基于多阶信息融合的行为识别方法研究

张冰冰 [1] 葛疏雨 [1] 王旗龙 [2] 李培华 [1]

摘 要 双流卷积神经网络能够获取视频局部空间和时间特征的一阶统计信息, 测试阶段将多个视频局部特征的分类器分数平均作为最终的预测. 但是, 一阶统计信息不能充分建模空间和时间特征分布, 测试阶段也未考虑使用多个视频局部特征之间的更高阶统计信息. 针对这两个问题, 本文提出一种基于二阶聚合的视频多阶信息融合方法. 首先, 通过建立二阶双流模型得到视频局部特征的二阶统计信息, 与一阶统计信息形成多阶信息. 其次, 将基于多阶信息的视频局部特征分别进行二阶聚合, 形成高阶视频全局表达. 最后, 采用两种策略融合该表达. 实验表明, 本文方法能够有效提高行为识别精度, 在 HMDB51 和 UCF101 数据集上的识别准确率比双流卷积神经网络分别提升了 8 % 和 2.1 %, 融合改进的密集点轨迹 (Improved dense trajectory, IDT) 特征之后, 其性能进一步提升.

关键词 行为识别, 双流卷积神经网络, 多阶信息融合, 二阶聚合

引用格式 张冰冰, 葛疏雨, 王旗龙, 李培华. 基于多阶信息融合的行为识别方法研究. 自动化学报, 2021, **47**(3): 609−619

DOI 10.16383/j.aas.c180265

Multi-order Information Fusion Method for Human Action Recognition

ZHANG Bing-Bing[1] GE Shu-Yu[1] WANG Qi-Long[2] LI Pei-Hua[1]

Abstract The classical two-stream convolutional neural network (CNN) can capture the first-order statistics of the local spatial and temporal features from an input video, while making final predictions by averaging the softmax scores of the local video features. However, the first-order statistics can not fully characterize the distribution of the spatial and temporal features, while higher-order information inherent in local features is discarded at the test stage. To solve the two problems above, this paper proposes a multi-order information fusion method for human action recognition. To this end, we first introduce a novel two-stream CNN model for capturing second-order statistics of the local spatial and temporal features, which, together with the original first-order statistics, forms the so-called multi-order information. We perform individually second-order aggregation of these extracted local multi-order information to compute global video representations. Finally, two strategies are proposed to fuse video representations for prediction. The experimental results demonstrate that our proposed method significantly improves recognition accuracy over the original two-stream CNN model, i.e., 8 % and 2.1 % gains on the HMDB51 and UCF101, respectively. The performance of our method is further improved by combining traditional IDT (improved dense trajectory) features.

Key words Human action recognition, two-stream convolutional neural network, multi-order information fusion, second-order aggregation

Citation Zhang Bing-Bing, Ge Shu-Yu, Wang Qi-Long, Li Pei-Hua. Multi-order information fusion method for human action recognition. *Acta Automatica Sinica*, 2021, **47**(3): 609−619

行为识别在智能监控、人机交互和视频检索等领域中得到了广泛的应用, 引起了众多研究者的关注. 由于行为视频的拍摄视角、背景和尺度等方面具有多样性, 使得不同行为的类间差异较小, 相同行为的类内差异较大, 所以基于视频的人体行为识别是非常具有挑战性的研究课题[1−3].

2012 年, Hinton 带领的团队在大规模图像竞赛 ILSVRC (ImageNet large scale visual recognition challenge) 中凭借卷积神经网络 (Convolutional neural network, CNN) 模型 AlexNet[4] 赢得了该年度比赛的冠军. 此后, 基于卷积神经网络的方法在图像分类、物体检测、图像分割和人脸识别等计算机视觉领域的研究中占据了重要的位置. 虽然卷积神经网络在处理静态图像任务中的表现令人印象深刻, 但是由于无法建模视频中的时序变化信息, 基于卷积神经网络的行为识别方法在相当长的一段时间内仍然一直无法超越基于人工设计特征的方法[5]. 2014 年, Simonyan 等[6] 提出了双流卷积神经网络模型, 将基于卷积神经网络的方法较好地拓展到视频分析领域. 该模型由两个独立的空间信息网络和时间信息网络构成. 空间信息网络的输入为视频的

