的生成模型,然后使用这一得到的生成模型对训练集以外图像中人物运动进行识别.但是,这一方法忽略了图像中高维异构特征之间固有存在的组效应.实际上,不同类型异构特征在图像语义理解过程中具有不同区别性,例如手臂特征对人挥手这一动作最具有区别性.为了识别图像中人物运动进而对其进行标注,提出了通过 Group LASSO 从高维异构姿势特征中选择最具区别性特征,最终学习得到生成模型的方法.实验结果表明,该方法对姿态变化较大动作进行识别时取得了更好结果.

关键词: Group LASSO;生成模型;组效应动词标注

* Supported by the National Natural Science Foundation of China under Grant Nos.60833006, 61070068 (国家自然科学基金); the China Postdoctoral Science Foundation under Grant No.20090451448; the National Key Technology R&D Program of China under Grant No.2007BAH11B05; the Fundamental Research Funds for the Central Universities of China under Grant No.KYJD09008

Received 2010-07-20; Accepted 2010-11-03

© 中国科学院软件研究所 http://www.jos.org.cn

近年来,互联网图像资源增长速度迅猛.各类新闻网站、社交网站、博客和微博上每天都会有大量新的图像被上传而共享.为了有效地对海量互联网图像信息进行检索,图像语义标注成为当前学术热点而被深入研究.在 Web2.0 时代,互联网图像或多或少会存在若干伴随文本来描述其所蕴含的对象、行为和事件等丰富语义.因此,挖掘图像底层视觉特征与其伴随文本之间关联性,建立识别模型,从而对图像中语义进行理解和标注成为一个热点问题.

先前图像语义标注工作主要集中在人物和物体识别上[1−3].由于图像中人物动作反映了客体行为,因此对其识别和标注,进而深化图像语义理解,就变得十分重要.如何高效识别图像中人物动作,并对其进行标注,从而对图像中运动进行解释和检索,是一件非常有意义的研究工作,这也是本文要研究和解决的问题.

现有人物动作识别的主要研究工作之一是如何提取人体姿势的良定特征,以便更好表达人体运动.这一方面的早先工作集中于对简单背景中人体姿势特征进行提取[4],这些方法难以处理复杂背景和较复杂动作.后续研究通过更好特征选择方法(如 HOG[5],Pictorial Structure[6,7])以及更好图像分割算法[8]对复杂背景下人物姿势进行有效表达[9].但是,由于动作的复杂性,单纯的从静态图像中对动作进行识别结果还是难以精确,因此文献[10]提出对图像中的名词-动词联合建模的方法,利用人物和动作信息的关联性,降低信息熵值,来同时识别人物和其动作.文献[10]的实验结果表明,结合人物和动作联合建模相对单一的建模方式准确率有较大提升.

然而,文献[10]的方法并未考虑图像特征之间固有差别性,在图像语义识别过程中未能根据具体情况对所得到视觉特征进行区别性选择.实际上,从图像中可提取全局特征(如颜色、纹理和形状等)、局部特征(如 SIFT, Shape Context 和 GLOH)等以及对局部特征进行聚类或量化处理而得到的视觉单词(visual words),这些不同类型特征在特定图像语义理解中所起作用不同,呈现出“组效应(grouping effect)”.所谓组效应指在某一特定图像语义识别过程中,某些重要特征会被同时选择出来表示这一特定语义,成为这一特定语义的区别性特征,而其他特征由于重要程度低,将不会被挑选出来表达这一语义.因此,如何在图像语义理解过程中对区别性特征进行选择,是将压缩感知(compressed sensing)和稀疏表达(sparse representation)等研究领域中理论和方法引入图像理解要解决的关键问题.如为了克服样本中特征相关性过大而造成过学习问题,文献[11]在稀疏差别性分析(sparse discriminant analysis)中引入等相关矩阵作为惩罚因子,给出了图像视觉特征之间在一定程度上存在“组效应”的理论证明,并将其拓展到图像多标注领域;针对图像中不同视觉特征在表示特定高层语义时所起重要程度不同,文献[12]提出了一种对异构特征进行组间和组内选择的机制,这一机制利用异构特征所存在的结构性组稀疏(structural grouping sparsity)特点,能够选择某一语义所对应的重要特征.

