基于深度学习的人体行为识别算法综述

朱 煜 [1] 赵江坤 [1] 王逸宁 [1] 郑兵兵 [1]

摘 要 人体行为识别和深度学习理论是智能视频分析领域的研究热点, 近年来得到了学术界及工程界的广泛重视, 是智能视频分析与理解、视频监控、人机交互等诸多领域的理论基础. 近年来, 被广泛关注的深度学习算法已经被成功运用于语音识别、图形识别等各个领域. 深度学习理论在静态图像特征提取上取得了卓著成就, 并逐步推广至具有时间序列的视频行为识别研究中. 本文在回顾了基于时空兴趣点等传统行为识别方法的基础上, 对近年来提出的基于不同深度学习框架的人体行为识别新进展进行了逐一介绍和总结分析; 包括卷积神经网络 (Convolution neural network, CNN)、独立子空间分析 (Independent subspace analysis, ISA)、限制玻尔兹曼机 (Restricted Boltzmann machine, RBM) 以及递归神经网络 (Recurrent neural network, RNN) 及其在行为识别中的模型建立, 对模型性能、成果进展及各类方法的优缺点进行了分析和总结.

关键词 行为识别, 深度学习, 卷积神经网络, 限制玻尔兹曼机

引用格式 朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述. 自动化学报, 2016, **42**(6): 848−857

DOI 10.16383/j.aas.2016.c150710

A Review of Human Action Recognition Based on Deep Learning

ZHU Yu[1] ZHAO Jiang-Kun[1] WANG Yi-Ning[1] ZHENG Bing-Bing[1]

Abstract Human action recognition is an active research topic in intelligent video analysis and is gaining extensive attention in academic and engineering communities. This technology is an important basis of intelligent video analysis, video tagging, human computer interaction and many other fields. The deep learning theory has been made remarkable achievements on still image feature extraction and gradually extends to the time sequences of human action videos. This paper reviews the traditional design of action recognition methods, such as spatial-temporal interest point, introduces and analyzes different human action recognition framework based on deep learning, including convolution neural network (CNN), independent subspace analysis (ISA) model, restricted Boltzmann machine (RBM), and recurrent neural network (RNN). Finally, this paper summarizes the advantages and disadvantages of these methods.

Key words Action recognition, deep learning, convolution neural network (CNN), restricted Boltzmann machine (RBM)

Citation Zhu Yu, Zhao Jiang-Kun, Wang Yi-Ning, Zheng Bing-Bing. A review of human action recognition based on deep learning. *Acta Automatica Sinica*, 2016, **42**(6): 848−857

基于机器视觉的人体行为识别是将包含人体动作的视频添加上动作类型的标签. 近年来, 随着视频采集传感器及信息科学技术的不断发展, 这方面的研究在视频监控、人机接口、基于内容的视频检索等方面逐渐成为一个具有广泛应用前景的研究课题. 自动化监控对生产生活产生很大的影响, 可以应用在商场、广场以及工业生产的监控中; 作为人机交互的关键技术, 可以将其作为智能家居的一部分应用在家庭中, 如监护小孩或者老人的危险行为等; 传统的视频检索方法都是人工对其进行标定, 其中有很多主观因素, 如果能够将人体行为识别方法应用到该领域, 将大大提高建立索引的效率及搜索效果.

人体行为识别工作主要分为两个过程: 特征表征和动作的识别及理解. 图 1 为动作识别的原理框图. 特征表征是在视频数据中提取能够表征这段视频关键信息的特征, 这个过程在整个识别过程起了关键的作用, 特征的好坏直接会影响到最终的识别效果. 动作识别及理解阶段是将前一阶段得到的特征向量作为输入经过机器学习算法进行学习, 并将在测试过程或应用场景中得到的特征向量输入到上述过程得到的模型中进行类型的识别.

图 1 动作识别原理框图

Fig. 1 The flowchart of action recognition

人体行为识别特征提取方法早期有基于人体几

收稿日期 2015-10-31 录用日期 2016-04-18
Manuscript received October 31, 2015; accepted April 18, 2016
国家自然科学基金 (61370174, 61271349), 中央高校基本科研业务费专项资金 (WH1214015) 资助
Supported by National Natural Science Foundation of China (61370174, 61271349) and the Fundamental Research Funds for the Central Universities (WH1214015)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 华东理工大学信息科学与工程学院 上海 200237
1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237

何特征的计算方法[1]、运动信息的特征提取方法[2]; 随着 HOG (Histogram of oriented gradient)[3]、SIFT (Scale-invariant feature transform)[4] 等具有先验知识的多尺度特征提取算法的提出, 结合视频序列信息的 HOG3D (Histogram of gradients 3D) 等基于时空兴趣点的特征提取方法得到了长足发展[5−11].

以上方法在特征提取之后通常采用常见的模式识别算法如支持向量机 (Support vector machine, SVM) 等进行分类识别. 近年来随着深度学习 (Deep learning) 理论的提出[12−14], 为设计无监督的自动特征学习方法奠定了基础, 其理论框架应用于行为识别也得到了长足发展. 本文在介绍传统算法的基础上, 重点分析深度学习算法在行为识别中的研究进展.