收稿日期 2018-04-28 录用日期 2018-11-05
Manuscript received April 28, 2018; accepted November 5, 2018
国家自然科学基金 (61971086, 61806140, 61471082) 资助
Supported by National Natural Science Foundation of China (61971086, 61806140, 61471082)
本文责任编委 赖剑煌
Recommended by Associate Editor LAI Jian-Huang
1. 大连理工大学信息与通信工程学院 大连 116033 2. 天津大学智能与计算学部 天津 300350
1. School of Information and Communication Engineering, Dalian University of Technology, Dalian 116033 2. College of Intelligence and Computing, Tianjin University, Tianjin 300350

单帧彩色图像, 是视频中的环境、物体的空间位置信息的载体. 时间信息网络的输入是堆叠光流灰度图像, 代表时序变化信息, 用来建模行为的动态特征. 通过融合两路网络 softmax 输出的分数, 得到最后的识别结果. 双流卷积神经网络模型对于行为识别任务十分有效, 研究者们基于此模型提出了多种融合双流网络的方法. Feichtenhofer 等[7] 在最后一个卷积层融合视频序列中连续多帧图像的空间和时间特征, 然后对融合后的时空特征进行 3D 卷积和 3D 池化操作. Feichtenhofer 等[8−9] 进一步研究了使用残差网[10] 作为双流模型基本架构时的融合方法, 提出了在空间流和时间流之间加入短连接, 将时间流信息注入到空间流之中, 以增强双流之间的时空交互. 其中, ST-ResNet[8] 采用直接注入的方式, 而 ST-multiplier[9] 的时间流信息会先经过乘法门函数. 在增强了时空信息的交互的同时, 这两个工作中都将网络中 2D 卷积核拓展成了 3D 卷积核, 扩大了视频局部特征建模时序的范围. Wang 等[11] 引入了空间和时间二阶统计信息, 并在最后一个卷积层以金字塔的形式融合双流网络, 形成了更有效的视频局部时空特征. Wang 等[12] 将视频片段分成 N 段, 利用一阶双流网络分别提取每一段的特征, 最后对每一段的特征进行加权融合, 得到最终的视频表达.

上述工作主要研究基于 RGB 视频的行为识别. 除此之外, 学者们也研究了基于 RGB-D 视频的行为识别问题, 即采集的视频图像中包含深度 (Depth) 信息. Hu 等[13] 提出了一种异质特征融合方法, 通过融合动态骨架特征、动态颜色模式特征和动态深度特征, 在 4 个 RGB-D 行为数据库上取得领先性能. Shahroudy 等[14] 提出了一种基于深度自编码的共享特定特征分解网络, 将输入的多模态信号分解成不同的组成成分, 并提出使用混合范数作为多特征的正则项, 可以选择不同组合的特征, 该方法在 5 个 RGB-D 行为数据库取得较好结果. 与 Hu 等[13] 和 Shahroudy 等[14] 的工作不同, 本文主要研究基于 RGB 视频的行为识别方法.

目前基于双流卷积神经网络的工作中, 仅融合了视频空间和时间特征的一阶统计信息, 没有考虑更高阶的统计信息. 以上的融合方法虽然在训练时获得了视频局部空间和时间特征的一阶统计信息或二阶统计信息, 但是没有同时利用视频局部特征的一阶和二阶信息. 尽管在训练阶段都扩大了局部特征建模时序的范围, 但在测试阶段仅考虑融合多个视频局部特征的分类器分数, 没有考虑局部特征之间的统计信息. 为了解决双流卷积神经网络方法中存在的问题, 同时受到多种模态特征融合方法的启发, 本文提出了基于二阶聚合的视频多阶信息融合方法.

