基于片段关键帧的视频行为识别方法

李鸣晓[1]，庚琦川[1]，莫红[1]，吴威[1]，周忠[1]

（[1]北京航空航天大学 虚拟现实技术与系统国家重点实验室，北京市 100191）

摘　要：视频行为识别是智能视频分析的重要组成部分。近年来，深度学习方法在该领域有了显著的进步，目前得到最佳效果的方法都使用了双流卷积神经网络。现有的行为识别方法大多以均匀采样得到的视频帧作为输入，这可能损失采样间隔中的重要信息。本文提出了一种用于视频行为识别的片段划分和关键帧提取方法，并将两种方法与双流网络结合，在 UCF101 数据集 split1 上追平了目前最高的 94.2%准确率。

关键词：深度学习；行为识别；视频片段划分；关键帧提取；光流；边缘检测

文章编号：1004-731X (2002) xx-xxxx-x　　**中图分类号**：　　**文献标识码**：

A method of key-frame based video action recognition

Li Ming-Xiao[1]，*Geng Qi-Chuan*[1]，*Mo Hong*[1]，*Wu Wei*[1]，*Zhou Zhong*[1]

([1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, 100191, China)

Abstract：Video action recognition is an important part of intelligent video analysis. In recent years, deep learning methods, especially the two-stream convolutional neural network achieved state-of-the-art performance. However, most methods simply use uniform sampling to get frames, which may cause the loss of information in sampling interval. In this paper we proposes a segmentation method and a key-frame extraction method for video action recognition, and combine them with a two-stream network. Our framework achieves a 94.2% accuracy at UCF101 split1, which is the same as the state-of-the-art method's performance.

Keywords：deep learning; action recognition; video segment; key-frame extraction; optical-flow; edge detection

0 引　言[1]

视频行为识别是智能视频分析的重要组成部分，其核心是把视频拆分为图像序列，通过提取序列的时空特征进行分类。早期该领域的研究主要采用人为设计的时空特征描述子进行行为分类[5]。近年来随着深度神经网络在图像识别领域的巨大成功，使用深度神经网络进行视频行为识别也取得了显著效果。

视频和静态图像的主要区别之一在于视频拥有时间域。为了提取时间域特征，视频行为识别需要处理大量视频帧并得到其中隐含的联系。现有的研究通常将视频划分为若干片段，对每一个片段选取待检测帧提取特征，最后将片段特征进行融合。早期的方法使用单独的卷积神经网络（Convolutional Neural Network，CNN）对静态视频帧进行特征提取，对提取结果求均值得到分类结果。双流网络引入了光流作为辅助输入，使用两个CNN结构分别提取片段内的时间空间信息[2]。LRCN[6]引入了循环神经网络（Recurrent Neural Network，RNN）提取时间特征。上述方法使用均匀采样得到待检测帧。现有的大多数研究关注如何提取时间信息，很少关注待检测帧的质量。待检测帧是否包含当前行为的关键信息是一个影响行为识别结果的重要因素。这里我们称包含关键信息的视频帧或光流序列为关键帧，称采样间隔为一个视频片段。

视频和静态图像的区别不仅仅是时间域。由于视频需要连续拍摄，容易存在运动模糊、焦距丢失等干扰，甚至视频压缩也会导致噪声，不同视频帧之间存在信息量的差异；同时由于自然行为运动不均匀，不同视频片段间包含的运动信息量往往差别巨大。传统的视频关键帧提取方法大多基于亮度直方图与梯度直方图等全局特征[13]，难以找出场景变化较小的行为视频中含有关键信息的片段。

为了解决上述问题，我们提出了一种基于运动信息量的

收稿日期：2017-07-30　　**修回日期**：2017-10-09

基金项目：国家“八六三”高技术研究发展计划 项目基金(No.2015AA016403)、国家自然科学基金项目(No.61572061、No. 61472020)

