文章编号:1671-4598(2019)03-0155-04 DOI:10.16526/j.cnki.11—4762/tp.2019.03.031 中图分类号:TP391.4 文献标识码:A

基于场景理解的人体动作识别模型

张嘉祺，赵晓丽，张 翔
（上海工程技术大学 电子电气工程学院，上海 201620）

摘要：为了满足在复杂环境下对人体动作识别的需求，提出了一种基于场景理解的双流网络识别结构；将场景信息作为辅助信息加入了人体动作识别网络结构中，改善识别网络的识别准确率；对场景识别网络与人体动作识别网络不同的融合方式进行研究，确定了网络最佳识别结构；通过分析不同参数对识别准确率的影响，最终确定了双流网络的所有结构参数，设计并训练完成了双流网络结构；通过在 UCF50，UCF101 等公开数据集上实验，分别取得了 95%，93%的准确率，高于典型的识别网络结果；对其他一些典型识别网络加入同样场景信息进行了研究，其实验结果证明了此方法可以有效改善识别准确率。

关键词：双流网络结构；场景识别；人体动作识别

Human action recognition model based on scene understanding

Zhang Jiaqi，Zhao Xiaoli，Zhang Xiang
（Department of Electrical and Electronic Engineering，Shanghai University of Engineering Science，Shanghai 201620，China）

Abstract：In order to meet the needs of human action recognition in complex environments，a dual—flow network recognition structure based on scene understanding is proposed. The scene information is added as auxiliary information to the human action recognition network structure to improve the recognition accuracy. The different fusion modes of the scene recognition network and the human action recognition network are studied，and the network optimal identification structure is determined. By analyzing the influence of different parameters on the recognition accuracy，all the structural parameters of the dual—flow network are finally determined. Through experiments on public data sets such as UCF50 and UCF101，95% and 93% accuracy were obtained，respectively，which is higher than the typical identification network results. Some other typical identification networks have been studied by adding the same scene information. The experimental results show that this method can effectively improve the recognition accuracy.

Keywords：dual stream network structure；scene recognition；human action recognition

0 引言

现阶段，监控领域中往往需要专人在监控显示器前监视，或在异常事件发生后进行人工筛选视频，这样做不仅效率低下，且不能对可能发生的危险视做出快速反应并及时报警。如何对视频内容的识别，自动化分析处理视频数据，成了一个亟需待解决的问题。[1]

对视频内容的识别，首先需要提取视频中有效的特征信息。目前，深度学习已成功应用于图片分类，人脸识别，物体检测等静止图像的识别领域中[2]，已被证明可以有效提取图像的特征信息。但在视频领域中，通常的深度卷积神经网络表现并不突出，主要由于视频不仅具有图像信息还具有时间运动信息，因此会受到光照，视角，背景，动作快慢等众多因素的影响。传统的卷积神经网络结构只能有效提取单一、静止图片的特征，而面对视频这种，前后具有很强的时间相关性和空间相关性的时空特征，很难有效提取视频的表达特征。

收稿日期:2018-08-21； **修回日期**:2018-09-26。
基金项目:国家自然科学基金项目(61461021)；上海市科教委项目(15590501300)。
作者简介:张嘉祺(1994-)，男，上海人，硕士，主要从事研究数字图像处理方向的研究。

为了能同时提取时间与空间信息，一种直观的做法是将 2D 卷积替换为 3D 卷积。Du 等人提出了 3D 卷积神经网络[3]，将连续几帧图片作为输入。他们的工作表明了 3D 卷积比 2D 卷积更加适合于视频视觉任务中，但是 3D 卷积只能提取很短一段时间（几帧）内的时间信息。为了提取解决这个问题 Simonyan 等人提出了一种双流网络结构[4]，在 3D 卷积网络的基础上又加入了人体运动的光流信息作为辅助。作者把两种网络结构的输出，通过简单的损失函数进行特征合并。这对于长视频来说这不是一种很好的聚合信息的方法。为了解决这个问题，Diba 等人提出了在时间轴上进行编码，随后压缩到低维空间中[5]。通过这种方法可以在较长时间内聚合时间信息。在动作过程中，时间和动作是相辅相成的，空间与时间的交互信息对动作识别来说也是一种非常重要的信息。但是双流结构是两个独立的网络结构，即一个网络只提取空间信息，另一个网络只提取时间信息，缺少了两种信息的交互。为了使两种网络能相互传递信息，Feichtenhofer 等人提出了一种位于两个网络之间的融合层[6]，通过这个层，可以使得时间信息与空间信息相互交换。另一种解决方法是，在卷积神经网络顶层加入循环神经网络组成的混合网络，充分利用了两种网络的优势。Danhue 等人在多帧卷积神经网络之后加入了长短期记忆模型（LSTM）循环神经网络[7]，使得可以处理较长