在人体行为识别过程中主要遇到以下几方面的挑战:

1) 类内和类间数据的差异

对于很多动作, 它们本身就具有很大的差异性, 例如不同人物或者不同时刻的行走动作在速度或者步长都可能具有差异. 不同动作之间又可能具有很大的相似性. 例如 KTH 数据库中的慢跑和跑步.

2) 场景和视频采集的条件

背景复杂甚至是动态变化的, 或者在动作过程中光照、天气等发生变化都会对特征提取算法的选择和算法的计算结果产生很大的影响. 其次, 视频采集条件等其他因素也会对其产生影响, 例如摄像头晃动等.

目前国内外有多个人体行为数据库供广大科研人员下载使用, 使用公共数据库能够方便地验证相关算法的可行性及对比不同算法的性能.

1) Weizman 行为数据库[15]

该人体行为数据库是在以色列 Weizman 科学研究所录制拍摄的, 包含 10 种动作 (走路、快跑、向前跳、侧身跳、弯腰、挥单手、挥双手、原地跳、全身跳、单腿跳). 每个动作由 10 个人来演示. 该数据库背景固定并且前景轮廓已经包含在数据库中, 视角固定. 如图 2 为 Weizman 数据库部分动作示例.

2) KTH 数据库[5]

该人体行为数据库包括 6 种动作 (走、跳、跑、击拳、挥手和拍手), 是由 25 个不同的人执行的, 分别在四个场景下, 一共有 599 段视频. 除了镜头的拉近拉远、摄像机的轻微运动外, 背景相对静止. 如图 3 为 KTH 数据库部分动作示例.

3) UCF Sports 数据库[16−17]

该人体行为数据库包含 150 个视频序列, 这些都是从各种广播体育频道如 BBC 和 ESPN 上收集得到的, 该数据库涵盖很广的场景类型和视角区域.

图 2 Weizman 数据库部分动作示例

Fig. 2 Examples of Weizman database

图 3 KTH 数据库部分动作示例

Fig. 3 Examples of KTH database

这个数据库中由 10 类行为动作组成: 跳水、打高尔夫、踢腿、举重、骑马、跑步、滑板、摇摆、侧摆和走路. 人体图像边界框在数据库中已给出. 在视频中有一定的人体外形、视角、光照和背景的变化及摄像头的移动. 如图 4 为 UCF Sports 数据库部分动作示例.

4) Hollywood 数据库[18]

该人体行为数据库是从 32 部好莱坞电影中采集得到的, 包含 8 个类别的动作: 接电话、下车、握手、拥抱、接吻、坐下、坐着、站起来, 总共有 1 707 个视频. 如图 5 为 Hollywood 数据库部分动作示例. Hollywood 2 将 Hollywood 数据库的动作类别扩展到了 12 类.

本文各章节内容安排如下, 首先主要介绍了课题的研究背景及常用的数据集. 第 1 节介绍传统的基于人工设计特征提取方法的研究成果. 第 2 节介绍了多个深度学习算法的理论基础及在人体行为识别上的研究进展. 最后对论文做了总结, 分析了基于深度学习算法的优缺点.

1 传统的特征提取方法

传统特征提取方法一般是通过人工观察和设计, 手动设计出能够表征动作特征的特征提取方法. 人体行为识别特征提取方法主要分为两部分: 基于人体几何或者运动信息的特征提取方法和基于时空兴趣点的特征提取方法.

1.1 基于人体几何特征或运动信息的特征提取方法

根据人体的几何形状进行行为识别是最直接的方法, Fujiyoshi 等[1] 使用四肢和头部 5 个顶点表示的星状图来表示当前帧的人体姿态, 并使用 5 个特征点与重心构成的矢量作为动作的特征向量; Yang 等[19] 从人体深度图像中采集关节点的三维坐标, 将这些关节点形成的人体轮廓作为特征进行行为识别. 使用人体几何形状的方法受限于人体几何形状的建模, 而运动中的人体形状具有一定的柔性, 不能用简单的数学模型来描述运动过程中的人体形状. 在此基础上有人提出了基于运动信息的人体行为的表征方法.

基于运动信息的人体行为的表征方法主要考虑了每帧图像在时间维度上的变化. 基于光流场的方法是基于运动信息表征方法中典型的方法. Chaudhry 等[2] 将两个方向的光流场半波整流成上下左右四个方向的运动矢量, 进行归一化并形成最终的运动描述符, 如图 6 所示. Bobick 等的研究工作延续了这一思路, 但抽取了不同的特征用于识别. 他们采用运动能量图像 (Motion energy images, MEI)[20] 和运动历史图像 (Motion history images, MHI)[21] 来解释图像序列中人的运动[22]. 基于人体几何形状或者运动信息的人体动作表征方法都是在以人体为核心的感兴趣区域内进行的. 在 Weizman 数据库中感兴趣区域已经给出, KTH 数据库虽然没有给出但是背景相对变化不大, 通过运动检测方法容易得到感兴趣区域. 所以在普通的场景下, 识别效果较好, 但在复杂场景下, 因不能得到人体的准确位置, 效果急剧下降. 表 1 总结了这两种方法在各个数据库上的结果.