作者简介：**李鸣晓**(1993-)，男，硕士，研究方向为深度学习、行为识别；**庚琦川**(1989-)，男，博士，研究方向为图像语义理解；**莫红**(1988-)，女，博士，研究方向为深度学习、图像描述；**吴威**(1961-)，男，教授，中国计算机学会(CCF)高级会员，博士，研究方向为研究领域网络与信息安全技术、分布式虚拟现实与可视化技术、多智能体系统；**周忠**(1978-)，男，教授，中国计算机学会(CCF)高级会员，博士，研究方向为虚拟现实等。

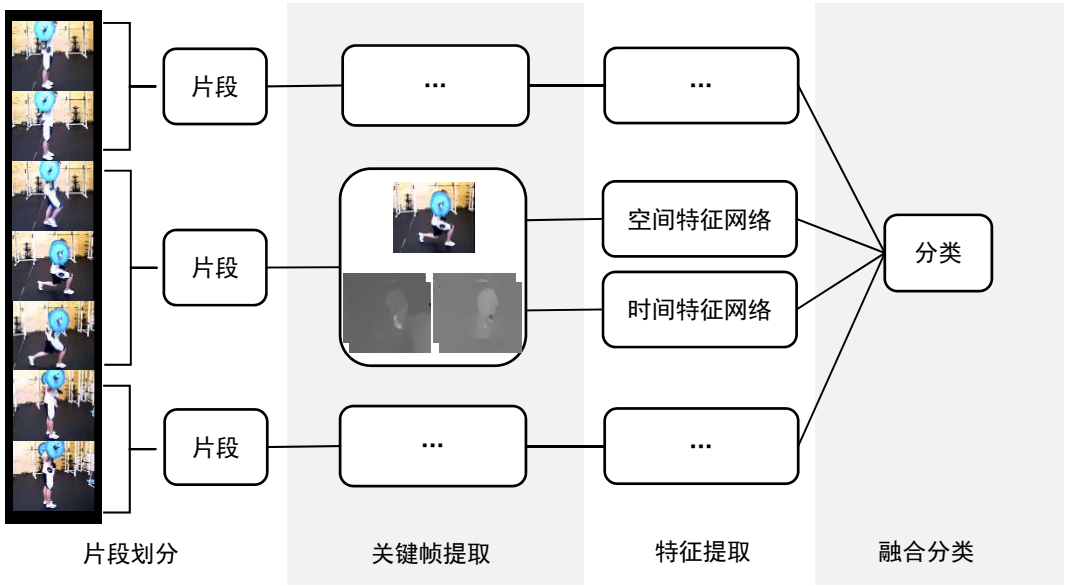

图 1　本文采用的视频行为识别系统结构。

视频片段划分方法和一种基于图像信息量的片段内关键帧提取方法。在两者的基础上，结合双流网络，设计了一个使用多尺度光流网络的行为识别系统。该系统在 UCF101 数据集 split1 上表现良好，达到了 94.2%的准确率，和 Temporal Inception[8]准确率相同。

1　相关研究

随着深度卷积神经网络在图像识别领域取得巨大成功，研究者开始将视线转移到视频行为识别领域。使用深度网络进行行为识别已经取得了较大的进展。当前的深度学习方法在均匀划分视频片段后，通常由两部分组成：视频片段特征提取，视频级别的特征融合。