在图像中人物运动的识别过程中,可依据组效应对异构特征进行提取和表示,即将从人体各个不同部位所提取出特征归入不同组别,对于给定人体运动,进而学习得到对这一运动最具区别性的特征组别,将会提高人体运动识别性能.比如,给定挥手这一人体运动,从手臂所提取的姿态特征将在挥手这一人体运动识别中起到重要作用,是其最具区别性特征,而其他特征所起作用相对而言就比较小.

研究表明:图像理解过程中考虑不同种类特征在表达图像语义中所起的不同作用可有效提高识别精度[13].因此,在对图像中运动和行为进行识别时,根据人体姿态特征异构成组这一特点,有效挑选最具区别性的特征组,会极大提高和改善运动和行为识别性能.基于这样的考虑,本文提出了通过 Group LASSO[14]对异构姿态特征成组选择,来训练生成图像运动和行为识别模型的方法.

1 异构姿态特征成组选择与生成模型训练

实验中用于训练生成模型的数据集是一组表示不同人物动作的静态图像,每张图像包含相应伴随文本描述该幅图像中人物所进行动作.

本文用 D 表示包含 M 幅图像的数据集,记 $D=\{D_1,\ldots,D_M\}$,其中 $D_i(1\leqslant i\leqslant M)$表示图像 l_i 及其所对应文本 T_i.对于给定文本T_i,先用语言解析器[15]提取T_i中所有名词-动词对 V_i,如奥巴马-演讲、艺术家-弹奏等,则 $V_{i,p}\in V_i$ 为图像l_i中人物-动作的候选名词——动词对.同时,本文进一步假定对l_i中人物动作进行阐述的名词—动词均来自V_i,或者为 Null.

© 中国科学院软件研究所 http://www.jos.org.cn

在训练识别某一人体运动的生成模型时,首先手工将V_i中的名词-动词对作为l_i中人物和动作的ground-truth(基准值,即对图像中人物和动作进行标注).然后对图像中人物脸部和上身进行检测[7,16],并提取脸部和姿势特征.对于姿势特征本文按照其异构性进行分组,并用 Group LASSO 算法计算出每组特征的选择系数$\hat{\beta}_\lambda$.如果某一组姿态特征所对应选择系数不为 0,表示这一组特征被选择出来表示其对应的人体运动,而其他选择系数为 0 的姿态特征均不予选择.

通过上述异构姿态特征成组选择,得到最具区别性特征组后,假设 D 中元素各自独立,可得到公式(1):

$$P(I\mid H,\varphi)=\prod_{i=1}^{M}P(I^i\mid H_i,\varphi)=\prod_{i=1}^{M}\prod_{I^{1,D}\in I^1}P(I^{i,p}\mid h_{i,p},\varphi) \tag{1}$$

其中,φ为表示图像中人物视觉特征的参数,$I^{i,p}$表示第i幅图像中第p个人物脸部和姿态特征.$H=\{H_1,...,H_M\}$是隐含变量,表示动词-名词对 V 和图像人物-动作的关联赋值,$H_i=\{h_{i,1},...,h_{i,p_1}\}$是$V_i$对图像中每一人物-动作的标注,其中$P_i$表示图像$I_i$中所含人物的数目.生成模型训练目标是寻找合适$\varphi$,使得$P(I|\varphi)$最大,这一模型训练建立具体过程将在第 2.3 节中详述.

在训练得到某一动作生成模型后,给定训练集以外的测试图像 I',该生成模型将被用来去检测I'中是否出现了某一运动,即 $H_{res}=\arg\max_H P(I'\mid H)$.

1.1 Group LASSO

在数据分析过程中,当样本特征维数远远大于样本数目时($P\gg N$,P 和 N 分别是样本特征维数和样本个数),传统方法难以准确进行数据预测与识别,基于 l_1 范式约束的 LASSO(least absolution shrinkage and selection operator)思想因此被提出[17,18],促使被选择出来的特征尽可能稀疏,以保证结果稳定性和提高数据处理结果可解释性(interpretable).