图 4 UCF Sports 数据库部分动作示例

Fig. 4 Examples of UCF Sports database

图 5 Hollywood 数据库部分动作示例

Fig. 5 Examples of Hollywood database

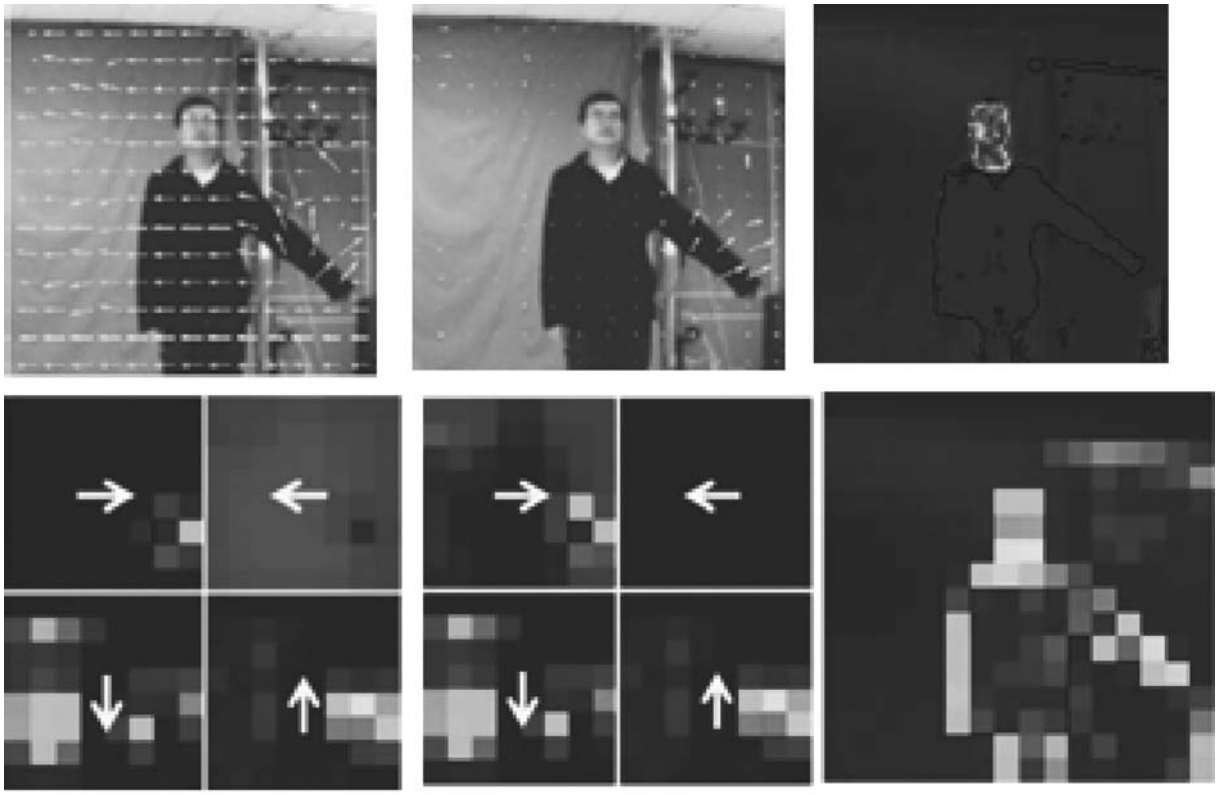


图 6　基于光流法的运动信息表征方法

Fig. 6　Movement information representation method based on optical flow method

表 1　基于几何形状或基于运动信息的识别结果 (%)

Table 1　The results of recognition methods based on geometric shapes or motion information (%)

	Fujiyoshi 等[1]	Chaudhry 等[2]
Weizman	–	100
KTH	92.73	95.77

1.2　基于时空兴趣点的特征提取方法

在背景相对复杂的情况下基于时空兴趣点的行为识别方法取得了比较好的效果. Schuldt 等[5] 将 Harris 的空域特征点扩展到三维的时空兴趣点, 通过在三维时空上进行对应的高斯模糊和局部角点提取, 获取时空兴趣点并在时空兴趣点周围进行像素直方图的统计最终形成描述动作的特征向量. 但是 Dollar 等指出这种方法检测出来稳定兴趣点的数量太少, 因此 Dollar 等提出在时间维度和空间维度上采用 Gabor 滤波器进行滤波[6], 这样检测出来的兴趣点数目就会随着局部邻域块的尺寸大小的改变而改变. Rapantzikos 等提出在 3 个维度上分别应用离散小波变换[7], 通过每一维低通和高通滤波响应来选择感兴趣的时空点. 同时为了嵌入颜色和运动信息, Rapantzikos 等又加入了彩色和运动信息来计算显著时空点. 局部时空块可以用网格来描述, 一个网格包括了观察到的局部邻域像素, 并将其看作是一个特征块, 由此减少了时空局部变化的影响. Knopp 等[8] 将二维 SURF (Speeded up robust features) 特征扩展到三维, 3D SURF 特征的每个单元包含了全部 Harr-wavelet 特征; Kláser 等[9] 将局部梯度方向直方图 HOG 特征扩展到三维形成 HOG3D, HOG3D 的每个块都是由规则多面体组成, 并且 HOG3D 可以在多尺度下对时空块进行快速密度采样, 算法流程可参考图 7; Wang 等[10] 在文献中比较了各种局部特征描述子 (HOG3D、HOG/HOF[11]、Extended SURF), 发现整合梯度与光流信息的描述子实验效果较好, 在这几个描述子中, HOG3D 的效果最好, 表 2 为各种方法在 KTH、UCF Sports 及 Hollywood 数据库上的结果.

