文章编号:1002-2082(2018)05-0743-08

基于时空双流卷积神经网络的红外行为识别

吴雪平[1,2],孙韶媛[1,2],李佳豪[1,2],李大威[1,2]

(1.东华大学 信息科学与技术学院,上海 201620;2.东华大学 数字化纺织服装技术教育部工程研究中心,上海 201620)

摘 要:针对红外视频人体行为识别问题,提出了一种基于时空双流卷积神经网络的红外人体行为识别方法。通过将整个红外视频进行平均分段,然后将每一段视频中随机抽取的红外图像和对应的光流图像输入空间卷积神经网络,空间卷积神经网络通过融合光流信息可以有效地学习到红外图像中真正发生运动的空间信息,再将每一小段的识别结果进行融合得到空间网络结果。同时将每一段视频中随机抽取的光流图像序列输入时间卷积神经网络,融合每一小段的结果后得到时间网络结果。最后再将空间网络结果和时间网络结果进行加权求和,从而得到最终的视频分类结果。实验中,采用此方法对包含 23 种红外行为动作类别的红外视频数据集上的动作进行识别,正确识别率为 92.0%。结果表明,该算法可以有效地对红外视频行为进行准确识别。

关键词:人体行为识别;卷积神经网络;信息融合;红外视频;视频分段

中图分类号:TN219 **文献标志码**:A **DOI**:10.5768/JAO201839.0506002

Infrared behavior recognition based on spatio-temporal two-stream convolutional neural networks

Wu Xueping[1,2], Sun Shaoyuan[1,2], Li Jiahao[1,2], Li Dawei[1,2]

(1. School of Information Science and Technology, Donghua University, Shanghai 201620, China;
2. Engineering Research Center of Digitized Textile & Fashion Technology(Ministry of Education), Donghua University, Shanghai 201620, China)

Abstract: Aiming at the recognition of human behavior in infrared video, an infrared human behavior recognition method based on spatio-temporal two-flow convolutional neural network was proposed. In this method, first the entire infrared video is equally segmented, and then the infrared image extracted randomly and the corresponding optical flow image in each video segment are input into the spatial convolutional neural network, and the spatial network can effectively learn which part of the infrared image is actually the action by merging the optical flow information. Next the recognition results of each small segment are merged to get the spatial network results. At the same time, the randomly selected optical stream image sequence in each segment of the video is input into the temporal convolutional neural network, and the result of the temporal network can be obtained by fusing the result of each small segment. Finally, the results of spatial network and the temporal network are weighted and summed to obtain the final video classification results. In the experiment, the action on the infrared video data set containing 23 kinds of infrared behavior action categories was identified by this method, and the correct recog-

收稿日期:2018-05-11; **修回日期**:2018-06-06

基金项目:上海市科委基础研究项目(15JC1400600);国家青年自然科学基金(61603089);上海市青年科技英才扬帆计划(16YF1400100)

作者简介:吴雪平(1993—),男,四川巴中人,硕士,主要从事图像处理与深度神经网络方面的研究工作。E-mail: kopingwu@163.com

导师简介:孙韶媛(1974—),女,博士,教授,主要从事夜视机器视觉方向研究。E-mail: shysun@dhu.edu.cn

nition rate was 92.0%. The results show that the algorithm can effectively identify the infrared video behavior.

Key words: human action recognition; convolutional neural network; information fusion; infrared video; video segmentation

引言

随着视频人体行为识别(human action recognition)在视频安防监控、人体姿态识别、视频检索和人机交互等领域展现的巨大潜力,越来越多的机器视觉研究工作者开始重视如何使计算机自动准确地理解视频人体行为这一问题。