假设对于某一人体运动,采集到了 N 个样本,从每个样本中提取 P 维特征,则训练样本构成一个 $N\times P$ 矩阵 X.Lasso 通过线性回归对这P维特征进行选择,以得到最具区别性特征.假设特征选择系数记为 $\beta\in R^p$,则可通过如下优化函数求取这 P 维特征对应的选择系数:

$$\hat{\beta}_\lambda=\arg\min(\|Y-X\beta\|_2^2+\lambda\sum_{j=1}^{P}|\beta_j|) \tag{2}$$

其中,$Y\in R^N$,β_j表示第 j 个特征的选择系数.对于向量 $u\in R^N$,定义 $\|u\|_2^2=\sum_{i=1}^{N}u_i^2$.$\lambda$是调节被选择特征稀疏度大小的变量参数,$\lambda$越大,选择系数 $\hat{\beta}_\lambda$ 中值为 0 的越多,也就是被挑选出来的特征越少.

虽然 LASSO 在进行特征选择时的算法复杂度和选择结果均取得了不错效果,但是它没有考虑变量之间的相关性,应用到存在组效应的高维异构特征选择时有很大局限性.同时 LASSO 的结果对于特征之间正交化方式较为敏感,即对特征施以不同正交变换,其选择结果将极为不同.为了解决这一问题,Yuan 等人提出了 Group LASSO[14]方法.本质上说,Group LASSO 扩展了 LASSO 方法中对特征进行选择的惩罚机理.公式(3)给出了 Group LASSO 的目标优化函数:

$$\hat{\beta}_\lambda=\arg\min\left(\left\|Y-\sum_{j=1}^{I}X_j\beta_j\right\|_2^2+\lambda\sum_{j=1}^{I}\|\beta_j\|_{K_j}\right) \tag{3}$$

其中,X_j 是对应第 j 组特征的 $N\times P_j$ 矩阵,P_j是第 j 组特征的维数:$\hat{\beta}_\lambda=(\beta_1',...,\beta_I')'$.对于向量 $\mu\in R^N$,$\|\mu\|_K=(\mu' K\mu)^{1/2}$,$K$ 为 $d\times d$ 的对称正定矩阵,可以令 $K_j=I_{p_j}$.在 Group LASSO 中,对特征选择过程进行的约束可被认为介于 l_1 范式(LASSO)和 l_2 范式(ridge regression)之间.Group LASSO 有一个很好特性,即在不同正交变换(如 ridge regression 等)下,其特征选择结果保持一致性.对于第 j 组特征,其选择系数可用公式(4)求解:

$$\beta_j=\left(1-\frac{\lambda\sqrt{P_j}}{\|S_j\|}\right)_+S_j \tag{4}$$

其中,$S_j=X_j'(Y-X\beta_{-j})$,$\beta_{-j}=(\beta_1',...,\beta_{j-1}',0',\beta_{j+1}',...,\beta_I')$.结合公式(4)和公式(3),可迭代得到每一组异构特征的选择系数 $\hat{\beta}_\lambda$.

© 中国科学院软件研究所 http://www.jos.org.cn

1.2 脸部特征提取

对给定的图像人物,本文用 multi-view 算法[16]检测得到脸部区域后,用文献[19]中的方法得到脸部 9 个不同位置的 SIFT 特征[20].将这个 9 项特征组合在一起作为脸部的特征向量.对于特征向量相似度的计算使用如下公式: $sim(a,b)=\frac{a^T b}{\|a\|\|b\|}$.而特征向量距离可定义为 $dis(a,b)=1-sim(a,b)$.

Fig.1 The extracted facial features from example images

图 1 人物脸部特征提取结果

1.3 姿态特征提取与分组

1.3.1 姿势特征提取

本文所提取的姿态特征是基于图像中人物身体各个部分空间位置的边际概率分布而得到.在姿态特征提取这一过程中,使用文献[6,7]中的算法获得人物上半身姿势 E,即身体 6 个部分(头、躯干、左/右,上/下手臂)空间位置(横坐标 x、纵坐标 y 以及方向)概率分布,记为 $E=\{E_i\}_{i=1..6}$.具体而言,E_i 是每个身体部位的边际概率分布,$E_i=p(l_i=(x,y,\theta))$.由于 E 包括了身体部分空间信息,可从 E 中提取出身体部位运动方向、相对位置和夹角等 3 种属性信息[9]来表示人物的运动特性,本文称其为姿态特征.