时间的依赖时间关系，相比于原始的双流网络结构，网络更加紧凑，计算效率更高。

过去的研究方法都是针对于如何有效提取视频的表达特征，但是一张图片中往往暗含了多个提示信息。在动作识别中，场景信息也是另一种重要的提示信息，例如在湖上的场景更有可能是划船，而不太可能是在打篮球。因此可以将场景信息作为一种提示信息加入识别网络结构中。基于此，本文提出了一种基于场景识别的双流网络结构模型，充分考虑了场景信息对识别的提示作用。本文提出的场景识别双流网络模型总体结构如图 1 所示，下面具体详细说明。

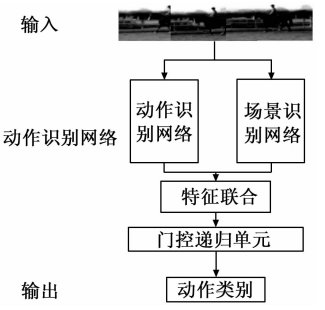

图 1 网络总体结构图

1 本文的方法

本章主要的讲述了文所采用的网络结构，以及详细的参数设置。

1.1 动作识别网络结构

浅层网络的模型泛化能力有限，网络越深，提取的特征越抽象，对识别的结果也有很大的改善，但同时参数成倍增加，消耗大量的计算资源，变得难以训练。为了解决这个问题 GoogLeNet 网络提出了 inception 模块[8]，使得在网络层数增加的同时，而不大量增加参数。

由于视频帧与帧之间图像基本相似，直接使用 GoogLeNet 网络会收敛于一个局部最优值，无法将特征最大化区分。为了解决这个问题，本文在网络中加入了批归一化层（batch normalization）[9]。批归一化层可以使得这一层的特征进行约束，使得相似的特征尽可能减少之间的距离，从而最大化不同特征间的距离，同时也避免了因为数据产生的微小变动，经过数层网络层而影响不断变大，从而产生梯度爆炸或梯度消失的问题。批归一化层的算法如下：

输入：m 维数据：$X=\{x\_(i...m)\}$

训练的参数：γ，β

输出：$\{y_i=BN_{\gamma,\beta}(x)\}$

$$\mu_X \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \text{ // 计算样本的均值} \tag{1}$$

$$\sigma_X \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_X)^2 \text{ // 计算样本方差} \tag{2}$$

$$\dot{x_i} \leftarrow \frac{x_i-\mu_X}{\sqrt{\sigma_X^2+\in}} \text{ // 正则化} \tag{3}$$

$$y_i \leftarrow \gamma \dot{x_i}+\beta \text{// 尺度平移变化} \tag{4}$$

本文将批归一化层加入 inception 模块中，加入后的 inception 模块结构如图 2 所示。加入了批归一化层后的 GoogLeNet 的详细网络结构如表 1 所示。

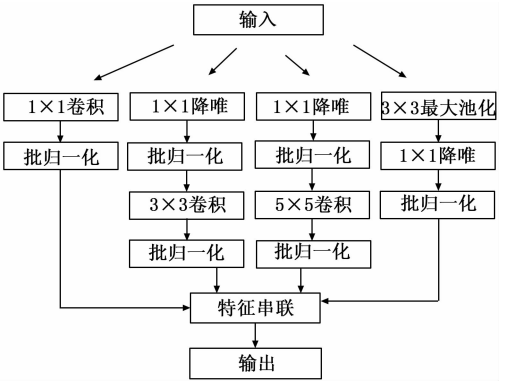

图 2 加入批归一化层后的 inception 模块结构图

表 1 加入批归一化层之后的 GoogLeNet 详细参数

类型	窗口大小/步长	输出维度
输入		224x224x3
卷积	7x7/2	112x112x64
最大池化	3x3/2	56x56x64
卷积	3x3/1	56x56x192
最大池化	3x3/2	28x28x192
Inception1		28x28x256
Inception2		28x28x320
Inception3		14x14x512
Inception4		14x14x512
Inception5		14x14x512
Inception6		14x14x528
Inception7		14x14x576
Inception8		7x7x832
Inception9		7x7x1024
平均池化	7x7/1	1x1x1024
droput		1x1x1024
全链接		1x1x2048