目前,对视频人体行为识别的研究方法大致可以分为两类。第1类是将视频图像的全局或局部特征输入到支持向量机或随机树等分类器中进行动作分类。传统的局部特征主要有:人体空域行为特性的方向梯度直方图(histogram of oriented gradient, HOG)和人体时域行为特性的光流方向直方图[1](histogram of oriented gradient, HOF)、时空兴趣点描述子[2](spatial-time interest point, STIP)、运动边界描述子[3](motion boundary histogram, MBH)及其在三维方向上的扩展,如HOG3D[4]、SIFT3D[5]等。全局特征主要有:人体骨架特征[6-7],包含运动位置信息的运动能量图(motion energy images, MEI)和包含运动位置以及时间信息的运动历史图[8](motion history images, MHI)。最近提出的密集轨迹描述子[9](improved dense trajectories, IDT)由于融合了HOG、HOF、MBH以及特征编码等方法,目前在手工提取特征的算法中效果最为显著。第2类方法是使用最近在目标识别、场景分类等领域具有显著效果的卷积神经网络[10-12]。Tran[13-14]等在网络中使用3D卷积层来直接学习视频的时间和空间信息,再通过全连接层后对动作进行分类。Simonyan[15]等将视频提取为包含空间信息的彩色图像以及包含帧间信息的光流图像后,分别输入2个使用2D卷积层的网络后,再将结果加权求和,也取得了显著的提升效果。Wang[16]等在双流网络的基础上,采用了分段结构,从而解决了视频长度对双流网络分类的影响。Feichtenhofer[17]等人尝试将彩色图像和光流图像在双流网络的中部开始交叉影响,而不是在分类的最后才进行融合。Diba[18]等充分考虑了图像的帧间联系,将各条网络输出的图像结果用双线性特征编码的方式进行分类。

近年对人体行为识别的研究主要集中在可见光领域,对夜视领域的研究相对较少。而很多涉及到人身安全的人体行为动作往往发生在夜晚及黑暗无光的地方,因此,针对夜视情况下基于红外热成像的人体行为识别具有重要意义。

由于红外人体行为识别的研究极少,因此该类数据集的行为种类不够丰富。目前仅有文献[19]中使用的含12种基本红外行为动作的数据集。针对此问题,本文首先构建了含23种红外行为动作类别的数据集,然后实现了基于双流卷积神经网络的红外人体行为识别模型。本模型的输入分别是包含视频空间内容及位置信息的红外图像,以及包含视频时间变化信息的红外光流图像。但该模型与目前文献中的双流网络不同。首先针对传统双流网络较难识别长视频的问题,本模型首先将视频平均分为若干段,然后将每一小段视频中抽取的红外图像和其对应的光流图像输入空间卷积神经网络,之后将每一小段的结果进行融合得到空间网络结果。同时将每一小段视频中抽取的光流图像序列输入时间卷积神经网络,融合每一小段的结果后得到时间网络结果。最后再将两条网络流的结果进行加权求和得到视频行为分类结果,从而充分地利用了整个视频的信息并解决了视频的长短不一问题。其次,空间卷积神经网络所要学习的特征应该是红外图像中发生了运动的那部分信息,为了使空间网络可以更好地关注这部分信息,本模型所构建的空间卷积神经网络由2条网络流组成,将红外图像和对应的光流图像分别作为2条网络流的输入,在学习红外图像空间信息的同时,融合了对应的光流信息特征,这比单纯地识别红外图像的正确率要高。

1 时空双流卷积神经网络

由于从红外视频中抽取的红外图像包含视频图像的空间静态内容信息和位置信息,有助于了解视频的基本内容;光流图像则包含视频帧间动

态变化信息,提供了视频随时间产生的变化信息。因此,本文构建了一个时空双流卷积神经网络,该网络包含2条网络流,分别学习红外视频的空间信息和时间信息,最后再将2条网络流的结果进行加权求和,从而得到对整个红外视频的人体行为分类。

1.1　双流卷积神经网络

由于视频片段的长短不同,因而类似文献[14]所采用的截取部分视频片段输入到网络中的方法,很容易因为未恰好截取到行为动作的视频或者视频中包含了部分其他动作而导致网络结构的错误判断。同时,文献[15]则充分对视频采取了密集采样的方式,导致网络学习的视频信息大幅度增多。事实上,视频中紧邻的图像帧间变化相对极小,因此本文所采用的网络结构,首先对视频进行稀疏采样,即首先将视频平均分为若干段,然后再将若干段视频通过网络后的结果进行融合,既可以充分地利用到整个视频信息,又能有效地降低需要学习的视频信息量。本文所采用的时空双流卷积神经网络结构图如图1所示。