用 l_i 表示图像中某个人物身体部位 $i(1\leqslant i\leqslant 6)$,$l_i$ 运动方向为 θ 的概率为

$$p(l_i^0=\theta)=\sum_{(xy)}p(l_i=(x,y,\theta)) \tag{5}$$

图像中某个人物所对应身体中任意两个部位夹角为 ρ 的概率为

$$p(r(l_i^0,l_j^0)=\rho)=\sum_{(\theta_i,\theta_j)}p(l_i^0=\theta_i)p(l_j^0=\theta_j)l(r(\theta_i,\theta_j)=\rho) \tag{6}$$

其中 $l(r(\theta_i,\theta_j)=\rho)$ 是判断 θ_i 和 θ_j 之间夹角是否为 ρ 的条件参数,如果是,则为 1,否则为 0;

图像中某个人物所对应身体中任意两个部位相对位置 $\delta=(\delta_x,\delta_y)$ 的概率 $p(l_i^{xy}-l_j^{xy}=\delta)$ 可通过公式(6)类似得到.

身体部位运动方向、相对位置和夹角这 3 类异构特征可通过如下方法对其进行分组,以突出这些异构成组特征在不同运动识别过程中所起得不同重要性:

1) 公式(5)计算得到 6×24 维表示身体各个部位的运动方向划分为 24 个角度;

2) 对于身体两两部位之间相对夹角,用公式(6)计算得到的 15×24 维特征;

3) 在求取身体两两部位之间相对位置时,为降低特征维数,将图像划分为 8×8 大小子块,得到 15×8×8 个特征表示其相对位置.

这样,图像中每个人物总共可得到 1 464 维异构姿态特征,这些高维异构特征被自然分为 36 组(6+15+15).图 2 给出了 3 幅图像中人物姿态特征提取结果.

© 中国科学院软件研究所 http://www.jos.org.cn

Fig.2 Example images with pose features extracted. Different body parts are represented different colors. The deeper of the color, the more possible the pixel belong to a part

图 2 图像中人物姿态特征(即姿势边际概率)提取结果.不同颜色代表不同身体部位,颜色越深表示特征值越大(即这一姿态特征出现概率越大)

1.3.2 异构姿态特征分组

在上一节中,所提取的人体姿势特征被自然分成了 36 组(身体每个部位运动角度概率分布为 6 组,身体两两部位之间夹角概率分布和相对位置概率分布各为 15 组).由于人体运动是整体和局部的统一,为了更好表示不同身体部位之间在运动过程中相互联系和协调性,可对姿势特征进一步分组.

注意到人体上半身可分为左侧、中间和右侧 3 个区域,且很多动作只对某个区域敏感(如挥右手仅对人体上半身右侧区域敏感),因此可按此区域划分对特征进一步分组.

若用 R 表示人体部位分成的 3 个区域,即 R_1={左上臂、左下臂},R_2={头、躯干},R_3={右上臂、右下臂}.对于 R_i,将 l∈R_i 的方向特征分为一组,则得到 3 组特征;将 $l_i^m \in R_i$ 和 $l_i^n \in R_i$ 之间夹角和相对位置各分为一组,则得到 6 组特征.对于R_i和 R_j,将 $l_i \in R_i$ 与 $l_j \in R_j$ 之间夹角和相对位置各分为一组,共有 6 组.

这样,最终将得到的高维异构姿态特征分为 3+6+6=15 组.在每一组内部,姿态特征被认为是同构的;不同组之间的姿态特征被认为是异构的.