1.2 场景识别网络结构

将深度学习技术运用于场景识别中也是未来的趋势所在。本文采用了监督学习方法。先对每个视频按照室外，屋内，体育场，马场，湖上等进行人为分为 11 类场景。场景识别网络应该选取较小型的网络，主要的原因在于：1. 不造成整个动作识别网络参数量过大。参数量过大，不容易训练，并且可能会造成过拟合等问题。2. 若场景识别网络复杂，会导致整个网络倾向于场景识别，而不是动作识别。因此，本文选取简单的 5 层卷积神经网络结构模型，如图 3 所示。

1.3 门控递归单元

循环神经网络与卷积神经网络最大的不同是，循环神经网络引入了状态变量。在一个序列中，循环神经网络当

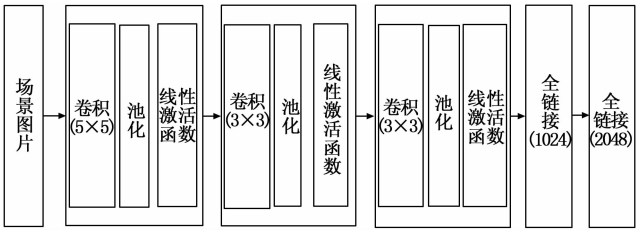

图 3　场景识别网络结构图

前时刻的状态不仅保存了过去时刻的信息，还与当前时刻的输入共同决定了当前时刻的输出，适合用于处理序列数据，在自然语言处理中也有被大量使用。但是循环神经网络缺少了遗忘的设计，使得早期的信息随着时间推移而逐渐消失，从而导致了梯度往往消失或爆炸，这使得基于梯度的优化方法变得非常困难，难以捕捉时间跨度较大的依赖关系。目前主要有两种解决方法，一种是发现更好的优化算法例如使用梯度裁剪等等[10]。第 2 种是设计一个更优的网络结构，也是当前的研究重点。这方面第一个尝试是 SakH 提出的长短期记忆模型（LSTM）[11]。与之前循环神经网络模型结构最大的不同是，长短期记忆模型加入了遗忘门和输出门，信息不再是简单输入，而是通过遗忘门进行控制。一旦 LSTM 发现了重要的信息，可以携带这些信息跨越较长时间，从而避免了一些较早期的信息丢失。

与 LSTM 类似，Kyunghyun 等人提出了门控递归单元（GRU）模型[12]。GRU 和 LSTM 主要的区别在于输入门的位置。LSTM 的输入门只针对输入进行单独控制，而 GRU 的更新门是前一步的隐状态与输入共同决定更新门的状态。在视频内容识别这个任务中，GRU 相比于 LSTM 更加合适，主要的原因之一是大部分的空间特征是相似的，对识别来说这些信息是冗余的。GRU 可以根据前一步的隐状态，对输入的信息流进行只提取区别较大的特征，进而减少这些冗余的信息量，加快了收敛的速度，提高了训练的效率。GRU 的网络示意图如图 4 所示。

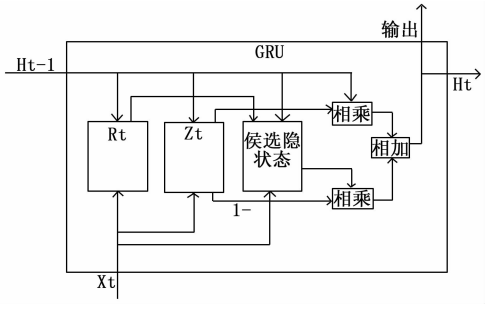

图 4　门控递归单元结构图

门控递归单元的算法过程如下：

输入：X_t，上一步的隐状态 $H_(t-1)$。

输出：当前时刻隐状态 H_t。

第一步：计算重置门和更新门的状态：

$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r) \text{ // 计算重置门的状态} \quad (5)$$

