基于半马尔可夫和 Large-margin 的动作识别

汪 力 叶 桦 夏良正
(东南大学自动控制系, 南京 210096)

摘 要 如果一个人做了一系列连续动作,并被拍摄成一段视频,那么如何通过这段视频对动作进行分割和识别是人们要考虑的问题。为了对视频中的人的动作进行有效识别,基于半马尔可夫模型框架,提出了一个对人的动作进行识别的方法,该方法通过输入-输出空间的一组特征值来抓住与 2 个动作相邻的帧的特征,以及相邻的 2 个动作段之间的特征。为了提高算法的效率,提出了一个类似于 Viterbi 的算法,该算法被用来解决优化问题。不同数据集上的实验结果表明,该方法是有效的。

关键词 动作识别 半马尔可夫 支持向量机
中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2009)11-2304-07

Discriminative Human Action Recognition Using Semi-Markov Model and Large-margin

WANG Li, YE Hua, XIA Liang-zheng
(*Automation Institute*, *Southeast University*, *Nanjing* 210096)

Abstract Given an input video sequence of one person who conducted a sequence of continuous actions, we consider the problem of jointly segmenting and recognizing actions. To recognize the activities in videos, we propose a discriminative approach to this problem within a semi-Markov model framework, where we are able to define a set of features over input-output space that captures the characteristics on boundary frames, action segments and neighboring action segments, respectively. A Viterbi-like algorithm is devised to help efficiently solve the induced optimization problem. Experiments on a variety of datasets demonstrate the effectiveness of the proposed method.

Keywords activity recognition, semi-Markov, support vector machine(SVM)

1 引 言

所谓动作识别就是将一个较长的动作,分割成基本动作,并进行识别。如一个人先跑,然后走到黑板前,再在黑板上写字,就可以分割成 3 个基本的动作。

这在人的动作分析中,属于较基本的问题,其在监控、视频检索、智能界面等方面有着广泛的用途。

然而,由于人的外貌、形状特征变化比较大,加上遮挡等问题,使得动作的识别非常困难,尤其它需要在对动作进行分割的同时进行识别,就更显得困难了。

模型 本文考虑的是一个通过学习来识别动作的方法,为了更好地说明识别模型,将在下面给出 3 个统计模型,这些模型可以被用于动作识别,例如图 1 中的 3 种统计模型。

第 1 种模型(见图 1 最上一行,如 K 近邻(KNN),支持向量机模型(SVM))。该种模型由于假设每一帧和其他帧相互独立,从而忽略了帧与帧之间的关系。

这就限制了该模型的识别能力,尤其在某些视

收稿日期:2008-06-06;**改回日期**:2008-10-16

第一作者简介:汪 力(1974~),男。东南大学博士研究生。主要研究方向是动作识别、机器学习。E-mail:wang.wanglinet@gmail.com

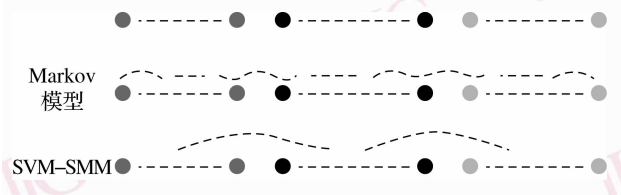
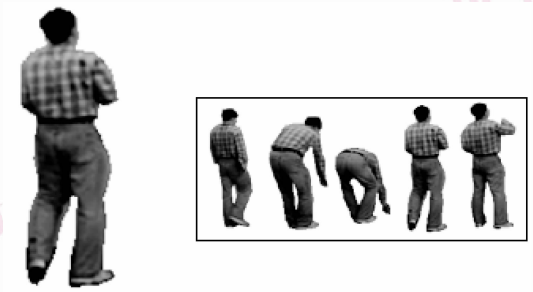


图 1　3 种不同统计模型的比较

Fig. 1　Comparison of three types of statistical models

频序列中存在有比较难判别的帧(如图 2(a)所示),这就造成在不知道上下文的情况下,很难判别出该帧属于哪个动作。



(a) 视频序列中一帧　　(b) 一组序列图像

图 2　一帧和一组序列图像的动作识别

Fig. 2　Action recognition of one frame and a series of frames

这个问题在文献[1]~[3]中已经部分解决,其解决方式是通过每个动作序列的时间、空间特征来进行识别,并使用和第 1 种模型类似的方法来识别一个有限的动作序列是属于哪个动作,然而,它们的局限性是需要对动作进行预先分段。