1.4 生成模型的训练

给定公式(1),令 $\varphi=(\varphi_{\text{name}},\varphi_{\text{verb}})$,每一个 $\varphi_{\text{name}}^p \in \varphi_{\text{name}}=(\varphi_{\text{name}}^1,...,\varphi_{\text{name}}^p,\beta_{\text{name}})$ 是一属于人物 p 的脸部特征向量,β_{name} 是对应人物模型为 null 时标量.每一个 $\varphi_{\text{verb}}^v \in \varphi_{\text{verb}}=(\varphi_{\text{verb}}^1,...,\varphi_{\text{verb}}^v,\beta_{\text{verb}})$ 是一属于动作 v 的姿态特征向量,β_{verb} 是对应动作模型为 null 时标量.

公式(1)可如下改写:

$$P(I\mid H,\varphi)=\prod_{i=1}^{M}\prod_{I^{1,D}\in I^i}P(I^{i,p}\mid\varphi_{\text{verb}},h_{i,p})P(I^{i,p}\mid\varphi_{\text{name}},h_{i,p}) \tag{7}$$

我们首先考察姿态特征的模型:

$$P(I^{i,p}\mid\varphi_{\text{verb}})=\sum_{k=(1,...V,NULL)}\delta(h_{i,p},k)P(I^{i,p}\mid\varphi_{\text{verb}}^k) \tag{8}$$

其中 φ_{verb}^k 是第 k 个动作模型参数; $\delta(h_{i,p},k)$ 在 $h_{i,p}=k$ 时为 1,否则为 0.因为在训练过程中已经做了标记,因此 $h_{i,p}$ 是唯一的,设 $h_{i,p}=k'$.故公式(8)可以写成:

$$P(I^{i,p}\mid\varphi_{\text{verb}})=P(I^{i,p}\mid\varphi_{\text{verb}}^{k'}) \tag{9}$$

在该概率模型的学习训练过程中,现有工作大多采用混合高斯模型的建模方法[1,3].由于姿势特征往往是非正态分布[10],因此本文采用了文献[10]中提出的如下模型:

$$P(I^{i,p}\mid\varphi_{\text{verb}}^k)=\begin{cases}\dfrac{1}{z_{\varphi_{\text{verb}}}}\mathrm{e}^{-d(I^{1,D},\varphi_{\text{verb}}^k)}, & \text{if } k\in\{knownverb\}\\ \dfrac{1}{z_{\varphi_{\text{verb}}}}\mathrm{e}^{-\beta_{\text{verb}}}, & \text{if } k=NULL\end{cases} \tag{10}$$

其中 $z_{\varphi_{\text{verb}}}$ 是归一化向量, $d(I^{1,p},\varphi_{\text{verb}}^k)$ 是待识别运动的姿态特征 $I^{i,p}$ 与学习训练得到运动 k 姿态特征向量之间的

© 中国科学院软件研究所 http://www.jos.org.cn

距离,标量β_{verb}表示运动模型为 NULL 时情况.

公式(10)同样适用于人物脸部特征模型,脸部特征向量距离可用第 2.2 节的方法得到.而对于姿态特征的距离,可以用核函数来计算两个姿势特征之间相似度 $sim(a,b)=K(a,b)=\Phi(a)\Phi(b)$,然后基于相似度如下计算得到:

$$d(I^{i,p},\varphi^k_{\text{verb}})=\left\|\Phi(a)-\frac{\sum_{b\in\pi^k}\omega(b)\Phi(b)}{\sum_{b\in\pi^k}\omega(b)}\right\|^2=K(a,a)-\frac{2\sum_{b\in\pi^k}\omega(b)K(a,b)}{\sum_{b\in\pi^k}\omega(b)}+\frac{\sum_{b\in\pi^k}\omega(b)\omega(d)K(b,d)}{\left(\sum_{b\in\pi^k}\omega(b)\right)^2} \tag{11}$$

其中 $\mu^k=\frac{\sum_{b\in\pi^k}\omega(b)\Phi(b)}{\sum_{b\in\pi^k}\omega(b)}$ 是动作 k 的姿态特征向量,π^k 是所有属于动作 k 的姿势特征聚类数目,$\omega(b)$是节点 b 权重,即 b 对于μ^k的贡献程度,表示了 b 属于 k 的可能性,该模型可看作是 k 均值聚类[21]改进.可发现,在实际计算时,第 1 项和第 3 项都是常数,因此只需要计算第 2 项.