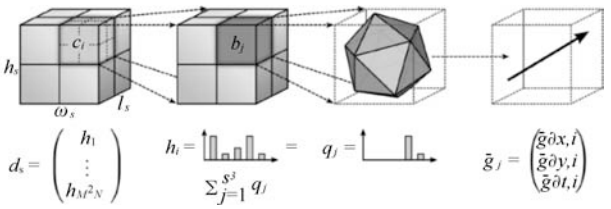


图 7　3D 梯度方向直方图获得过程

Fig. 7　HOG3D descriptor

表 2 基于时空兴趣点的特征提取方法在 KTH、UCF Sports 及 Hollywood 数据库上的结果 (%)
Table 2 The results of methods based on the interest of time and space on the KTH, UCF Sports and Hollywood databases (%)

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris 3D[5]	89 / 80 / 44	92 / 78 / 45	81 / 71 / 33	92 / 75 / 43	–	–
Cuboids[6]	90 / 83 / 46	89 / 78 / 46	82 / 73 / 39	88 / 77 / 43	89 / 77 / 45	–
Hessian[8]	85 / 79 / 41	89 / 79 / 46	78 / 66 / 36	89 / 75 / 43	–	81 / 77 / 38
Dense[11]	85 / 86 / 45	86 / 82 / 47	79 / 77 / 39	88 / 83 / 46	–	–

2 基于深度学习的特征提取方法

由于深度网络[12−14] 可以无监督地从数据中学习到特征, 而这种学习方式也符合人类感知世界的机理, 因此当训练样本足够多的时候通过深度网络学习到的特征往往具有一定的语义特征, 并且更适合目标和行为的识别. 深度学习算法可以分为四个体系: 有监督的卷积神经网络、基于自编码 (AutoEncoder) 的深度神经网络、基于限制玻尔兹曼机 (Restricted Boltzmann machine, RBM) 的深度置信网络 (Deep belief networks, DBN)[23−24] 和基于递归神经网络 (Recurrent neural network, RNN) 的深度神经网络.

2.1 基于 3D 卷积神经网络的行为识别

卷积神经网络 (Convolution neural network, CNN)[25−28] 是基于深度学习理论的一种人工神经网络, 它主要利用权值共享来减小普通神经网络中的参数膨胀问题并在前向计算过程中使用卷积核对输入数据进行卷积操作, 将得到的结果通过一个非线性函数作为该层的输出, 这样的层称为卷积层, 卷积层和卷积层之间会出现下采样层, 下采样层主要用于获取局部特征的不变性, 同时降低特征空间的尺度. 一般在卷积层和下采样层之后是一个全连接的神经网络用于最终的识别.

Ji 等[29] 将传统 CNN 拓展到具有时间信息的 3DCNN, 在视频数据的时间维度和空间维度上进行特征计算. 在卷积过程中的特征图与多个连续帧中的数据进行连接, Chéron 等[30] 使用单帧数据和光流数据, 从而捕获运动信息. 这个卷积神经网络的第一层是硬编码的卷积核, 包括灰度数据, x、y 方向的梯度, x、y 向的光流, 还包括 3 个卷积层, 2 个下采样层和 1 个全连接层, 其结构图如图 9 所示. Varol 等[31] 在定长时间的视频块内使用 3DCNN. Karpathy 等[32] 使用多分辨率的卷积神经网络对视频特征进行提取. 输入视频被分作两组独立的数据流: 低分辨率的数据流和原始分辨率的数据流. 这两个数据流都交替地包含卷积层、正则层和抽取层, 同时这两个数据流最后合并成两个全连接层用于后续的特征识别, 结构图如图 9 所示. Simonyan 等[33] 同样使用两个数据流的卷积神经网络来进行视频行为识别. 他们将视频分成静态帧数据流和帧间动态数据流. 静态帧数据流可使用单帧数据, 帧间动态的数据流使用光流数据, 每个数据里都使用深度卷积