第 2 种模型是 Markov 模型(图 1 中间一行,如隐马尔可夫模型(HMM), CRF(conditional random field) 和 SVM-HMM[4-6]。该种模型考虑了帧与帧之间的统计相关性,并在预先分割好的视频序列中取得了较高的识别率,但是该模型并不能很好地应用于本文考虑的视频序列,这是因为:首先,连续的动作识别本质也是一个分割问题,而且动作的识别要考虑到一个动作的开始,持续一段时间到转向另一个动作的过程;其次,Markov 模型只是考虑了相邻帧之间的相关性,而没有考虑视频段的特点,如视频段的长度。

因此本文提出了一种基于支持向量机-半马尔可夫模型(SVM-SMM, 图 1 的 底行)的动作识别方法,该模型考虑了动作分割的问题,它的特点是抓住了单独动作段的特征和不同长度的动作段之间的特征。

具体地说,SVM-SMM 方法使用了以下 3 种类型的特征:(1) 每个动作段的边沿帧;(2) 动作段本身的特征;(3)相邻动作段的交互特征。

本文的贡献主要有以下 2 个方面:

(1) 提出在半马尔可夫模型下,利用 large-margin 识别方法来解决动作的分割和识别问题,并采用一个类似 Viterbi 的算法来做推导。

(2)采用了基于对象的 codebook 表示方法,并通过综合尺度不变特征变换(SIFT)[7] 和 shape context[8]得到了一组特征函数,该组函数是在输入-输出空间得到的,并能表示边沿帧和视频段的特点。

相关工作简介　一般的统计方法,尤其是 Markov 模型[4,5,9-12]在人的动作分析和建模中被广泛地使用,如 HMMs 及其派生的方法,如关联 HMMs[4,9]。除此之外,文献[5]将 HMMs 和 AdaBoost 用于动作序列的分割和识别;文献[11]利用了 2 层 SMM 来建立人的日常动作模型,而文献[12]则利用可变长度的马尔可夫模型来进行 2 维跟踪和 3 维动作捕获。

最近,基于 large-margin 的分类识别方法[13]被用到具有结构化输出的情况下[14-16](如 SVM-HMM 可以用于时间序列输出),在生物信息和自然语言处理领域也出现了很多有说服力的识别结果。就笔者所知,在视频动作分析领域,这方面的工作还不是很多。SMM 方法在文献[16]中被用于基因结构的判断。该文还提出了一个 2 步的学习算法,即第 1 步利用 SVM 对动作序列进行分割,第 2 步利用分割后的局部视频段进行识别。这个步骤和本文第 2 节提出的类似 Viterbi 的方法有很大的不同。其不同之处在于本文所提的方法是同时完成动作的分割和识别的。CRF 是另一种输出结构化信息的模型。它在文献[2],[6]中被用于动作识别,近来文献[17]又利用半马尔可夫 CRF 进行了自然语言的处理。

2　方　法

首先定义一组动作标识 $\boldsymbol{\rho}=\{1,\cdots,C\}$,一组人的标识为 $\boldsymbol{\psi}=\{1,\cdots,C\}$。在不失一般性的前提下,假设在每个给定的视频序列中只有一个人,即 $P\in\boldsymbol{\psi}$。本文把动作识别的问题表示成概率图模型的优化问题。

图模型的定义：考虑一个图模型，该模型是定义在一个动作序列 $\boldsymbol{Y}$ 和一个人的标识 $P\in\boldsymbol{\psi}$ 上。具体地说，当考虑一个半马尔可夫模型时，该图模型上的一个节点对应一个视频序列中具有相同标识的一段视频，而连接节点之间的边则表示了相邻段之间的影响。给定一个长度为 m 的视频序列 $\boldsymbol{X}=\{x_k\}_{k=0}^{m-1}$。如果给该序列加上一个空的节点，$x_m$，$l$ 表示段的个数，则一组段的边沿可用 $\{n_k\}_{k=0}^{l-1}$ 表示，其中 $n_{k-1}<n_k<n_{k+1}$，$\forall k$。并且使 $n_0=0$，$n_l=m$，以满足边沿条件。最后可以看出，第 1 段是 $[0,n_1)$，最后一段是 $[n_{l-1},m)$。

动作标识序列可以同样表示为 $\boldsymbol{Y}=\{(n_k,c_k)\}_{k=0}^{l-1}$，其中每一对 (n_k,c_k) 表示一个起始位置和相应的动作标识，该标识可对应到第 k 段 $[n_k,n_{k+1})$。

在训练中，使用了 T 段视频序列 $\boldsymbol{\chi}=\{\boldsymbol{X}_t\}_{t=1}^{T}$，以及相对应的动作标识 $\boldsymbol{\gamma}=\{\cdots,\boldsymbol{Y}_t,\cdots\}$。进一步假设，给定一个视频序列 $\boldsymbol{X}=\boldsymbol{X}_t$，标识 $\boldsymbol{Y}$ 的条件分布可以用一个指数模型表示，即

$$\log p(\boldsymbol{Y}\mid\boldsymbol{X},\boldsymbol{W})=\langle\boldsymbol{W},\boldsymbol{\Phi}(\boldsymbol{X},\boldsymbol{Y})\rangle-A_{\boldsymbol{W}}(\boldsymbol{X}) \tag{1}$$