在求取姿势特征相似度 $sim(a,b)$时,可求取人体身体 6 个部分运动方向分布、两两部分间夹角和相对位置的 Bhattacharyya 相似度,对其平均得到[22]:

$$K(a,b)=\frac{\sum_{1=\lambda}^{G}sim(\varphi_{a_1},\varphi_{b_1},\beta_1)}{G} \tag{12}$$

$$sim(\varphi_{a_1},\varphi_{b_1},\beta_1)=\sum_{j\in G_1}\begin{cases}\sqrt{\varphi_{a_{i,j}}\varphi_{b_{i,j}}}, & \text{if } \beta_{i,j}\neq 0\\ \sqrt{\varphi_{a_{i,j}}\varphi_{a_{i,j}}}, & \text{if } \beta_{i,j}=0\end{cases} \tag{13}$$

其中 $\beta_i\in\hat{\beta}_\lambda$,其是通过用 Group LASSO 所得到的异构组特征对应的选择系数.

生成模型训练的目的是在给定 H 和φ后,使得 $P(I|H,\varphi)$最大化.综合以上分析,最优化公式为

$$P(I\mid H,\varphi)=\prod_{i=1}^{M}\prod_{I^{1,D}\in I^1}P(I^{i,p}\mid\varphi^{k'}_{\text{verb}}) \tag{14}$$

最大化上述概率函数,相当于最小化以下目标函数:

$$\begin{aligned}\tau=&\sum_{i,p,h_{1,D}\neq NULL}d(I^{i,p},\varphi^k_{\text{name}})+\sum_{i,p,h_{1,D}=NULL}\beta_{\text{name}}+\sum_{i,p}\log Z_{\varphi_{\text{name}}}+\\&\sum_{i,p,h_{1,D}\neq NULL}d(I^{i,p},\varphi^k_{\text{verb}})+\sum_{i,p,h_{1,D}=NULL}\beta_{\text{verb}}+\sum_{i,p}\log Z_{\varphi_{\text{verb}}}\end{aligned} \tag{15}$$

由于 $\sum_{i,p}\log Z_{\varphi_{\text{name}}},\sum_{i,p}\log Z_{\varphi_{\text{verb}}}$ 计算复杂,可将其近似作为常数而不予考虑.公式(15)可用文献[10]中的 EM 算法,先将$\omega(b)$全部置为 1,然后迭代求解得到.

2 实 验

2.1 数据集和评价指标

互联网上很多新闻图像存在伴随文本,比如文献[1,2]所使用的数据集.但是,这些数据集并不适合本文实验,因为这些数据集中图像大多是人物头像,没有人物身体信息,无法提取比较好的姿势特征.因此本文在对比实验中采用了文献[10]中所使用的图像数据集.该数据集是用一些体育界和政治界著名人物名字和动作组合(如“克林顿”+“握手”、“科比 布莱恩特”+“扣篮”),提交给 Google 图像搜索引擎得到的.

该数据集被分成了训练集和测试集两部分,训练集有 1 610 张带伴随文本的图像,包含了 25 个人物,75 个动作,共 2 627 个名字-动作对;测试集有 103 张带伴随文本的图像,包含了 6 个人物,5 个动作,其中,每张图像中人脸部分所占图像面积的比例小于 5%.

本文采用动作识别精度作为对实验结果的评价指标.动作识别精度的计算公式如下:

$$\text{动作识别精度(ARC)}=\frac{\text{正确识别数}}{\text{识别总数}}.$$

2.2 动作识别结果及分析

对训练集图像提取出其中所包含人物运动的姿态特征后,调节 Group LASSO 中参数λ,可选择稀疏程度不

© 中国科学院软件研究所 http://www.jos.org.cn

同的异构特征组,用得到的异构特征组训练得到生成模型,以识别某一人物运动,进而实现其标注.

为了比较组效应特征选择算法在运动识别中有效性,本文方法与对所有特征不进行选择方法进行了对比.表 1 展示了在最好识别结果下的识别总数和正确识别数,图 3 则显示了详细实验结果.