$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \text{ // 计算更新门的状态} \quad (6)$$

其中 σ 为 sigmoid 函数使得重置门和更新门的值都在［0，1］之间，作为门的开关程度。其中（W_{xr}，W_{hr}，b_r）和（W_{xz}，W_{hz}，b_z）分别为重置门和更新门所需训练的参数

第二步：计算候选隐状态：

$$\dot{H}_t = tanh(X_t W_{xh} + R_t H_{t-1} W_{hh} + b_h) \quad (7)$$

这里使用到了第一步计算的重置门，若重置门为 0 则将上一个隐藏状态全部丢弃。这个设计有助于捕捉时间序列里短期的依赖关系。

第三步：计算当前隐状态：

$$H_t = Z_t H_{t-1} + (1 - Z_t) \dot{H}_t \quad (8)$$

Z_t 为第一步计算的更新门，更新门控制了当前隐状态应当如何与包含当前的时间信息的候选隐状态所更新，若更新门为 1 则将这一时刻的信息全部丢弃，沿用上一个时刻的隐状态。这个设计可以减少循环神经网络中的梯度衰减问题，并更好的捕捉时间序列中时间跨度较大的依赖关系。

2　实验结果与分析

本章重点讲述了本文的训练过程与结果，并对结果进行分析与比较

2.1　数据集

本文实验数据采用几个公开的视频动作识别数据集 UCF50，UCF101 和 HMDB51。UCF50 是采集了来自于网络的真实视频一共有 6680 个视频，每段视频的像素分辨率较低，一共含有 50 类动作。UCF101 数据集是对 UCF50 的扩充，将视频数量扩充到了 13320 个视频，并将动作扩展到了 101 类，包含了诸如，化妆，剪头发，打太极，游泳等几个常见动作。视频中存在相机运动，对象外观和姿势，对象尺度，视点，杂乱的背景，照明条件等的大变化，是最具有挑战性的数据集之一。HMDB51 数据集，共有 6849 个视频，大部分来自于电影片段，包含了人与人交互类，比如拥抱，亲吻等，人与物交互类如拔剑，骑马等 51 类动作。

2.2　训练过程

本文在 Linux 系统下利用 tensorflow 平台搭建网络结构。与通常训练神经网络不同的地方在于，本文对动作识别网络和场景识别网络单独训练，而不是联合在一起训练。主要的原因在于：1. 场景信息是作为提示信息，需要尽可能给予正确信息。2. 场景信息与动作识别信息属于不同层次信息，损失函数值代表的意义不同，因此采取对两个网络单独训练法。

场景识别网络的训练使用了 ImageNet 的参数作为初始化参数，并对数据随机采取了旋转，平移，缩放等数据扩充方法。最后对每张图片裁成 224x224 大小，方便卷积计算。训练采用 Adam 优化算法，最优化交叉熵损失函数。训练结果如图 5 所示。可以看到当迭代次数达到 600 多次时，损失函数值趋向于稳定。此时准确率为 98%，并将此

时的参数固定。随后训练动作识别网络。

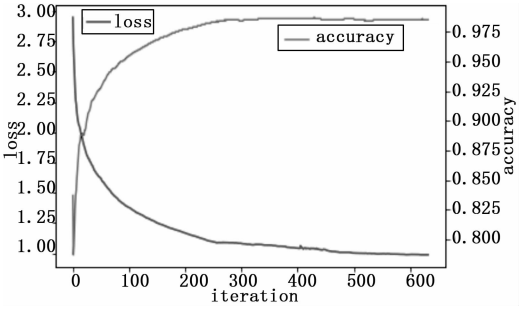

图 5 场景识别网络的损失函数值与准确率

本文将 80%的数据集用于训练动作识别网络，剩余的 20%数据集用于测试。以视频帧作为网络的输入，每次送入 16 个视频作为一个 batch，对每个数据集完整训练 20 遍。为了选取最好的学习率，本文分别测试了学习率为 10^{-3}，10^{-4}，10^{-5}3 种情况收敛时的 loss 值，如表 2 所示，可以看到当学习率为 10^{-4}时 loss 函数最低，因此本文的学习率采取 10^{-4}。