现有工作大多将CNN作为视频的特征提取方式[1, 2, 3, 4]。他们划分视频片段，从片段中提取一个或多个视频帧作为CNN的输入，再对片段的特征进行融合。早期研究以单帧作为CNN的输入。这种做法对时间信息的利用不够充分。部分研究尝试直接使用CNN提取视频特征[3]。Tran等人训练了大规模的 3DCNN网络[4]。这类做法在不同层次增加了CNN的维度，以视频序列直接作为CNN的输入。受限于卷积网络的大小，3DCNN不能处理不定长度的视频，本质上仍然无法避免视频的分段识别。Donahue等人提出了LRCN结构[6]，在使用CNN提取独立视频帧特征的基础上，引入了LSTM（一种更有效的RNN结构）以融合每一帧提取到的特征。LSTM与3DCNN的训练的时间与存储代价较大。为了提取时间信息，Karen等人提出了双流法[2]。该方法以稠密光流[17]作为辅助输入，使用两个独立的卷积神经网络分别提取单帧原始图像和多帧光流图像特征，在最后的评分层次进行融合。双流法通常使用图像识别任务使用的网络，计算规模和图像识别任务相当。光流的引入使得行为识别准确率有了显著的提高，双流网络成为主流。此后的研究大多在双流网络的基础上改进深度神经网络，或在双流网络特征的基础上使用各种方法融合特征。2016 年Wang等人提出的TSN（Temporal Segment Networks）是双流网络的一个有力改进[1]，现有的方法大多以此为衡量标准。该框架引入了光流的预训练模型；同时提出了一种预先划分视频片段，片段间使用池化融合的训练方法。但TSN结构采用均匀划分片段，忽略了片段间的信息量差异。

部分人尝试使用关键帧的思想提高行为识别的效果。Hu等人采用光流差分的方式提取视频中的关键帧，在KTH数据集取得了一定效果[7]。KVMF一文提出了关键卷的概念，证实了不同的视频片段选取对于最终视频分配结果有一定贡献[9]。但是上述工作未对完整视频的片段划分方式进行研究，也未对关键帧对行为识别的准确率影响进行研究。

本文以双流的TSN结构为基础，提出了一种基于运动信息量的片段划分方法，可用于网络的训练与检测；设计了一种基于图像信息量的片段内关键帧提取方法。本文对传统光流网络进行了改进，设计了多尺度的光流网络。

2　片段划分与关键帧提取

为了进行特征的提取，首先需要选定待检测帧。我们通过划分视频片段后选取片段内关键帧的方式确定待检测帧。首先使用运动信息量进行视频的片段划分，然后使用图像信息量评价的方法选定待检测帧。