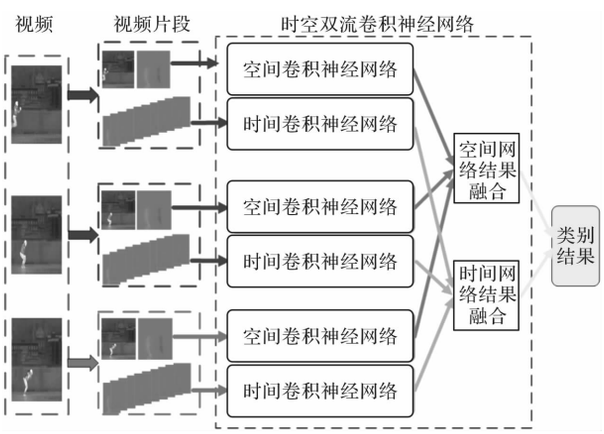

图1　时空双流卷积神经网络结构图

Fig. 1　Structure diagram of spatio-temporal two-stream convolutional neural network

图1中,本文首先将输入的红外视频平均分为K段,可以得到K个长度相同的片段(V_1,V_2,…,V_K),空间卷积网络的结果S_{spatial}可以表示为

$$S_{\text{spatial}}=G(f(t_1;W),f(t_2;W),\cdots,f(t_k;W)) \tag{1}$$

式中:t_1是从V_1中随机抽取的灰度图像和与其对应的光流图像,相应地(t_1,t_2,…,t_K)为从(V_1,V_2,…,V_K)中随机抽取的图像序列;W为空间卷积网络中对应的网络结构参数,通过调整W的值来不断学习视频空间特征;$f(t_k;W)$则代表第K段视频的空间网络输出结果;G函数则是将所有通过空间卷积神经网络后的结果进行融合,本文所采取的融合方式为平均值融合,即将每一段的结果$f(t_k;W)$求和之后取其平均值,从而得到空间卷积网络的结果。时间卷积网络的结果S_{temporal}也可以同理得到:

$$S_{\text{temporal}}=G(f(t_1;W),f(t_2;W),\cdots,f(t_k;W)) \tag{2}$$

最后的视频分类结果S_{labels}则由S_{spatial}和S_{temporal}通过加权求和的方式得到。其中k_1和k_2的取值为正整数,具体数值是使S_{labels}值为最大值的组合,即:

$$S_{\text{labels}}=k_1\cdot S_{\text{spatial}}+k_2\cdot S_{\text{temporal}} \tag{3}$$

1.2　时间卷积神经网络

时间卷积神经网络的输入是包含红外视频动态变化信息的光流图像。本文通过光流提取算法[20]提取出对应的关于横轴和纵轴2个方向的光流图像,因此一帧视频图像将有2张相应的光流图像。由于深度卷积神经网络[21-22]目前在机器视觉中取得了较为显著的效果,因此本文的时间卷积神经网络的基础网络使用了深度卷积网络BN-Inception[21]。BN-Inception网络的详细配置如表1所示。

表1　BN-Inception网络结构的参数配置

Table 1　Parameter configuration of BN-Inception network structure

网络层	窗口大小/滑动步长	输出特征图大小	深度	#1×1	单层 #1×1	单层 #3×3	双层 #1×1	双层 #3×3	池化+卷积
卷积层1	7×7/2	112×112×64	1						
池化层1	3×3/2	56×56×64	0						
卷积层2	3×3/1	56×56×192	1		64	192			
池化层2	3×3/2	28×28×192	0						
Inception(3a)		28×28×256×3	3	64	64	64	64	96	Avg+32
Inception(3b)		28×28×320	3	64	64	96	64	96	Avg+64

续表 1

网络层	窗口大小/滑动步长	输出特征图大小	深度	#1×1	单层 #1×1	单层 #3×3	双层 #1×1	双层 #3×3	池化+卷积
Inception(3c)	stride 2	28×28×576	3	0	128	160	64	96	Max+pass through
Inception(4a)		14×14×576	3	224	64	96	96	128	Avg+128
Inception(4b)		14×14×576	3	192	96	128	96	128	Avg+128
Inception(4c)		14×14×576	3	160	128	160	128	160	Avg+128
Inception(4d)		14×14×576	3	96	128	192	160	192	Avg+128
Inception(4e)	Stride 2	14×14×1 024	3	0	128	192	192	256	Max+pass through
Inception(5a)		7×7×1 024	3	352	192	320	160	224	Avg+128
Inception(5b)		7×7×1 024	3	352	192	320	192	224	Max+128
池化层 3	7×7/1	1×1×1 024	0						