表 2 不同学习率对 loss 值的影响

学习率	Loss 值
10^{-3}	1.32
10^{-4}	0.65
10^{-5}	0.78

2.3 测试结果

两种不同的网络会存在兼容性的问题，为了探究动作识别网络与场景识别网络最好的加权权值。本文对每个网络的置信度分别设置为 1：0.5，1：1 及 1：2 其结果如表 3 所示。从第一行，第二行可以看出，随着场景识别信息的增多，识别的准确率有较大的提升，从而说明了场景信息对动作识别的准确率有促进作用。当场景信息增大到 1：2 时，准确率略有下降，过多的场景信息反而可能会影响动作的识别。从表 4 中的结果可以得到动作识别网络与场景识别网络最佳的权值为 1：1。

表 3 动作识别网络与场景识别网络不同加权值的准确率

网络的加权值	UCF50(%)	UCF101(%)	HMDB51(%)
1:0.5	95.25	92.04	64.11
1:1	95.73	92.81	64.72
1:2	95.66	92.59	64.61

表 4 本文方法与其他一些典型方法比较

方法	UCF50(%)	UCF101(%)	HMDB51(%)
C3D[3]	90.61	83.04	52.44
TwoStream[4]	92.37	88.13	58.21
C3D+lstm[7]	93.59	90.42	61.16
本文的方法	95.73	93.81	65.72

表 4 给出了本文的方法和动作识别中其他一些典型的方法在 UCF50，UCF101 和 HMDB51 数据集上的识别准确率比较。C3D 是典型的使用 3D 卷积神经网络，所以准确率在 UCF101 等大型数据集上表现较差。Two Stream 是通过构建时空与光流的双流网络结构，相比于直接使用 3D 神经网络有了较大的提升。C3D+lstm 是在 C3D 网络的基础上，在其顶部加入了循环神经网络组成的混合网络，改善了 Two Stream 的缺点。从表的结果可知，本文的方法相比于一些其他的方法，在动作识别准确率上有更好的性能。

为了探究场景识别信息对动作识别准确率的影响。本文也分别测试了一些典型方法加入场景识别信息后的准确率，其结果如图 6 所示。可以看到加入场景识别信息之后，网络的识别准确率均有不同程度的提升，从而证明了本文提出的加入场景信息可以有效改善动作识别网路的准确率。

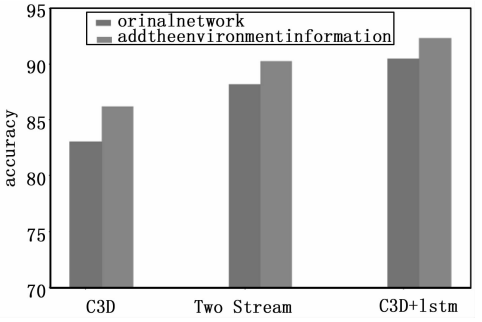

图 6 加入了场景信息后与原来的网络识别准确率对比

3 结束语

在人体动作识别任务上，本文提出了一种基于场景识别的双流网络结构。通过加入场景信息，可以有效改善网络的识别准确率。本文提出的方法在 UCF101 和 HMDB51 数据集上分别取得了 93.81%和 65.72%的准确率，优于一些典型的神经网络识别方法。由于现在网络还需要大量的参数，需要消耗很大的计算资源，还无法运用于摄像头等嵌入式设备中进行实时的识别，未来会在模型压缩方面与运算效率方面做深入研究。

参考文献：

[1] 于萧榕，席 屏，黄健荣. 监控系统预警视频的分布式检索设计与实现 [J]. 计算机测量与控制，2015，23 (7)：2511-2514.

[2] He K，Zhang X，Ren S，et al. Deep Residual Learning for Image Recognition [J]. 2015：770-778.

[3] Du T，Bourdev L，Fergus R，et al. Learning Spatiotemporal Features with 3D Convolutional Networks [A]. IEEE International Conference on Computer Vision [C]. IEEE，2016：4489-4497.

[4] Simonyan K，Zisserman A. Two－Stream Convolutional Networks for Action Recognition in Videos [J]. 2014，1 (4)：568-576.

(下转第 163 页)