2.1　基于运动信息量的片段划分

行为的运动信息往往在时间域上不均匀。如图 2 所示，

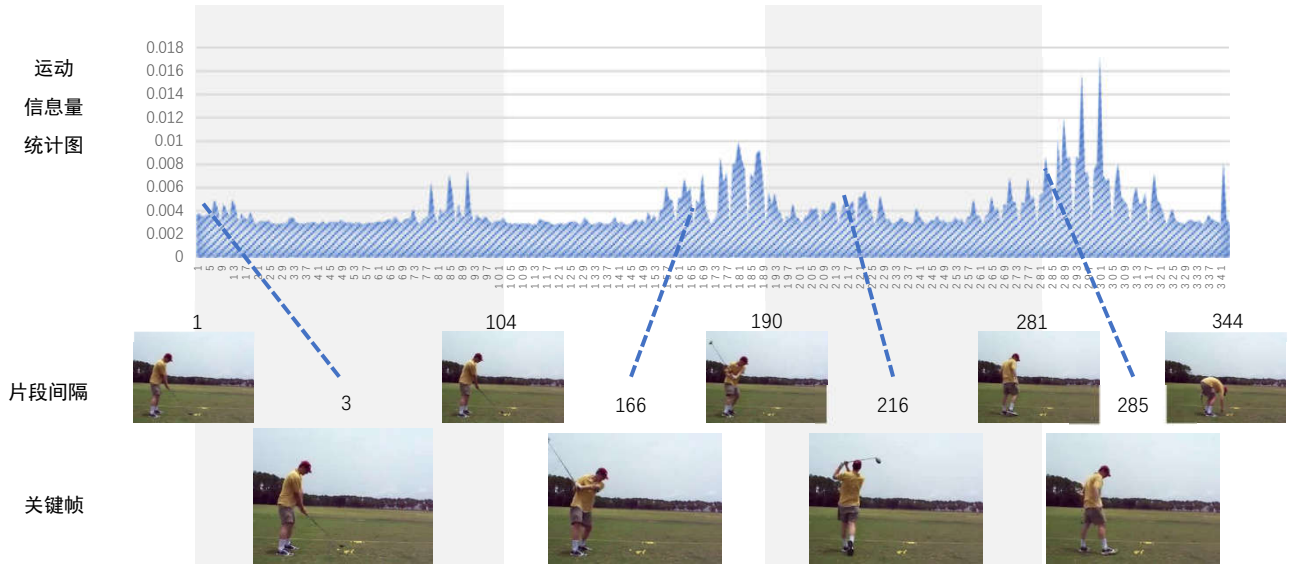

图 3 GolfSwing_g04_c05，使用信息量方法的完整视频的片段划分与图像关键帧提取。

在高尔夫球运动视频中，击球准备时间是持续了 156 帧，击球动作持续到 201 帧。均匀分段采样容易遗漏采样间隔中的击球行为。

图 2 UCF101，GolfSwing_g04_c05，视频信息不均

广义的视频行为识别包含三部分：检测，追踪和识别[16]。本文只讨论识别。在已经检测到行为主体的情况下，主体将占据有效视频的绝大部分空间，此时稠密光流能够较好的表达主体的运动状态。稠密光流的定义式如下[17]：

$$T(x,y) = I(x+u, y+v) \tag{1}$$

其中T为过去帧，I当前帧，u、v分布表示像素在x、y轴上的偏移量。光流图像为u、v对应坐标x、y的矩阵。当主体运动明显时，光流图像应有较高的绝对值；同时主体显著运动时，往往会包含更多的行为信息。

行为信息在视频中的分布是不均匀的。为了使得视频分片段的运动信息量较为均匀，在片段数N_s固定的情况下，需要最小化片段的方差D_M。D_M的定义如下所示：

$$D_M = \sum_i (M_i - \bar{M})^2 / N_s \tag{2}$$

上式M_i表示第i个片段的的运动信息量，$\bar{M}$为平均片段信息量，每个片段的信息量M由下式表出：

$$M = \sum_i \sqrt{\sum_c \sum_{x,y} flow_i(x,y,c)^2} \tag{3}$$

其中$flow_i(x,y,c)$是片段内对应的二通道光流图像，c表示光流的通道。

使用动态规划能够得到精确解，但通常情况下光流会随着视频质量波动，引入误差。为了运算效率，我们使用贪心算法得到了近似解。

2.2 基于图像信息量的关键帧提取

深度神经网络的预训练模型能够使网络的效果得到极大提高。一般认为这是因为网络学习了大型图像数据集的特征。大型图像数据集的图像都较为清晰，对模糊图像特征的学习较少。模糊图像可能会影响最终的识别结果。此外，含有更多物体的图像通常含有更多信息。为了解决上述问题，我们使用图像信息量评价的方法，在视频片段中找到含有最大信息量的图像帧。

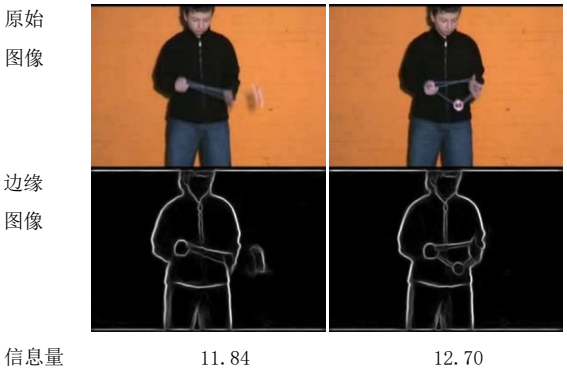

图 4 UCF101，YoYo_g03_c02，边缘强度对比

大量研究表明，人类视觉比较关注图像的边缘部分。许多基于图像边缘提取的对象检测方法取得了良好的效果[11]。我们认为图像的平均边缘强度能够较好的代表图像的信息量。定义E_i为第i帧图像的信息量：

$$E_i = \frac{\sum_{x,y} edge_i(x,y)}{img\_size} \tag{4}$$

其中$edge_i(x,y)$为由视频帧得到的灰度边缘图像，

img_size为图像像素数。我们使用Dollar等人提出的边缘提取方法[10]得到灰度边缘图像。该方法运行较快，同时图像的灰度值能够平滑的表示物体的模糊程度。一个提取效果和信息量评价的示例如图 4 所示，左图存在较大的运动模糊，右图较为清晰。在主体几乎不变的情况下，右图的打分较高。