假设动作序列之间是相互独立的，那么在所有训练序列上的联合概率可以分解为 $p(\boldsymbol{\gamma}\mid\boldsymbol{\chi},\boldsymbol{W})=\prod_t p(\boldsymbol{Y}_t\mid\boldsymbol{X}_t,\boldsymbol{W})$。其中 $A_{\boldsymbol{W}}(\boldsymbol{X})$ 是用来进行归一化的常数，该常数用于确保 $p(\boldsymbol{Y}\mid\boldsymbol{X},\boldsymbol{W})$ 表示的是一个有效的概率分布，$\boldsymbol{W}$ 表示参数向量。$\boldsymbol{\Phi}(\boldsymbol{X},\boldsymbol{Y})$ 表示输入-输出空间中的特征映射。基于 SMM 的图结构（如图 1 最底部）可以被进一步分解为

$$\boldsymbol{\Phi}(\boldsymbol{X},\boldsymbol{Y})=\begin{pmatrix}\sum_{i=0}^{l-1}\varphi_1(\boldsymbol{X},n_i,c_i),\\ \sum_{i=0}^{l-1}\varphi_2(\boldsymbol{X},n_i,n_{i+1},c_i),\\ \sum_{i=0}^{l-1}\varphi_3(\boldsymbol{X},n_i,n_{i+1},c_i,c_{i+1})\end{pmatrix} \tag{2}$$

就像在论文开始提到的那样，φ_1 和 φ_2 抓住了当前段中观测量和标识的依赖关系，并且 φ_2 表示了该段的特征。相邻段之间的影响可以用 φ_3 表示，而 $\boldsymbol{W}$ 也可以用同样的方式进行分解。

下面，将给出 large-margin 识别方法应用于结构化数据统计意义上的解释[14-15]。在判断一个视频的各种动作时，它的动作序列可以通过找视频序列的最大条件概率得到。

$$\begin{aligned}\boldsymbol{Y}^*&=\arg\max_{\boldsymbol{Y}}\log p(\boldsymbol{Y}\mid\boldsymbol{X},\boldsymbol{W})\\&=\arg\max_{\boldsymbol{Y}}F(\boldsymbol{X},\boldsymbol{Y})\end{aligned} \tag{3}$$

其中，$F(\boldsymbol{X},\boldsymbol{Y})=\langle\boldsymbol{W},\boldsymbol{\Phi}(\boldsymbol{X},\boldsymbol{Y})\rangle$ 是识别函数。

$\boldsymbol{Y}^*$ 的赋值可以通过找识别函数的最大值得到。

SVM-SMM 的学习可以通过找归一化的向量参数 $\boldsymbol{W}$ 的最优值来解决：通常希望 $\boldsymbol{W}$ 取值在一定范围之内，以避免过拟合，并能够对条件概率的 log 比值进行极小值最大化的处理

$$\min_{\boldsymbol{W}}\frac{\|\boldsymbol{W}\|^2}{2}\ \text{s. t.}\ \log\frac{p(\boldsymbol{Y}_t\mid\boldsymbol{X}_t,\boldsymbol{W})}{p(\boldsymbol{Y}\mid\boldsymbol{X}_t,\boldsymbol{W})}\geqslant\Delta(\boldsymbol{Y}_t,\boldsymbol{Y})\ \forall t,\boldsymbol{Y} \tag{4}$$

视频序列中的时间为 $\{t:t\in1,\cdots,T\}$，其中标识误差 $\Delta(\boldsymbol{Y}_t,\boldsymbol{Y})$ 是指 2 个标识之间的距离（margin）。

得到式（1）后，再加上一个松散变量 ξ 用来解决不可分的情况。由于归一化的部分被抵消，因此优化问题可以表示为

$$\begin{aligned}&\min_{\boldsymbol{W},\xi}\frac{\|\boldsymbol{W}\|^2}{2}+\frac{\eta}{T}\sum_t\xi_t\\&\text{s. t.}\quad\langle\boldsymbol{W},\boldsymbol{\Phi}(\boldsymbol{X},\boldsymbol{Y})\rangle\geqslant\Delta(\boldsymbol{Y}_t,\boldsymbol{Y})-\xi_t,\\&\qquad\eta>0\quad\forall t,\boldsymbol{Y}\end{aligned} \tag{5}$$