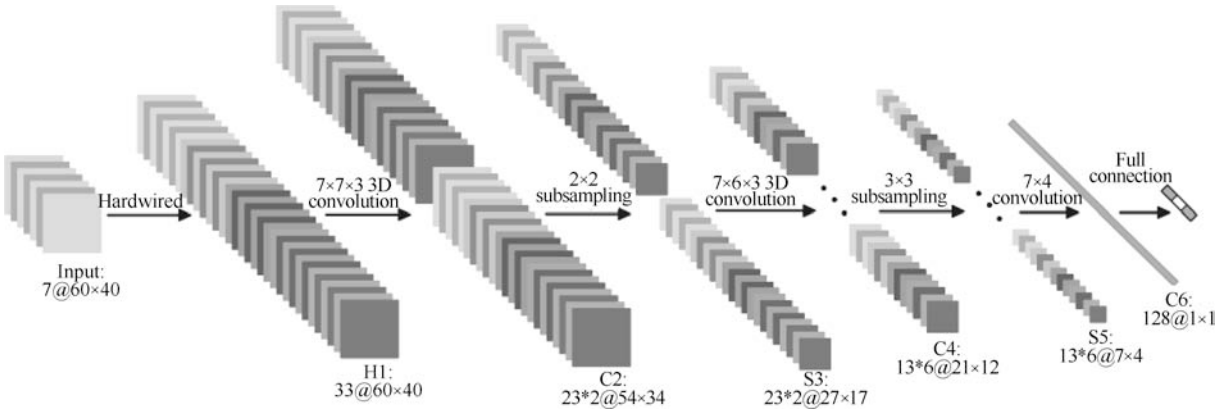


图 8 3DCNN 结构图
Fig. 8 The structure of 3DCNN

神经网络进行特征提取. 最后将得到的特征使用 SVM 进行动作的识别. 他们提出只使用人体姿势的关节点部分的相关数据进行深度卷积网络进行特征提取, 最后使用统计的方法将整个视频转换为一个特征向量, 使用 SVM 进行最终分类模型的训练和识别. 表 3 为各种方法在 KTH、UCF101 数据库上的结果, 其中, UCF101 行为识别数据库是从 YouTube 上的现实生活视频中收集得到的, 共 101 类.

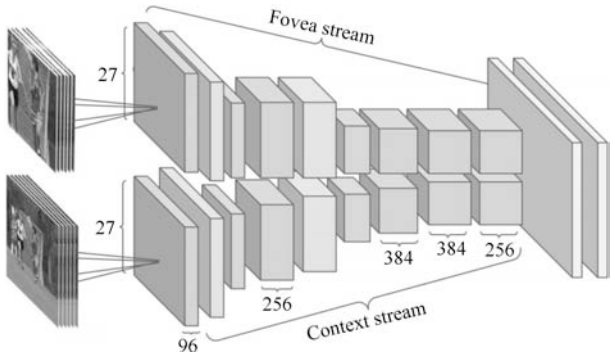


图 9　多分辨率卷积神经网络结构图

Fig. 9　The structure of multiresolution convolution neural network

表 3　基于 CNN 的行为识别算法结果 (%)

Table 3　The results of action recognition based on CNN (%)

	KTH	UCF101
Ji 等[29]	90.2	–
Simonyan 等[33]	–	88.0

2.2　基于自动编码机的无监督行为识别

自动编码机 (AutoEncoder)[26, 34−35] 是一种无监督的学习算法, 利用反向传播算法, 让目标值等于输入值, 如图 10 所示.

AutoEncoder 试图学习一个函数 $h_{w,b}(x)$, 使得 $h_{w,b} \approx x$, 也就是说它试图学习得到一个等值函数, 使得该模型的输出几乎与输入相等. Le 等[36] 将独立子空间分析 (Independent subspace analysis, ISA)[37] 扩展到三维的视频数据上, 使用无监督的学习算法对视频块进行建模. 这个方法首先在小的输入块上使用 ISA 算法, 然后将学习到的网络和较大块的输入图像进行卷积, 将卷积过程得到的响应组合在一起作为下一层的输入, 如图 11 所示. 将得到的描述方法运用到视频数据上, 这个方法同时在三个著名的行为识别库上做了实验, 表 4 为其在 KTH、UCF Sports、Hollyword 2 数据库上的结果. 可以看出, ISA 算法在具有复杂环境的 Hollywood 2 数据集上获得了更优异的性能, 较时空兴趣点算法高近 10 %.

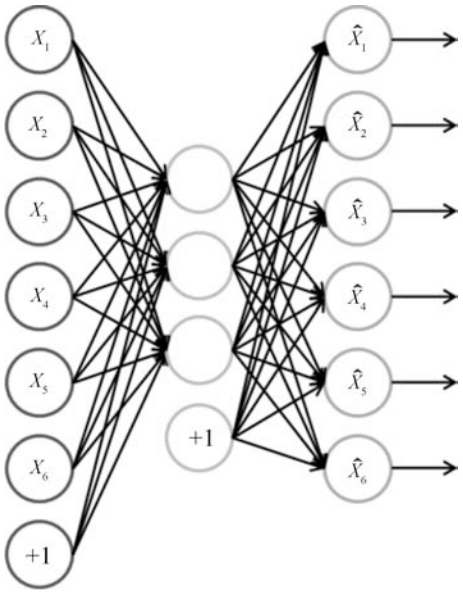


图 10　AutoEncoder 结构图

Fig. 10　The structure of AutoEncoder

表 4　ISA 在三个数据库上的结果统计 (%)

Table 4　The results of ISA on three databases (%)

	KTH	UCF Sports	Hollyword 2
Le 等[36]	93.9	86.5	53.3

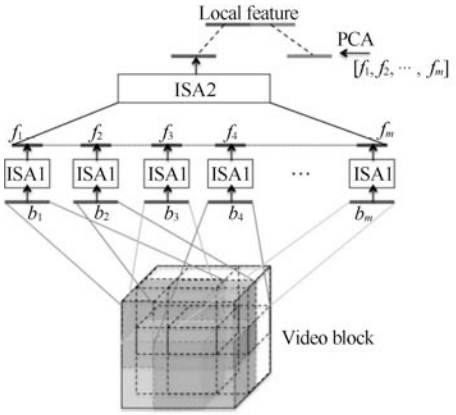


图 11　ISA-3D 结构图

Fig. 11　The structure of ISA-3D

2.3　限制玻尔兹曼机及其扩展模型

限制玻尔兹曼机 (RBM)[38−40] 是一个关于输入 (可见) 神经元 v 和输出 (隐藏) 神经元 h 之间的概率生成模型. 可见层和隐藏层的神经元之间通过一个权值矩阵 w 和两个偏置向量 c 和 b 连接. 在可见层神经元之间或者隐藏层神经元之间都没有连接.