最终取片段内信息量最大的帧作为关键帧，关键帧的片段内序号为：

$$Index = argmax(E) \tag{5}$$

3 行为识别系统结构

3.1 特征提取网络架构

在实际实验中，深度神经网络容易受限于设备的性能。C3D等三维卷积网络的显存占用与计算量是同深度的图像网络的数倍乃至数十倍，LRCN等RNN结构的开销受到时间步长影响过大。双流法分别使用图像识别网络构建空间特征网络与时间特征网络，不仅在行为识别的方面表现强大，同时显存占用和计算规模和常用的图像分类网络规模相当。双流法能够使用在大型图像数据集上的预训练模型，这不仅大幅提升了分类效果，还能够防止数据的过拟合并加速收敛。本文采用双流网络提取视频特征，以视频中每个片段的关键帧作为输入。

3.2 时空特征网络的构建

3.2.1 空间特征网络

空间特征网络即图像网络。该网络选用在图像分类任务上表现良好的BN-Inception结构[14]。Inception结构使用了不同尺度的卷积核，这使得Inception网络拥有较强的多尺度细节识别能力。批正则化（Batch Normalization，BN）的引入使得该结构收敛更加迅速、准确。该网络输入为固定尺寸的单帧RGB图像。本文使用了在Pascal VOC 2012 数据集[18]上得到的预训练模型。

3.2.1 时间特征网络

时间特征网络即光流网络。该网络选用修改的BN-Inception结构。在原网络的基础上，对网络的第一个卷积层权值进行通道拓展，使得输入支持更多通道。具体步骤为如下：首先将原始模型在第一个卷积层（conv1）上的卷积核参数沿通道求和；将得到的参数和除以目的通道数，沿通道复制叠加，作为新的conv1 层参数。该网络输入为固定的 10 通道，由 5 帧的光流图像堆叠得到。

由于现有的光流网络仅提取连续有限帧的信息，我们引入了多时间尺度的光流网络用于提取不同时间尺度的特征。通过在不同时间跨度上均匀采样，我们得到了同一结束时刻的多个等长光流序列，将其分别叠加，作为输入训练光流网络。在测试时，将多尺度光流序列分别测试，使用均值池化得到最终结果。如图 5 所示，对光流图像序列在不同时间尺度上进行采样。

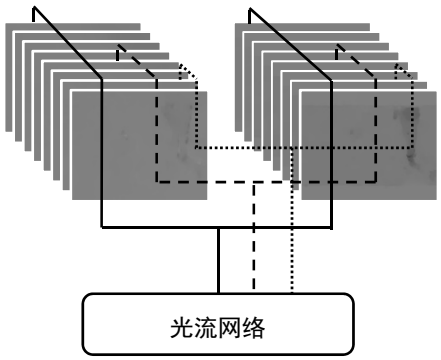

图 5　多时间尺度光流网络示意图

3.3 特征融合

时空网络的训练结构如图 6 所示，在每个视频片段中选取帧作为输入，以对应网络最后一个卷积层的输出作为该网络的输出，使用均值池化进行融合。

在训练时，基于运动信息量划分视频片段，从片段中随机选取视频帧作为输入，对每个片段得到的时空特征，即卷积层输出，分别进行均值池化，将池化结果作为损失函数的输入计算损失，进行反向传播。我们设计的时空网络的损失函数均使用$softmax$交叉熵，使用随机梯度下降(Stochastic gradient descent，SGD)进行训练。

在测试时，划分更多的片段，片段内根据图像信息量选取关键帧，对关键帧进行图像过采样并进行检测，对得到的分类结果取均值池化，最终经过$argmax$函数得到预测结果。