时间卷积神经网络由 1 条 BN-Inception 网络组成,本文以 BN-Inception 网络识别彩色图像的参数作为初始参数进行训练,输入为每一小段视频随机截取的 n 张红外图像所对应的光流图像序列。由于每一张红外图像有对应的横轴和纵轴 2 个方向的光流图像,因此输入光流图像序列的长度为 $2n$。时间卷积神经网络通过对连续的 $2n$ 张包含视频动作变化信息的光流图像进行学习,从而得到时间卷积网络的结果 S_{temporal}。

1.3 空间卷积神经网络

空间卷积神经网络主要是学习视频图像的静态内容信息,由于与红外图像相对应的光流图像包含了该图运动部分的信息,可以帮助空间网络理解图像中的哪一部分正在运动,因此本模型的空间网络输入为每一小段分割视频中随机抽取的红外图像和对应的光流图像。本文提出的空间卷积神经网络结构图如图 2 所示。

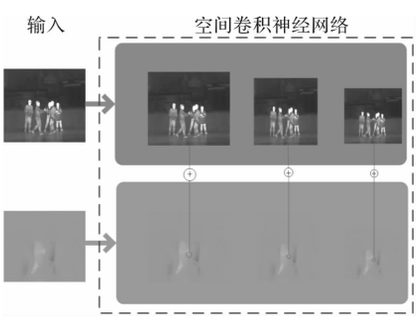

图 2 空间卷积神经网络结构图

Fig. 2 Structure diagram of spatial convolutional neural network

由图 2 可知,尽管输入的红外图像中有较多的人,但是实际发生运动的只有中间 2 个正在握手的人,因此光流图像中只有这 2 个人的光流信息比较明显。因此空间网络通过融合红外图像的光流信息可以很好地关注到图像中真正发生运动的部分信息。

空间卷积神经网络由 2 条 BN-Inception 网络组成,上面一条网络的输入为红外图像(t;224×224×3),下面一条网络的输入为对应的光流图像(t;224×224×3)。在 BN-Inception 网络的传输过程中,将光流图像的运动信息与对应位置的灰度图像信息进行融合,从而帮助空间网络去学习红外图像中真正发生动作的内容信息,可表示为

$$\hat{x}_{l+1}^{a}=f(x_l^a,w_l^a)+f(x_l^m,w_l^m) \tag{4}$$

式中:x_l^a 和 x_l^m 分别代表红外图像网络流和光流图像网络流的第 l 层的输入;w_l^a 和 w_l^m 分别表示各自网络流的学习权重参数;$\hat{x}_{l+1}^{a}$ 表示红外图像网络流第(l+1)层的输入是由红外图像内容特征和光流图像运动特征融合得到的。本文选取的融合点为 BN-Inception 网络的 3c,4e,5b 这 3 个网络输出节点,从而得到空间卷积神经网络的结果 S_{spatial}。

2 实验及结果分析

2.1 实验配置

本实验所使用的硬件及软件配置如表 2 所示。网络使用的是 Pytorch 深度学习框架,在此基础上进行网络训练及学习。

表 2 实验配置

Table 2 Experimental configuration

项目	CPU	内存	GPU	操作系统	CUDA
参数	Intel i5-6600	16 GB	Nvidia GTX 1070	Ubuntu16.10	CUDA8.0

2.2　实验数据

由于夜视领域的红外人体行为视频数据种类相对较少,并且较少将其应用到实际场景中。事实上,夜视红外视频人体行为识别在保护人身安全的功能上极具潜力,监控视频自动准确地判断出危害到人体行为安全的行为并迅速进行报警的话,将会大幅减小受害人的伤害。

基于此实际且又重要的安全应用,本文在文献[19]所创建的12类基本红外人体行为数据集的基础上,加入了7个涉及人身安全的人体行为动作以及4个相对应的未涉及人身安全的人体行为动作作为对比。因此,所建立的基于人身安全的红外人体行为数据集共包含23个动作类别,其中训练集共有690个视频,测试集有223个视频,具体类别见表3所示。