其中，$\boldsymbol{\Phi}(\boldsymbol{X}_t,\boldsymbol{Y})=\boldsymbol{\Phi}(\boldsymbol{X}_t,\boldsymbol{Y}_t)-\boldsymbol{\Phi}(\boldsymbol{X}_t,\boldsymbol{Y})$，它的 Dual 形式是

$$\begin{aligned}&\max_{\alpha}\sum_{t,\boldsymbol{Y}}\alpha_{t,\boldsymbol{Y}}\Delta(\boldsymbol{Y}_t,\boldsymbol{Y})-\frac{\eta}{2}\Big\|\sum_{t,\boldsymbol{Y}}\alpha_{t,\boldsymbol{Y}}\Delta\boldsymbol{\Phi}(\boldsymbol{X}_t,\boldsymbol{Y})\Big\|^2\\&\text{s. t.}\ \alpha_{t,\boldsymbol{Y}}\in\boldsymbol{\Psi}\quad\forall t\end{aligned} \tag{6}$$

$\boldsymbol{\Psi}$ 表示概率上的单一约束。利用 Representer 定理[18]就可以得到判别函数的一个 Dual 表示方法，即

$$F(\boldsymbol{X},\boldsymbol{Y})=\sum_{t,\tilde{\boldsymbol{Y}}}\alpha_{t,\tilde{\boldsymbol{Y}}}\langle\Delta\boldsymbol{\Phi}(\boldsymbol{X}_t,\tilde{\boldsymbol{Y}}),\boldsymbol{\Phi}(\boldsymbol{X}_t,\boldsymbol{Y})\rangle \tag{7}$$

F 能够分解为以下 3 个部分：

$$f_i(\boldsymbol{X},\boldsymbol{Y})=\langle w_i,\varphi_i(\boldsymbol{X},\boldsymbol{Y})\rangle,\ \forall i=\{1,2,3\}$$

从而得到

$$\begin{aligned}F(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=0}^{l-1}\big(&f_1(\boldsymbol{X},n_i,c_i)+f_2(\boldsymbol{X},n_i,n_{i+1},c_i)+\\&f_3(\boldsymbol{X},n_i,n_{i+1},c_i,c_{i+1})\big)\end{aligned} \tag{8}$$

值得注意的是，本文提出的模型具有很大的通

用性,几个已有的模型可以看成是这个通用模型的特例。假设 $M \geqslant 1$ 是一个动作段能够持续帧数的上限。如果使 $M=1$,并仅仅利用特征 φ_1 和 φ_2(如使 $\varphi_3=0$),那么就得到了一个 SVM 的多类问题。当 $M=1$,并使用所有 3 个特征,则就得到了 SVM-HMM 模型[14](图 1 中间一行)。

2.1　算法

为了得到模型参数,本文还使用了 cutting plane 方法[14],该方法能够在有限的时间内得到近似解。通过调换不同的项,式(5)中的第 t 个约束条件可以等价地表示为

$$\xi_t \geqslant \max_{(Y,P)\in\gamma} \Delta((\boldsymbol{Y}_t),(\boldsymbol{Y})) + \langle \boldsymbol{W}, \boldsymbol{\Phi}(X_t,\boldsymbol{Y})\rangle - \langle \boldsymbol{W}, \boldsymbol{\Phi}(\boldsymbol{X}_t,\boldsymbol{Y}_t)\rangle \tag{9}$$

违反约束条件(右边)的是以 ξ_t(左边)为上限的。给定当前的参数,cutting plane 方法(算法 1)可以被用来最大化目标函数,以便找到违反约束条件的最大约束,从而使下面的 column generation 问题得到解决。

无论是式(5)(primal)还是式(6)(dual)问题,它们实际上都无法求解,由于它们所需求解的空间 $\boldsymbol{\gamma}$ 耗费的时间为 $T\times C^m$,因此约束条件的数量是随训练序列长度的增长而增长的。然而这个问题可以通过称之为 column generation[14] 的优化技术来解决。

算法 1(cutting plane method)如下:

输入:数据 $\boldsymbol{X}_t$,标识 $\boldsymbol{Y}_t$,样本数量为 $T,\varepsilon>0$

初始化 $\boldsymbol{R}_t=0$;对所有 t.

repeat

　for $t=1$ to T do

　　$\boldsymbol{Y}^* = \arg\max_{\boldsymbol{Y}} \Delta(\boldsymbol{Y}_t,\boldsymbol{Y}) + F(\boldsymbol{X}_t,\boldsymbol{Y})$