给定一组 v 和 h, 可定义该模型的能量函数为:

$$E(v,h)=h^{\mathrm{T}}Wv+b^{\mathrm{T}}h+c^{\mathrm{T}}h$$

对应的联合概率密度是

$$P=\frac{1}{z}\mathrm{e}^{-E(v,h)}$$

其中 z 是一个配分函数, 来保证概率分布 P 是归一化的.

若可见层和隐藏层为二值 (0 或者 1), 在给定 v 的情况下 h 的概率分布和给定 h 的情况下 v 的概率分布分别是

$$p(h_j|v)=\sigma\left(b_j+\sum_i W_{ij}v_i\right)$$

$$p(v_i|h)=\sigma\left(c_i+\sum_j W_{ij}h_j\right)$$

其中 $\sigma(\cdot)$ 是激活函数, 可以选 $\sigma(x)=1/(1+\mathrm{e}^{-x})$ 或者 $\sigma(x)=\tanh(x)$ 等. 使用对比散度算法求重构误差最小值, 通过在数据上进行训练可得到概率分布的三个参数 W、b 和 c.

对于图像或视频来说, 它们都是实值数据. 使用二值分布对其建模是不合适的. 为使 RBM 能应用到此类数据上, 可将 RBM 的可见层替换成具有高斯噪声的线性变量[39, 41], 隐藏层仍然使用二值分布. 此时能量函数为:

$$E(v,h)=\sum_{i\in vis}\frac{(v_i-a_i)^2}{2\sigma_i^2}-\sum_{j\in hid}b_jh_j-\sum_{ij}\frac{v_i}{\sigma_i}h_jw_{ij}$$

其中 σ_i 是标准高斯分布的标准差.

相应的两个条件分布公式为

$$p(h_j=1|v,\theta)=\log istic\left(a_j+\sum_{i=1}^{V}w_{ij}\frac{v_i}{\sigma_i}\right)$$

$$p(v_i=1|h,\theta)=\mathrm{N}\left(b_i+\sigma_i\sum_{j=1}^{H}w_{ij}h_j,\sigma_i^2\right)$$

其中 $\mathrm{N}(\mu,\sigma^2)$ 表示均值为 μ、方差为 σ^2 的高斯分布.

条件限制玻尔兹曼机 (Conditional restricted Boltzman machines, CRBM)[40, 42] 是限制玻尔兹曼机在时间维度上的一个扩展, 它将过去时间点的可见层与当前时刻的隐含层建立连接, 因此对于二值数据来说两个条件分布公式分别为

$$p(h_j|v)=\sigma\left(b_j+\sum_i W_{ij}v_i+\sum_k\sum_i B_{ijk}v_i(t-k)\right)$$

$$p(v_i|h)=\sigma\left(c_j+\sum_j W_{ij}h_j+\sum_k\sum_i A_{ijk}v_i(t-k)\right)$$

参数 $\theta=\{W,b,c,A,B\}$ 同样可以通过对比散度算法进行优化.

Taylor 等[42] 提出将条件限制玻尔兹曼机用于人体行为识别的建模. Chen 等[43] 提出空间–时间深度信念网络 (Space-time deep belief network, ST-DBN), ST-DBN 使用卷积 RBM 神经网络将空间抽取层和时间抽取层组合在一起在视频数据上提取不变特征, 并在 KTH 数据库上获得了 91.13 % 的识别率.

2.4 递归神经网络及其扩展模型

在深度学习领域, 传统的前馈神经网络 (Feed-forward neural net, FNN) 取得了显著的成就. 但近年来随着研究的深入, FNN 模型对声音、文本、视频等信息表征时, 无法学习到信息的逻辑顺序. 为解决这一问题, 能够反映序列前后关联信息的递归神经网络 (Recurrent neural networks, RNN)[44−46] 发展迅速. RNN 将上几个时刻的隐含层数据作为当前时刻的输入, 从而允许时间维度上的信息得以保留. RNN 的网络结构如图 12 所示. 隐含层的结果 y_j 通过参数 w 作为系统输出, 同时上一时刻的 $y_j(t-1)$ 作为输入, 输入到当前时刻的系统中.

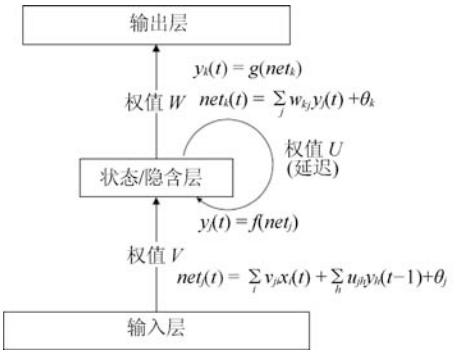


图 12 RNN 结构图

Fig. 12 The structure of RNN

长短时记忆 (Long short term memory, LSTM)[47−49] 型 RNN 模型是普通 RNN 模型的扩展, 主要用于解决 RNN 模型中的梯度消亡现象, 如图 13 所示. LSTM 接受上一时刻的输出结果, 当前时刻的系统状态和当前系统输入, 通过输入门、遗忘门和输出门更新系统状态并将最终的结果进行输出.