表3　基于人身安全的红外人体行为数据集

Table 3　Infrared human behavior data set based on personal safety

类别	拍手	握手	拥抱	慢跑	双脚跳	拳击	推	单脚跳
标签	0	1	2	3	4	5	6	7
类别	走路	单手挥	双手挥	报警	递东西	拿棍棒	摔倒	打架
标签	8	9	10	11	12	13	14	15
类别	掐脖子	用棍打	拽头发	下跪	晕倒	抢劫	扇耳光	
标签	16	17	18	19	20	21	22	

该数据集是第一个涉及人身安全的红外人体行为数据集,包含12个基本人体行为动作:拍手、握手、拥抱、慢跑、双脚跳、拳击、推、单脚跳、走路、单手挥舞、双手挥舞、打架;包含7个涉及危害到人身安全的红外人体行为动作:用棍棒等武器打人、拽头发、下跪、晕倒不起、抢劫、扇耳光;以及4个相对应的不涉及人身安全,仅作为对比干扰的红外人体行为动作:报警动作、正常地递东西、正常地拿着棍棒等武器、摔倒。

2.3　时间卷积神经网络实验

时间卷积神经网络的输入为n张红外图像所对应的$2n$张光流图像,因此网络的输入流大小为(224×224×2n)。由于BN-Inception网络的初始输入大小为(224×224×3)。本模型采用以BN-Inception网络的参数作为初始参数进行迁移学习,因此本模型第一层的初始参数为

$$\hat{w}_{i,j,k_i}=\frac{(w_{i,j,k_1}+w_{i,j,k_2}+w_{i,j,k_3})}{2n} \tag{5}$$

式中:w_{i,j,k_1}、w_{i,j,k_2}和w_{i,j,k_3}是BN-Inception网络的初始参数;$\hat{w}_{i,j,k_i}$(i=1,2,…,2n−1,2n)是本模型需要迁移学习时所设置的初始参数值。第1层之后的初始参数保持BN-Inception原网络的参数进行训练学习。

时间卷积神经网络所采用的学习率大小为0.001,随机失活参数Dropout为0.9,迁移学习的训练轮数为150轮。为了充分地利用多幅光流图像之间的帧间信息,本文分别测试了使用1张、5张和10张光流图像序列作为时间卷积神经网络的输入,实验结果如表4所示。

表4　时间卷积神经网络特征学习结果

Table 4　Results of temporal convolutional network feature learning

时间卷积神经网络输入	正确率/%
1张光流图像(t;224×224×2)	75.85
5张光流图像(t,t+1,t+2,t+3,t+4;224×224×10)	87.70
10张光流图像(t,t+1,…,t+8,t+9;224×224×20)	87.95

由表4可知,由于1张光流图像所能表示的视频运动信息较少,因此当光流图像序列取5张时的效果要明显好于1张光流图像的识别效果。与此同时,当加入过多的光流图像时,时间卷积神经网络并不能非常有效地进一步提高识别效果。因此本模型采用了5张光流图像序列作为时间卷积神经网络的输入。

2.4　空间卷积神经网络实验

空间卷积神经网络与时间卷积神经网络一样,首先将红外视频平均分为K段短视频,然后从每一段短视频中随机抽取相对应的红外图像和光流图像进入空间网络进行学习,最后将K段短视频的识别结果进行融合,从而得到空间卷积神经网络的识别结果。

由于红外视频的像素值和清晰度低,同时我们并不能确定人体行为动作究竟发生在视频的哪一个时间段内,因此合理地选取K值对实验有较

大的影响。同时考虑到本实验所使用的电脑配置,如果 K 值过大的话,可能导致网络的计算能力大幅度地提升,因此本文考虑了 K 值为 4、8、12、16 这 4 种情况。与此同时,由于 K 值的增大,网络学习过程中相应的批处理量(batch size)也会一定程度地减少。

空间网络的输入为每一段分割视频中随机抽取出的红外图像(t;224×224×3)和与其对应的光流图像(t;224×224×3),将它们分别输入 2 条 BN-Inception 基网络进行迁移学习。由于其输入与原始 BN-Inception 网络的输入相同,因此本模型的初始参数可以直接采用 BN-Inception 的原始参数作为训练学习。