　　$\xi = \max\left\{0, \max_{\boldsymbol{Y}\in\boldsymbol{R}_t} \Delta(\boldsymbol{Y}_t,\boldsymbol{Y}) + F(\boldsymbol{X}_t,\boldsymbol{Y}) - F(\boldsymbol{X}_t,\boldsymbol{Y}_t)\right\}$

　　if $\Delta(\boldsymbol{Y}_t,\boldsymbol{Y}^*) + F(\boldsymbol{X}_t,\boldsymbol{Y}^*) - F(\boldsymbol{X}_t,\boldsymbol{Y}_t) > \xi+\varepsilon$

　　then

　　增加约束集合 $\boldsymbol{R}_t \leftarrow \boldsymbol{R}_t \bigcup \{\boldsymbol{Y}^*\}$

　　使用优化方程(式(6))$\alpha_{t,\boldsymbol{Y}}$,其中 $\boldsymbol{Y}\in\boldsymbol{R}_t$

　　end if

　end for

until $\boldsymbol{R}=\{\boldsymbol{R}_1,\cdots,\boldsymbol{R}_t\}$ 在迭代中不再变化

这里并不是直接求解式(6),而是利用式(6)现有的解找到被违反的极值约束条件,并把该约束条件加进原来的优化问题,继续迭代,以寻找最优值。这个迭代过程肯定是收敛于一个优化的解[14],并且可以通过多项式次数的迭代,使所需得到的解达到任意精度。现在,针对 column generation 算法需要解决

$$\boldsymbol{Y}^* = \arg\max_{\boldsymbol{Y}\in\boldsymbol{\gamma}} \Delta(\boldsymbol{Y}_t,\boldsymbol{Y}) + F(\boldsymbol{X}_t,\boldsymbol{Y}) \tag{10}$$

当 $\boldsymbol{Y}^* \neq \boldsymbol{Y}_t$,则可以得到被违反的极值约束。这里设计了一个类似 Viterbi 的动态编程算法,如算法 1 所示。

此外,使用了 Hanming 距离来量度 $\Delta(\boldsymbol{Y},\tilde{\boldsymbol{Y}})$,从而得到下式

$$\sum_{k=0}^{m-1}(1-\delta(y_k=\tilde{y}_k))$$

$\delta(x)$ 是一个指示函数,只有当 $y_k=\tilde{y}_k$ 时,它的值才是 1。

对于任一段 i,和其相关的边界可分别表示为 $n_-=n_{i-1}$ 及 $n=n_i$。类似的,和其相关的标识可表示为 $c_-=c_{i-1}$ 和 $c=c_i$。同时可以得到一个局部的评估分数 $S(\boldsymbol{X},n,c)$,它是由该段 i 的累加和得到的(如从位置 0 开始,到 $[n_-,n)$ 段结束,其标识分别是 c_-(相应于 n_-)和 c(相应于 n)),它的定义如下:

$$\max_{c_-,\max\{0,n-M\}\leqslant n_-<n} \{S(\boldsymbol{X},n_-,c_-)+g(\boldsymbol{X},n_-,n,c_-,c)\} \tag{11}$$

增量 $g(\boldsymbol{X},n_-,n,c_-,c)$ 等于

$$f_1(\boldsymbol{X},n_-,c_-) + f_2(\boldsymbol{X},n_-,n,c_-) + f_3(\boldsymbol{X},n_-,n,c_-,c) + 1 - \sum_{k=n_-}^{n-1}\delta(y_k=c_-)$$

最后可以证明,式(10)中的 2 项加起来等于 $S(m;c_m)$。通过微小的改动,这个算法就可以在识别阶段解决式(3)中的 ML 问题。

这个 column generation 算法的效率是非常高的,其复杂度是 $O(mMC2)$,它的复杂度是随序列长度 m 成线性增加的,它的内存需求为 $O(m(C+2))$。本文算法是用 C++ 实现的,实验机器是 Intel Pentium 4 3.0ghz,内存是 512M,处理每帧图像平均花费 0.05s,这使得本文的算法能够高效地处理视频数据。

算法 2(column generation 算法)如下:

输入:长度为 m 的序列 $\boldsymbol{X}_t$,它的真实标识为 $\boldsymbol{Y}_t$,段的最大长度为 M

输出:得分 s,最优标识 $\boldsymbol{Y}^*$

初始化矩阵 $S \in \mathbf{R}^m \times C, J \in \mathbf{Z}^m$ 和 $L \in \mathbf{Z}^m$ 为 0, $Y* = 0$