如下公式所示, 输入门为 i_t, 遗忘门为 f_t, 输出门为 o_t, 遗忘门来决定上一时刻的状态信息中某部分数据需要被遗忘, 输入门来决定当前输入中某部分数据需要保留在状态中, 输出门来决定由当前时刻的系统输入、前一时刻的输入和状态信息组合的

信息某些部分可以作为最终的输出.

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o\right)$$

$$h_t = o_t \tanh\left(c_t\right)$$

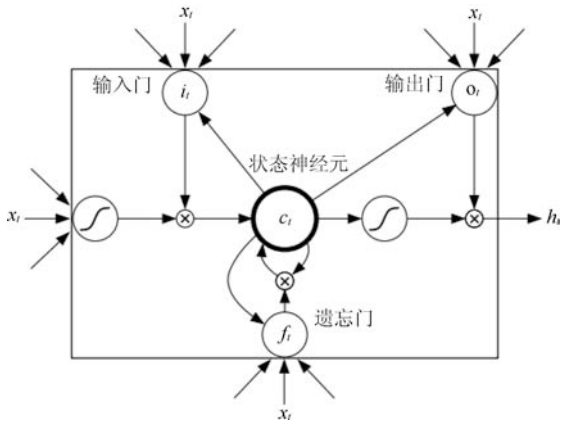


图 13　LSTM 单元

Fig. 13　The unit of LSTM

Ng 等[50] 使用 LSTM 对视频进行建模, LSTM 将底层 CNN 的输出连接起来作为下一时刻的输入, 在 UCF101 数据库上获得了 82.6 % 的识别率. Donahue 等[51] 提出了长时递归卷积神经网络 (Long-term recurrent convolutional network, LRCN), 这个网络将 CNN 和 LSTM 结合在一起对视频数据进行特征提取, 单帧的图像信息通过 CNN 获取特征, 然后将 CNN 的输出按时间顺序通过 LSTM, 这样最终将视频数据在空间和时间维度上进行特征表征, 在 UCF101 数据库上得到了 82.92 % 的平均识别率.

3　结论

本文对传统的行为识别方法和基于深度学习的人体行为识别方法进行了分析总结. 传统的方法对视频的环境或拍摄条件等有较高的要求, 并且特征提取方法是人工先验设计出来. 而基于深度学习的行为识别方法不需要像传统方法那样对特征提取方法进行人工设计, 可以在视频数据上进行训练和学习, 得到最有效的表征方法. 这种思路对数据具有很强的适应性, 尤其在标定数据较少的情况下能够获得更好的效果. 卷积神经网络在图像识别方面获得了比较好的成果, 因此基于卷积神经网络的方法一开始就获得了人们的注意, 推广至行为识别的 3DCNN 取得了不错的效果. 但是该方法属于有监督学习, 在整个学习训练过程中需要大量有标签的样本数据. 在行为识别领域无监督学习的深度学习算法, 如 ISA, 获得了比较好的效果. 基于 AutoEncoder 和 RBM 的方法可以在无标签的数据上进行无监督学习, 从而得到最佳的时空特征表示方法. 由于视频具有时间维度的信息, RNN 能够更好地适应视频的时间信息, 但是 RNN 的梯度消亡现象使得其不能很好地处理长时间的视频, LSTM 算法的提出解决了这个问题. 随着研究的深入, 相信将来会有更多更优的基于深度学习的人体行为识别方法框架被提出. 但是也应注意到, 基于深度学习的方法学习速度慢, 需要的样本数据量庞大, 这些问题的解决都期待算法的进一步研究和发展.

References

1 Fujiyoshi H, Lipton A J, Kanade T. Real-time human motion analysis by image skeletonization. *IEICE Transactions on Information and Systems*, 2004, **87-D**(1): 113−120

2 Chaudhry R, Ravichandran A, Hager G, Vidal R. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009. 1932−1939

3 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

4 Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra: IEEE, 1999. 1150−1157

5 Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition. Cambridge: IEEE, 2004. 32−36

6 Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing, China: IEEE, 2005. 65−72

7 Rapantzikos K, Avrithis Y, Kollias S. Dense saliency-based spatiotemporal feature points for action recognition. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009. 1454−1461

8 Knopp J, Prasad M, Willems G, Timofte R, Van Gool L. Hough transform and 3D SURF for robust three dimensional classification. In: Proceedings of the 11th European Conference on Computer Vision (ECCV 2010). Berlin Heidelberg: Springer. 2010. 589−602

9 Kláser A, Marszaéek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the 19th British Machine Vision Conference. Leeds: BMVA Press, 2008. 99.1−99.10

10 Wang H, Ullah M M, Klaser A, Laptev I, Schmid C. Evaluation of local spatio-temporal features for action recognition. In: Proceedings of the 2009 British Machine Vision Conference. London, UK: BMVA Press, 2009. 124.1−124.11

11 Wang H, Kláser A, Schmid C, Liu C L. Action recognition by dense trajectories. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI: IEEE, 2011. 3169−3176

12 Hinton G E. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 2007, **11**(10): 428−434

13 Deng L, Yu D. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 2014, **7**(3−4): 197−387

14 Schmidhuber J. Deep learning in neural networks: an overview. *Neural Networks*, 2015, **61**: 85−117

15 Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. In: Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005. 1395−1402

16 Soomro K, Zamir A R. Action recognition in realistic sports videos. *Computer Vision in Sports*. Switzerland: Springer. 2014. 181−208