空间网络所采用的学习率大小为 0.001,随机失活参数 Dropout 为 0.8,迁移学习的训练轮数为 200 轮。尽管红外图像能够很好地表达视频中的基本内容,但是对于人体行为动作来说,更应该关注的是红外图像中发生了运动的信息。因此本模型的空间网络采用 2 条网络流来进行信息融合,其与单独学习红外图像的实验以及 K 值变化的实验结果如表 5 所示。

表 5 空间卷积神经网络特征学习结果

Table 5 Results of spatial convolutional network feature learning

空间卷积神经网络输入	正确率/%
红外图像(t;224×224×3);K=4,batch=12	59.8
红外图像(t;224×224×3);K=8,batch=8	**68.5**
红外图像(t;224×224×3);K=12,batch=6	68.3
红外图像(t;224×224×3);K=16,batch=4	62.4
红外图像(t;224×224×3)和光流图像(t;224×224×3);K=4,batch=12	71.8
红外图像(t;224×224×3)和光流图像(t;224×224×3);K=8,batch=8	77.9
红外图像(t;224×224×3)和光流图像(t;224×224×3);K=12,batch=6	**81.0**
红外图像(t;224×224×3)和光流图像(t;224×224×3);K=16,batch=4	75.8

由于视频分割的次数越多,模型就越不容易丢失掉视频中动作发生时的图像。由表 5 可知,无论是采用单独识别红外图像还是使用本模型的方法,随着视频分割数 K 值的增大,识别的正确率都出现了一定的增长。同时,由于受实验电脑配置的原因,随着 K 值的增大,批处理量(batch size)的减少使得模型的过拟合现象可能加重。因此当 K 值到达 16 之后,正确率开始出现了一定程度的下降。因此本文所使用的分割数 K 值为 12。

由于本模型将红外图像对应的运动信息融合进网络进行学习,可以了解到红外图像中究竟哪些是静止的信息,哪些是动态的信息,从而可以帮助网络更准确地对动作进行分类。由表 5 可知,本模型比单纯地识别红外图像的准确率提高了近 13%。

2.5 双流卷积神经网络实验

本文所提出的时空双流卷积神经网络的输出结果是由空间卷积神经网络和时间卷积神经网络的结果进行加权求和得到。本实验所使用的空间网络结果(81.0%)和时间网络结果(87.7%)的权值比为 1∶1,因此最终得到的识别正确率为 92.0%。

图 3 为本模型在本文所建立的基于人体行为安全的红外行为数据集上的每一类动作的识别准确率。

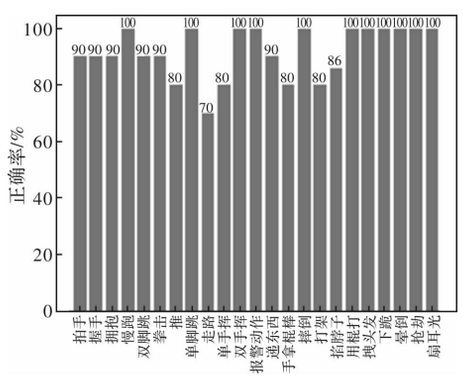

图 3 数据集的每一类动作的识别正确率

Fig. 3 Accuracy of recognition for each type of data set

由图 3 可知,本模型在具有实际应用的基于人身安全的红外数据集上取得了非常准确的效果,其中效果最差的是行走这一个类别,由于其他视频片段中,一般都包含行走这一个简单动作,因此其正确率最低,但是其正确率也达到了 70%。充分地说明了本模型无论是针对极具相似性的简单动作,还是针对实际中可能危害到人体行为的复杂动作都能做出较高的准确判断。

本文跟目前研究文献里的对比实验结果如表 6 所示。由表 6 可知,本文所提出的时空双流卷积神经网络模型相对于传统的最好的密集轨迹算法[9]以及目前的最优算法 TSN(temporal segment networks)[16]都有着不同程度的提升,充分地说明了本算法可以更高效地识别出红外视频人体行

为。同时采用的神经网络结构不仅可以省去大量的手工特征计算,而且可以大幅度地缩短识别时间。本文提出的算法不仅可以准确地识别极具相似性的人体简单动作,也可以在涉及人身安全等实际人体行为识别的场景中有较为准确的识别,从而可以通过视频监控分析采取措施,以减少对人身安全的伤害行为。