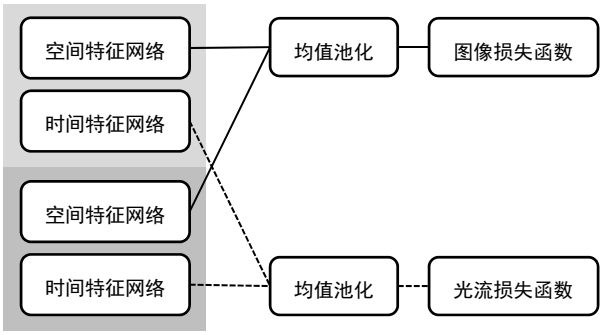

图 6　时空网络的训练结构图

4 实验与分析

UCF101 数据集是当前使用最多的视频行为识别数据集。该数据集包含 101 个分类，13320 段视频[12]。该数据集最常用的训练集测试集分组为split1。

4.1 数据预处理

4.1.1 重复帧的移除

低帧率的视频强行提高帧率会引入重复帧。重复帧一定程度上能表达部分时间特征，但会对光流的提取造成影响。使用重复帧提取的光流会近似空白。此前的工作对此未做特殊处理，仍将空白光流作为输入。为了有效划分视频片段，本文对重复帧进行了剔除，同时使用多尺度的光流网络防止时间信息丢失。

在训练参数相同，视频分段数为 25 时，是否移除空白帧的光流网络准确率对比见表 1；其中片段和网络间的融合方式为均值池化；该实验未使用预划分片段。

表 1 重复帧移除前后光流网络测试结果

包含重复帧	时间间隔（帧）	准确率
是	1	83.1%
否	1	83.1%
否	(1,2)	83.4%
是 + 否	1 + 1	84.0%
是 + 否	1 + (1,2)	84.1%

移除重复帧能够使网络学习到不同的特征，但对网络总体准确率提升并不大。在更大的数据量下，移除重复帧的多尺度光流网络应当能体现出更好的性能。

4.1.2 视频片段划分与关键帧提取

基于运动信息量的片段划分不依赖于神经网络提取的特征，能够为数据集提前生成固定的片段划分信息，可以作为一种预处理。在为训练集提取光流时，一并计算视频分段信息与图像信息量信息。

视频片段划分能够应用于训练和测试过程中。而片段内的关键帧提取仅应用于测试中。这是因为深度学习的训练过程中，需要对分段内的帧进行随机抽取训练，以学习更多的特征。关键帧提取会限制特征的学习。

4.1.3 光流关键帧提取

在双流网络中，空间特征网络和时间特征网络在特征提取过程中是独立的。图像关键帧的选取不影响光流关键帧的选取。因此也需要为光流序列设计关键帧选取方式。本文尝试使用滑动窗口的方式对光流序列求和，计算序列内的运动信息量最大值，然而实验证明，这种约束会明显降低分类效果，时间特征网络准确率降低 3%-5%，融合结果降低 1%-2%。寻找合适的光流关键帧提取方法将会是下一步的研究方向。

4.2 训练细节

许多实验证明，分段融合的训练方法能够有效提高网络表达能力[1,8]，但并非分段数越多效果越好。训练时会进行反向传播，中间数据需要保留，在显存一定时，视频片段数量和计算梯度时批次大小成反比。批次过小会导致收敛缓慢。训练时需要在片段数量与批次大小中寻求平衡。本文训练时使用的片段数为 3，批次大小为 32，占用约 9G显存。

本文使用caffe[19]框架进行实验，并在TensorFlow[20]下实现了接近的效果。

4.3 测试结果

4.3.1 视频测试结构

如表 2 所示，本文测试了片段数在 3、10、25 时的行为识别准确率。表中数据为融合双流网络后的结果：使用单时间尺度的光流网络，已移除空白帧；网络间使用 1:1 的均值融合。

表 2 UCF101 数据集Split1 测试结果

测试方法\片段数	3	10	25
均匀分段+取段首	92.2%	93.9%	93.7%
均匀分段+关键帧	92.7%	94.1%	93.9%
信息量分段+关键帧	92.7%	94.2%	94.0%