for $i = 1$ to m do

for $ci = 1$ to C do

$(J_i, L_i) = \arg\max_{j,c_j} S(j, c_j) + g(j, i, c_j, c_i)$

$S(i, c_i) = S(j*, c_{j*}^*) + g(j^*, i, c_{j*}^*, c_i)$

end for

end for

$c_m^* = \arg\max_{c_m} S(m, c_m)$

$s = S(m, c_m^*)$

$Y^* \leftarrow \{(m, c_m^*)\}$

$i \leftarrow m$

repeat

$Y^* \leftarrow \{(j_i, L_i), Y^*\}$

$i \leftarrow J_i$

until $i = 0$

3 特征表示

每帧图像中的前景对象是通过背景剪除来获取的。如果能通过 SIFT[7] 检测前景对象的特征点,那么每个对象就可以用这些特征点表示,每个特征点是用 128 维的特征向量表示的,由于该特征向量对于光照条件和视角变换具有相对的不变性,更重要的是,由于是在梯度空间中获取的局部图像特征,所以不受对象的颜色影响。另外,60 维的 shape context[8] 也可以表示每个特征点的属性,它大致可以看成是每个特征点相对于其他特征点的关系,而且这 2 组特征可以拼接成一个 188 维的向量。这些特征点的集合可以再次通过 K 均值方法转换成 50 维的 codebook,这类似于文献[19]中的虚拟单词方法。

现在,当得到新的一帧图像时,则上面的每一个特征点就可以通过归类映射到 codebook 空间,而每个对象就可以被表示成 50 维的直方图向量 $\boldsymbol{h}$。利用 codebook 得到的一些典型的聚类结果显示在图 3 的最下方,其中,选择了 4 种不同的 Codebook,并且把属于这 4 种 Codebook 的特征点的位置显示在图上。上面一行显示了数据集中的某些帧,下面一行随机选取 4 个 codebook,并在不同时间和人上,显示与这 4 个 codebook 相关的特征点在图像上的位置。显示的结果说明,每个聚类的位置能够抓住对象的类似的一块部位,而不随时间和不同的人的变化而变化。

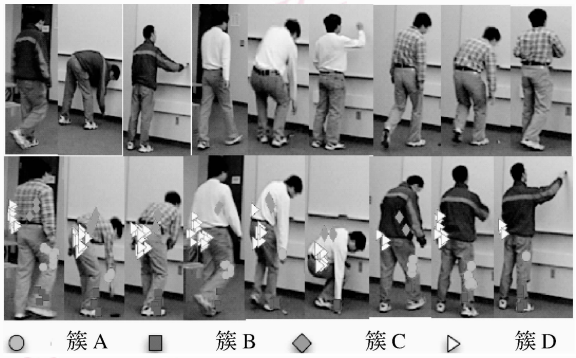


图 3 一个 Walk-Bend-Draw(WBD) 数据集

Fig. 3 A Walk-Bend-Draw(WBD) dataset

有了 codebook 的表示方式,就可以创建特征函数 φ_1, φ_2 和 φ_3。

(1)边沿帧的特征 $\varphi_1(\boldsymbol{X}, n_i, c_i) = \psi_1(\boldsymbol{X}, n_i) \otimes c_i$,其中 $\otimes$ 表示弹性乘积。ψ_1 是 2 个特征的组合。式中第 1 项是常量 1,可作为偏移项;第 2 项是从宽度为 w 的滑动窗口得到的,该滑动窗口是以边沿帧为中心的。当 $w = 1$ 时,它就是一个单独的直方图向量 $\boldsymbol{h}_{n_i}$。对于 Mobo 数据集[20],通过使 $w = 3$,就可将得到的一个由 3 个相连的直方图向量的组合作为第 2 项。

(2)段上的节点特征 节点特征被设计用来表示段的特征。φ_2 被定义为

$$\varphi_2(\boldsymbol{X}, n_i, n_{i+1}, c_i) = \psi_2(\boldsymbol{X}, n_i, n_{i+1}) \otimes c_i,$$

$\psi_2(\boldsymbol{X}, n_i, n_{i+1})$ 包含以下 3 个部分:段的长度、该段的直方图向量的均值和方差(如从第 n_i 帧到第 $n_{i+1} - 1$ 帧)。对于 Mobo 数据集,本文使用了如下的特征:段的长度,第 1 帧的直方图向量 $\boldsymbol{h}_{n_i}$、第 1 帧和中间 1 帧的距离 $\boldsymbol{h}_{n_i} - \boldsymbol{h}_{\lfloor \frac{n_i + n_{i+1}}{2} \rfloor}$,第 1 帧和最后一帧的距离 $\boldsymbol{h}_{n_i} - \boldsymbol{h}_{n_{i+1}-1}$。后 2 个特征是为了获取一个动作循环中的信息,例如,一个对象在第 1 帧中的特征与最后一帧的特征很像,但是与中间一帧却有很大的差别。

(3)相邻段的边的特征 实际上,人们是知道一个动作段至少持续的时间,假设一段的最少持续时间是 d。类似地可以得到

$$\varphi_3(\boldsymbol{X}, n_i, n_{i+1}, c_i, c_{i+1}) = \psi_3(\boldsymbol{X}, n_i, n_{i+1}) \otimes c_i \otimes c_{i+1}$$