图 2 中横坐标为λ值,λ为 0 时即不对特征进行挑选.纵坐标为动作识别精度.从图 2 可知:对高维异构特征通过 Group LASSO 进行组效应选择后,对于身体各个部分运动变化比较明显的动作(如 hit backhand,wave)而言,其识别率有了显著改善.但是,对于身体各个部分运动变化不很明显的运动识别效果改善不大(如 shake hand,hold).这是因为后者往往动作比较固定,各个部位的特征变化都不明显,而且这类动作在提取姿势特征中所产生的噪声也会比较小.而动作相对比较多样的动作,通过 Group LASSO 方法可挑选出来表征其运动的关键特征组,而其他特征对运动识别的影响较小.

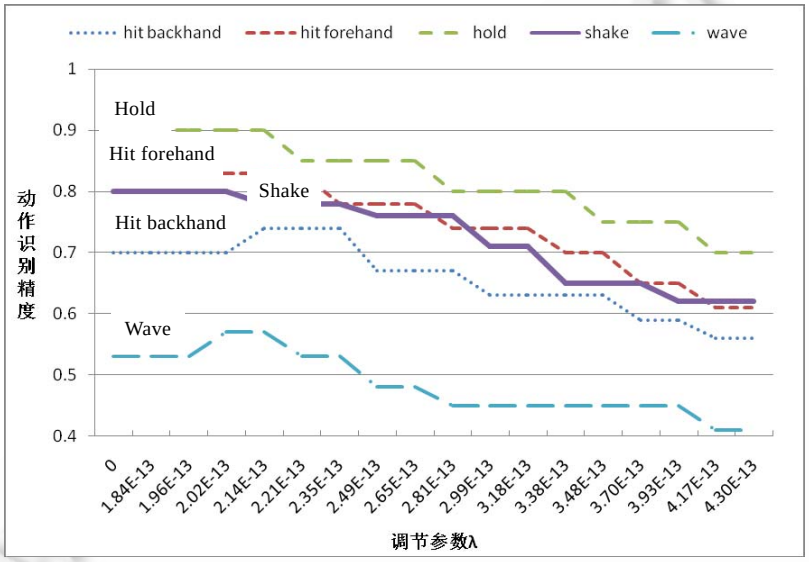


Fig.3 Comparison of recognition result of five motions(λ is the *X*-axis, when λ is zero, no selection will be done to features. The recognition accuracy is the *Y*-axis)

图 3 5 组动作识别结果对比(横坐标为λ值,λ为 0 时即不对特征进行挑选.纵坐标为动作识别精度)

另外,从图 3 也可以发现运动识别效果随着λ值增大而减小.由于λ是用于调整特征选择稀疏的参数(一般情况下λ越大,所选择出来特征越少,即选择结果越稀疏),λ越大,意味着被筛选掉的特征越多.这样,一定程度会将若干区别性特征丢失掉,从而影响了运动识别效果.

Table 1 The best recognition accuracy of five different motions

表 1 5 组动作在最好情况下的动作识别精度

	Hit backhand	Hit forehand	Hold	Shake	Wave
识别总数	27	23	20	41	33
正确识别数	20	19	18	33	19
识别精度	0.74	0.83	0.9	0.8	0.57

2.3 人物-动作识别结果

图 4 展示了测试样例图像在生成模型下的识别结果.图中测试图像的伴随文本为“Still alive ... Roger Federer gets ready to play a backhand to return Rafal Nadal’s serving”,候选名词-动词为“Federer-hit backhand, Nadal-serve”.通过本文算法得到人物为“Federer”的概率为 63.2%,为“Nadal”的概略为 58.8%.动作为“hit backhand”的概率为 71.2%,为“serve”的概率为 40.8%.联合概率“Federer-hit backhand”为 45.0%“Nadal-serve”为 24.0%.识别结果为“Federer-hit backhand”.

© 中国科学院软件研究所 http://www.jos.org.cn

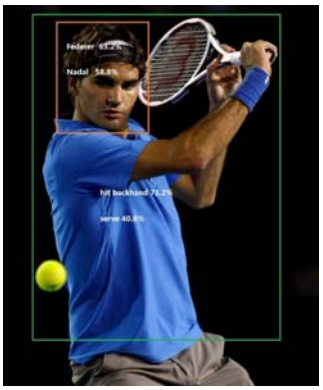


Fig.4 Example of recognition result

图 4 人物动作识别结果示例

3 总结和未来的工作

针对从图像中人物运动中所提取异构姿态特征可成组选择这一特点,本文使用 Group LASSO 来选择某一运动最具区别性的异构特征组,训练识别这一运动的生成模型以进行运动检测.实验结果表明,这一方法对姿态变化较大动作进行识别时取得了更好结果.