在分段数为 10 时网络得到最佳结果。在不同片段数下，基于信息量的分段和关键帧选取方法都能提升网络效果。

4.3.2 同类工作比较

表 3 UCF101，Split1，最佳效果的对比

方法名	年份	准确率
深度双流[15]	2015	90.9%
TSN[1]	2016	93.5%
Temporal Inception[8]	2017	94.2%
信息量分段+关键帧	2017	94.2%

引入信息量的片段划分方法与片段内关键帧提取的方法后，本文提出的方法达到了最高的 94.2%的准确率，超过了使用均匀分段去段首方法的TSN结构，和同样未使用额外数据的Temporal Inception（TI）方法持平。TI使用双流网络得到片段特征，通过训练一个时间域的Inception网络对特征进行融合，其核心在片段间特征融合。本文在仅使用均值融合的情况下达到了同样的准确率，说明本文使用的方法应当还有提升空间，有待进一步的实验发掘。

5 总结与展望

通过分析当前视频行为识别方法，我们提出了一种基于运动信息量的视频片段划分方法和一种基于图像信息量的片段内关键帧提取方法。在两者的基础上，结合双流网络，我们设计了一个使用多尺度光流网络的行为识别系统。在 UCF101 数据集上的实验证明了我们提出的片段划分和图像关键帧选取方法切实提高了行为识别的准确率，同时追平了数据集分片上的最佳结果。

我们设计的行为识别系统使用均值池化融合片段间特征。为了更好的利用时间域信息，我们将参考Temporal Inception等片段特征融合的工作，下一步研究视频分段与特征融合方法的相互影响。同时我们也会尝试寻找一个有效的光流关键帧提取方法。

参考文献:

[1] 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.

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

[3] Karpathy A, Toderici G, Shetty S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:1725-1732.

[4] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[J]. 2014:4489-4497.

[5] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893.

[6] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[M]// AB initto calculation of the structures and properties of molecules /. Elsevier, 2015:85-91.

[7] Hu Y, Zheng W. Human Action Recognition Based on Key Frames[M]// Advances in Computer Science and Education Applications. Springer Berlin Heidelberg, 2011:535-542.

[8] Ma C Y, Chen M H, Kira Z, et al. TS-LSTM and Temporal-Inception: Exploiting Spatiotemporal Dynamics for Activity Recognition[J]. 2017.

[9] Zhu W, Hu J, Sun G, et al. A Key Volume Mining Deep Framework for Action Recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2016:1991-1999.

[10] Dollár P, Zitnick C L. Structured Forests for Fast Edge Detection[J]. 2013:1841-1848.

[11] Zitnick C L, Dollár P. Edge Boxes: Locating Object Proposals from Edges[J]. 2014, 8693:391-405.

[12] Soomro K, Zamir A R, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. Computer Science, 2012.

[13] Zhuang Y, Rui Y, Huang T S, et al. Adaptive key frame extraction using unsupervised clustering[C]// International Conference on Image Processing, 1998. ICIP 98. Proceedings. IEEE, 2002:866-870 vol.1.

[14] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. Computer Science, 2015.

[15] Wang L, Xiong Y, Wang Z, et al. Towards Good Practices for Very Deep Two-Stream ConvNets[J]. Computer Science, 2015.

[16] Poppe R. A survey on vision-based human action recognition.[J]. Image & Vision Computing, 2010, 28(6):976-990.

[17] Farnebäck G. Two-Frame Motion Estimation Based on Polynomial Expansion[C]// Scandinavian Conference on Image Analysis. Springer-Verlag, 2003:363-370.

[18] Everingham M, Eslami S M A, Gool L V, et al. The Pascal, Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111(1):98-136.

[19] Jia Y, Shelhamer E, Donahue J, et al. Caffe:Convolutional Architecture for Fast Feature Embedding[J]. 2014:675-678.

[20] Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems[J]. 2016.