并且它是下列几项的连接:(1) 从第 n_i 帧到第 $n_{i+1} - 1$ 帧的直方图向量的均值;(2)从第 n_{i+1} 帧到第 $n_{i+1} + d$ 帧的直方图向量的均值;(3) 从第 n_i 帧到第 $n_{i+1} - 1$ 帧的直方图向量的均值;(4) 从第 n_{i-1} 帧

到第 $n_{i+1}+d$ 帧的直方图向量的方差。

针对 Mobo 数据集,还特别生成了一组不同的特征 φ_3,它是用来获取相邻动作循环的交互特征的,如:$\boldsymbol{h}_{n_{i+1}}-\boldsymbol{h}_{n_i}$,$\boldsymbol{h}_{n_{i+1}+d}-\boldsymbol{h}_{n_i}$,$\boldsymbol{h}_{n_{i+1}+d}-\boldsymbol{h}_{n_i+d}$ 和 $\boldsymbol{h}_{n_{i+1}+d}-\boldsymbol{h}_{\left\lfloor\frac{n_i+n_{i+1}}{2}\right\rfloor}$。这些特征的解释和 φ_2 中的特征是一致的。

4　实　验

为验证本文算法的识别效果,采用 2 个数据集进行了识别实验,并将本文方法(SVM-SMM)的实验结果和其他 5 种方法做了比较,这 5 种方法包括:KNN($K=1$, 3, 5), SVM multiclass 和 SVM-HMM。为了对测试序列的动作进行分割和识别,在所有的算法中,均使用了基于帧的分类策略。为了进行公平的比较,每种方法分别进行调试,以获取最优的性能。

Walk-Bend-Draw 数据集　除了标准的数据集[20](该数据集中的序列被预先分段,以使得每个序列只有一个动作),本文还建立了 Walk-Bend-Draw(WBD)数据集用来检验本文方法的性能,该数据集的每个序列都包含数个连续的动作(图 3 显示了其中的一些图像)。这个在室内拍摄的视频数据集包含 3 个人,每个人执行 6 个动作序列,每个动作序列是在帧率为 30 fps,分辨率为 720 ×480 的情况下拍摄的,每个序列包含:走, 弯腰和写 3 个连续的动作,每个动作大约持续 2.5 s。实验时,首先对每个序列进行采样,得到 30 个关键帧,并用手工标识好了每个动作,以作为参照值。

为了验证本文算法的性能,在每个用来比较的方法中使用了 6-fold 进行交叉验证。表 1 给出了在 WBD 数据集上不同方法的动作识别率,其中 SVM-SMM 方法取得了最好的识别结果。表 2 显示了识别结果的混淆程度矩阵,由表 2 可以看出,走和写动作很少混淆,但是它们和弯腰的动作都有可能产生误识别。

表 1　WBD 数据集不同方法的识别结果对比

Tab. 1　Comparing six methods on WBD

方法	1NN	3NN	5NN	SVM	SVM-HMM	SVM-SMM
动作识别率(%)	82 ±2	80 ±3	77 ±3	84 ±3	87 ±2	91 ±2

表 2　WBD 数据集上的识别混淆程度(%)矩阵

Tab. 2　Confusion matrix on WBD matrix

	走	弯腰	写
走	78%	22%	0%
弯腰	7%	91%	2%
写	3%	11%	86%

因为尽管走和写(见图 1)看起来动作更像一点,但是走、弯腰、写是按照固定顺序依次执行的,所以能被 SVM-SMM 更好地识别。

CMU Mobo 数据集　Mobo 数据集包含了 24 个不同的人在跑步机上的行走动作。每个人执行慢走、快走、上坡走、拿球走 4 种不同的动作。每个序列被预先分为几个循环动作,实验时还标识了这些循环的边沿。这个数据集上的任务是把一个长动作序列自动分割成循环动作,并且表示出序列中每一帧的动作类型;而实验则分别检验了动作识别和分割的性能。为了检验分割的准确性,还利用了 F_1-score 检验指标进行评判,它常用于信息检索中,其计算公式为$(2\times R_1\times R_2)/(R_1+R_2)$,其中 R_1 为正确率,R_2 为检出率。

表 3 给出了经过 6-fold 交叉验证后的结果。为了节省空间,省略了 3NN 和 5NN 方法的识别结果,因为它们和 1NN 的识别结果很类似。由表 3 可以看出,SVM-SMM 和 SVM-HMM 算法的识别性能都优于基准算法,包括 KNN($K=1,3,5$) 和基于 large margin 的 SVM 方法。此外,SVM-SMM 方法在动作表示和循环动作的分割方面要好于 SVM-HMM 方法。