表 6　对比实验结果

Table 6　Comparison experiment results

算法	正确率/%
密集轨迹算法[9]	69.8
自适应融合算法[18]	72.5
传统双流网络[14]	86.4
时序分段网络[15]	90.5
本文算法	**92.0**

3　结论

本文提出了一种基于时空双流卷积神经网络的红外人体行为识别方法,该方法解决了传统双流模型难以处理较长视频以及未充分考虑视频信息和光流信息之间关系的这 2 个问题。通过对视频进行平均分段并随机抽取视频图像序列的方法,使得模型可以有效地利用较长视频的所有信息。同时通过在空间卷积网络中融合了对应红外图像的光流信息,使得空间卷积神经网络可以更好地理解图像中真正发生运动的信息。实验结果表明,本文比现有的最优识别结果依然能提高 1.5%。在接下来的工作中,我们将考虑如何使双流网络真正融合为一个网络,使空间信息和时间信息能够在网络训练的时候就能够有效地融合,从而进一步提高红外视频的人体行为识别准确率。

参考文献:

[1] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies[C]// IEEE Conference on Computer Vision and Pattern Recognition. USA:IEEE,2008:1-8.

[2] LAPTEV I. On space-time interest points[J]. International Journal of Computer Vision, 2005,64 (2/3): 107-123.

[3] DALAL N, TRIGGS B, SCHMID C. Human detection using oriented histograms of flow and appearance [C]. Berlin:Springer, 2006.

[4] KLÄSER A, MARSZALEK M, SCHMID C. A spatio-temporal descriptor based on 3D-Gradients[C]. Leeds:DBLP, 2008.

[5] SCOVANNER P, ALI S, SHAH M. A 3-dimensional sift descriptor and its application to action recognition[C]. Leeds:DBLP, 2007:357-360.

[6] WU D, SHAO L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition[C]. USA:IEEE, 2014: 724-731.

[7] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition [C]. USA:IEEE, 2015: 1110-1118.

[8] BOBICK A F, DAVIS J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on Pattem Analysis & Machine Intelligence, 2001, 23(3):257-267.

[9] WANG H, SCHMID C. Action recognition with improved trajectories [C]. USA: IEEE, 2014: 3551-3558.

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. International Conference on Neural Information Processing Systems, 2012, 60 (2): 1097-1105.

[11] XU Lu, ZHAO Haitao, SUN Shaoyuan. Monocular infrared image depth estimation based on deep convolutional neural network[J]. Acta Optica Sinica, 2016, 36(7): 196-205.
许路,赵海涛,孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报,2016,36(7): 196-205.

[12] GAO Kaijun, SUN Shaoyuan, YAO Guangshun, et al. Semantic segmentation of night vision images for unmanned vehicles based on deep learning[J]. Journal of Applied Optics, 2017,38(3):421-428. .
高凯珺,孙韶媛,姚广顺,赵海涛. 基于深度学习的无人车夜视图像语义分割[J]. 应用光学,2017,38(3): 421-428.

[13] TRAN D, BOURDEV L D, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//International Conference on Computer Vision. USA: IEEE, 2015: 4489-4497.

[14] VAROL G, LAPTEV I, SCHMID C. Long-term temporal convolutions for action recognition [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2016, 99:1-1.

[15] SIMONYAN K, ZISSERMAN A. Two-stream convo-

lutional networks for action recognition in videos[J]. Computational Linguistics, 2014, 1(4):568-576.

[16] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[J]. Acm Transactions on Information Systems, 2016, 22(1):20-36.

[17] FEICHTENHOFER C, PINZ A, WILDES R P, et al. Spatiotemporal residual networks for video action recognition[C]//Neural Information Processing Systems. USA: ARXIV, 2016: 3468-3476.

[18] DIBA A, SHARMA V, VAN GOOL L, et al. Deep temporal linear encoding networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. USA:IEEE, 2017: 1541-1550.

[19] GAO C, DU Y, LIU J, et al. A New dataset and evaluation for infrared action recognition[C]. Berlin: Springer, 2015:302-312.

[20] GAO Y, BEIJBOM O, ZHANG N, et al. Compact bilinear pooling[C]//IEEE Conference on Computer Vision and Pattern Recognition. USA:IEEE, 2016: 317-326.

[21] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. USA:ARXIV, 2015: 448-456.

[22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2016: 770-778.