由于人体运动本质上可表示为时序数据,目前工作是从静态图像中进行动作识别和提取,未能充分利用运动中时序属性信息,因此今后工作将考虑如何将时序特征引入图像人物-动作识别框架,以提高识别精度.

References:

[1] Berg T, Berg A, Edwards J, Forsyth D. Who's in the picture? In: Proc. of the NIPS 2004. 2004.

[2] Duygulu P, Barnard K, de Freitas N, Forsyth D. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proc. of the ECCV 2002. 2002.

[3] Guillaumin M, Mensink T, Verbeek J, Schmid C. Automatic face naming with caption based supervision. In: Proc. of the CVPR 2008. 2008.

[4] Ioffe S, Forsyth D. Finding people by sampling. In: Proc. of the ICCV. Washington, 1999. 1092.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the CVPR, Vol.2. 2005. 886−893.

[6] Eichner M, Ferrari V. Better appearance models for pictorial structures. In: Proc. of the BMVC 2009. 2009.

[7] Ferrari V, Marin M, Zisserman A. Progressive search space reduction for human pose estimation. In: Proc. of the CVPR 2008. 2008.

[8] Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts. In: Proc. of the SIGGRAPH 2004. New York, 2004. 309−314.

[9] Ferrari V, Marin M, Zisserman A. Pose search: Retrieving people using their pose. In: Proc. of the CVPR 2009. 2009.

[10] Luo J, Caputo B, Ferrari V. Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In: Proc. of the NIPS. Vancouver, 2009.

[11] Wu F, Han YH, Tian Q, Zhuang YT. Multi-Label boosting for image annotation by structural grouping sparsity. ACM Multimedia, 2010,15−24.

[12] Han YH, Wu F, Jia JZ, Zhuang YT, Yu B. Multi-Task sparse discriminant analysis (MtSDA) with overlapping categories. In: Proc. of the 24th Conf. on Artificial Intelligence (AAAI). 2010. 469−474.

[13] Cao L, Luo J, Liang F, Huang T. Heterogeneous feature machines for visual recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). 2009.

[14] Ming Y, Yi L. Model selection and estimation in regression with grouped variables. J. R. Statist.Soc. B, 2006,68:49−67.

© 中国科学院软件研究所 http://www.jos.org.cn

[15] Deschacht K, Moens MF. Semi-Supervised semantic role labeling using the latent words language model. In: Proc. of the EMNLP. Morristown, 2009. 21−29.

[16] Rodriguez Y. Face detection and verification using local binary patterns [Ph.D. Thesis]. ´EcolePolytechnique F´ed´erale de Lausanne, 2006.

[17] Tibshirani R. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 1996,58:267−288.

[18] Tibshirani R. The lasso method for variable selection in the cox model. Statist. Med., 1997,16:385−395.

[19] Everingham M, Sivic J, Zisserman A. Hello! My name is... Buffy—automatic naming of characters in TV video. In: Proc. of the BMVC 2006. 2006.

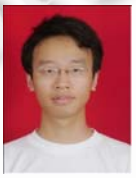
[20] Lowe D. Distinctive image features from scale-invariant keypoints. IJCV, 2004,60(2):91−110.

[21] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. In: Proc. of the IEEE PAMI. 2002.

[22] MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability. 1967.

[23] Dhillon I, Guan Y, Kulis B. Kernel *k*-means: Spectral clustering and normalized cuts. In: Proc. of the KDD. New York, 2004. 551−556.

邵健(1982－),男,浙江杭州人,博士,讲师,主要研究领域为跨媒体分析与检索,分布式计算.



赵师聪(1986－),男,硕士生,主要研究领域为多媒体搜索.

© 中国科学院软件研究所 http://www.jos.org.cn