17 Rodriguez M D, Ahmed J, Shah M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK: IEEE, 2008. 1−8

18 Marszalek M, Laptev I, Schmid C. Actions in context. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009. 2929−2936

19 Yang X D, Tian Y L. Effective 3D action recognition using EigenJoints. *Journal of Visual Communication and Image Representation*, 2014, **25**(1): 2−11

20 Bobick A, Davis J. An appearance-based representation of action. In: Proceedings of the 13th International Conference on Pattern Recognition. Vienna: IEEE, 1996. 307−312

21 Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 2006, **104**(2−3): 249−257

22 Bobick A F, Davis J W. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(3): 257−267

23 Sarikaya R, Hinton G E, Deoras A. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(4): 778−784

24 Ren Y F, Wu Y. Convolutional deep belief networks for feature extraction of EEG signal. In: Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN). Beijing, China: IEEE, 2014. 2850−2853

25 Bengio Y. Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2009, **2**(1): 1−127

26 LeCun Y, Ranzato M. Deep learning tutorial. In: Tutorials in International Conference on Machine Learning (ICML13). Atlanta, USA: Citeseer, 2013.

27 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, United States, 2012. 1097−1105

28 Bouvrie J. *Notes on Convolutional Neural Networks*. MIT CBCL Technical Report, 2006, 38−44

29 Ji S W, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(1): 221−231

30 Chéron G, Laptev I, Schmid C. P-CNN: pose-based CNN features for action recognition. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3218−3226

31 Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition. arXiV: 1604.04494, 2015.

32 Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F F. Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH: IEEE, 2014. 1725−1732

33 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc., 2014. 568−576

34 Poultney C, Chopra S, Cun Y L. Efficient learning of sparse representations with an energy-based model. In: Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006. 1137−1144

35 Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006.

36 Le Q V, Zou W Y, Yeung S Y, Ng A Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI: IEEE, 2011. 3361−3368

37 Hyvárinen A, Hurri J, Hoyer P O. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. London: Springer-Verlag, 2009.

38 Hinton G. A practical guide to training restricted Boltzmann machines. *Momentum*, 2010, **9**(1): 926

39 Fischer A, Igel C. An introduction to restricted Boltzmann machines. In: Proceedings of the 17th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Buenos Aires, Argentina: Springer. 2012. 14−36

40 Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines. In: Proceedings of the 25th International Conference on Machine Learning. New York: ACM, 2008. 536−543

41 Chen H, Murray A F. Continuous restricted Boltzmann machine with an implementable training algorithm. *IEE Proceedings-Vision, Image and Signal Processing*, 2003, **150**(3): 153−158

42 Taylor G W, Hinton G E. Factored conditional restricted Boltzmann machines for modeling motion style. In: Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009. 1025−1032

43 Chen B, Ting J A, Marlin B, de Freitas N. Deep learning of invariant spatio-temporal features from video. In: Proceedings of Conferrence on Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning. Whistler BC Canada, 2010.

44 Pineda F J. Generalization of back-propagation to recurrent neural networks. *Physical Review Letters*, 1987, **59**(19): 2229−2232

45 Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv: 1412.3555, 2014.

46 Omlin C W, Giles C L. Training second-order recurrent neural networks using hints. In: Proceedings of the 9th International Workshop Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. 361−366

47 Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv: 1402.1128, 2014.

48 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

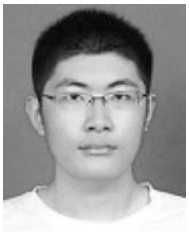
49 Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proceedings of the 2014 Annual Conference of International Speech Communication Association (INTERSPEECH). Singapore: ISCA, 2014. 338−342

50 Ng J Y H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: deep networks for video classification. arXiv: 1503.08909, 2015.

51 Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. arXiv: 1411.4389, 2014.

朱 煜 华东理工大学信息科学与工程学院教授. 1999 年获得南京理工大学博士学位. 主要研究方向为智能视频分析与理解, 模式识别方法, 数字图像处理方法及应用. 本文通信作者.
E-mail: zhuyu@ecust.edu.cn
(**ZHU Yu** Professor in the School of Information Science and Engineering, East China University of Science and Technology. She received her Ph. D. degree from Nanjing University of Science and Technology, China in 1999. Her research interest covers intelligent video analysis and understanding, pattern recognition, digital image processing methods and applications. Corresponding author of this paper.)

赵江坤 华东理工大学信息科学与工程学院硕士研究生. 主要研究方向为智能视频分析与模式识别.
E-mail: zhaojk90@gmail.com
(**ZHAO Jiang-Kun** Master student at the School of Information Science and Engineering, East China University of Science and Technology. His research interest covers intelligent video analysis and pattern recognition.)



王逸宁 华东理工大学信息科学与工程学院硕士研究生. 主要研究方向为智能视频分析与模式识别.
E-mail: wyn885@126.com
(**WANG Yi-Ning** Master student at the School of Information Science and Engineering, East China University of Science and Technology. His research interest covers intelligent video analysis and pattern recognition.)

郑兵兵 华东理工大学信息科学与工程学院硕士研究生. 主要研究方向为智能视频分析与模式识别.
E-mail: 13162233697@163.com
(**ZHENG Bing-Bing** Master student at the School of Information Science and Engineering, East China University of Science and Technology. His research interest covers intelligent video analysis and pattern recognition.)