表 3　CMU Mobo 数据集不同方法的识别结果比较

Tab. 3　Comparison on CMU Mobo dataset

方法	1NN	SVM	SVM-HMM	SVM-SMM
动作识别率(%)	65 ±2	67 ±3	75 ±6	75 ±3
F1-score(%)	16 ±5	15 ±3	43 ±1	59 ±3

5 结 论

本文给出了一种利用半马尔可夫识别方法对人的动作进行识别的方法,其目的是同时进行动作的分割和识别。通过类似于 Viterbi 的方法,不仅可以得到动作段的特征,并可提高识别的效率。在不同数据集上的实验结果显示,本文的识别方法可以灵活地适用于不同的数据集,并在识别效果上要好于一般的方法。

本文的研究在将来还可以扩展到几个方向。很有希望的一个方向是把点集的距离计算用 Dual 方式表示。还希望把该方法用于相关的问题,如在较长时间的视频数据中对突发动作的检测。

参考文献(References)

1 Schuldt Christian, Laptev Ivan, Caputo Barbara. Recognizing human actions: A local SVM approach[A]. In: Proceedings of IEEE International Conference on Pattern Recognition[C], Washington, DC, USA, 2004:32-36.

2 Wang L, Suter D. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model[A]. In: Proceddings of IEEE Conference on Computer Vision and Pattern Recognition[C], Minneapolis, MN, USA, 2007:1-8.

3 Niebles J C, Fei Fei L. A hierarchical model of shape and appearance for human action classification [A]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition [C], Minneapolis, MN,USA, 2007:1-8.

4 Brand M, Oliver N, Pentland A. Coupled hidden markov models for complex action recognition[A]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C], Washington, DC, USA, 1997:994.

5 Lv F, Nevatia R. Recognition and segmentation of 3D human action using hmm and multi-class adaboost[A]. In: European Conference on Computer Vision[C], Anchorage, AK, 2006:359-372.

6 Sminchisescu C, Kanaujia A, Li Z, *et al*. Conditional models for contextual human motion recognition[A]. In:Proceedings of IEEE International Conference on Computer Vision[C], Beijing, 2005: 1808-1815.

7 Lowe David. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, **60**(2): 91-110.

8 Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, **24**(4):509-522.

9 Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using hidden markov model[A]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition [C], Champaign, IL, USA, 1992: 379-385.

10 Kale A, Sundaresan A, Rajagopalan A, *et al*. Identification of humans using gait[J]. IEEE Transactions on Image Processing, 2004,**13**(9): 1163-1173.

11 Duong Thi V, Bui Hung H, Phung Dinh Q,*et al*. Activity recognition and abnormality detection with the switching hidden semi-markov model[A]. In: Procceeding of IEEE Conference on Computer Vision and Pattern Recognition [C], Washington, DC, USA, 2005: 838-845.

12 Galata Aphrodite, Johnson Neil, Hogg David. Learning variable length markov models of behaviour[J]. Computer Vision and Image Understanding,2001, **81**(3):398-413.

13 Vapnik V. The Nature of Statistical Learning Theory[M]. New York, USA: Springer,1995.

14 Tsochantaridis I, Joachims T, Hofmann T, *et al*. Large margin methods for structured and interdependent output variables[J]. Journal of Machine Learning Research, 2005,**6**:1453-1484.

15 Taskar B, Guestrin C, Koller D. Max-margin Markov networks[A]. In: Saul Thrun L, SchÄolkopf B, editors: Advances in Neural Information Processing Systems 16[M], Cambridge, MA,USA: MIT Press, 2004:25-32.

16 Ratsch Gunnar, Sonnenburg Soren. Large scale hidden semi-markov svms[A]. In: Schkopf Bernhard, Platt John, Hoffman Thomas, editors: Advances in Neural Information Processing(NIPS)[M], Cambridge, MA,USA: MIT Press, 2006: 1161-1168.

17 Sarawagi Sunita, Cohen William W. Semi-markov conditional random fields for information extraction[A]. In: Schkopf Bernhard, Platt John, Hoffman Thomas, editors: Advances in Neural Information Processing(NIPS)[M], Cambridge, MA,USA: MIT Press, 2004: 1185-1192.

18 Kimeldorf G S, Wahba G. Some resultson Tchebycheffian spline functions[J]. Journal Mathematic Analysis. Application, 1971, **33**(1):82-95.

19 Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos[A]. In: Proceedings of the International Conference on Computer Vision [C], Nice, France, 2003, **2**: 1470-1477.

20 Gross R, Shi J. The cmu motion of body(mobo) database[R]. Technical Report Tech. Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2001.