本文方法流程如图 1 所示, 主要分为两个阶段, 第 1 阶段为一阶和二阶双流网络的训练, 第 2 阶段是基于二阶聚合的多阶信息融合. 在第 1 个阶段中, 空间流和时间流都利用在 ImageNet 数据集[15] 上预训练的网络, 分别在目标数据集上进行微调, 微调后的双流模型可以提取视频局部空间和时间特征的一阶统计信息. 对于视频局部特征二阶统计信息的获取, 则受到了图像分类领域研究方法的启发. 在图像分类中, 特征分布的二阶信息有着较为广泛地应用, 对分类准确率的提升也起到重要作用. Lin 等[16−17] 提出了一种双线性池化卷积神经网络, 该网络将最后一层卷积层的输出特征进行外积计算, 从而得到特征分布的二阶信息, 该方法在精细粒度图像分类任务上取得了较高的准确率. Li 等提出了 MPN-COV 卷积神经网络[18] 及其快速算法 iSQRT-COV 卷积神经网络[19], 这两个网络通过对卷积层的输出进行协方差池化, 并对协方差矩阵进行幂正规化处

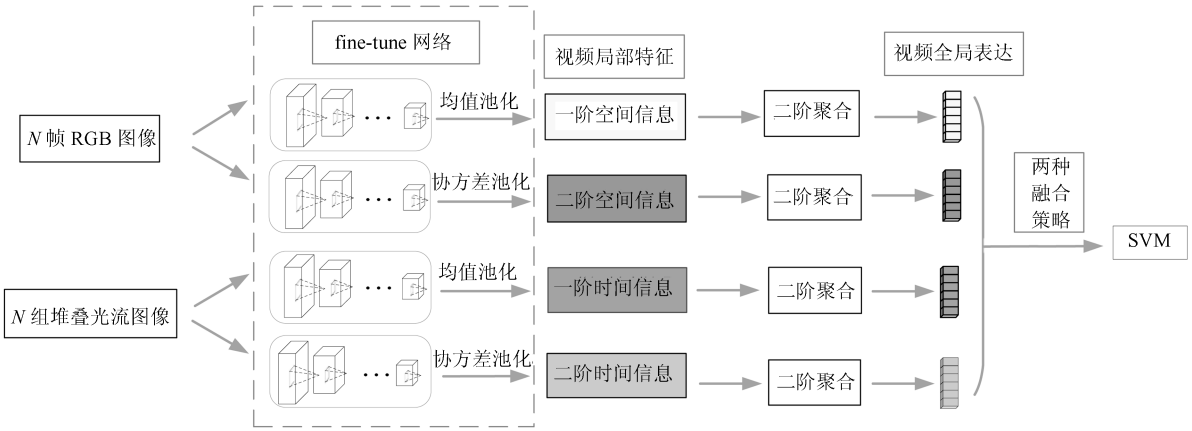


图 1 基于二阶聚合的多阶信息融合方法流程图

Fig. 1 The flow chart of multi-order information fusion based on second-order aggregation

理, 将正规化协方差矩阵进行取上三角矩阵并向量化, 作为图像的表达, 该表达包含了特征分布的二阶统计信息, 这一系列的工作在大规模图像分类任务以及精细粒度图像分类中取得了优异的性能. 考虑计算速度和收敛速度, 本方法基于 iSQRT-COV 卷积神经网络建立了二阶双流网络模型, 用来获取视频局部空间和时间特征的二阶统计信息.

在基于二阶聚合的多阶信息融合阶段, 对基于多阶信息的视频局部特征分别进行聚合. 为了获取视频局部特征之间的交互信息, 通过双线性池化[16−17] 这样的二阶聚合方式处理视频局部特征, 但是双线性池化后得到的表达维度较高, 将带来较大的计算和存储代价. 为了在降低维度的同时不损失多阶信息的表达能力, 本文使用压缩双线性池化方法[20] 对基于多阶信息的视频局部特征分别进行聚合, 形成高阶视频全局表达. 最后, 使用表达级和分类器分数级两种不同策略融合 4 种视频全局表达.

1　一阶和二阶双流卷积神经网络的训练过程

第 1.1 节阐述使用一阶双流卷积神经网络模型获取视频空间和时间一阶信息的过程. 第 1.2 节是建立二阶双流卷积神经网络模型的过程, 获得了视频局部特征的二阶信息. 本文选择 ResNet-50 作为一阶和二阶双流模型的基本架构. 对于双流网络的训练, 一般使用迁移学习的方法在 ImageNet 数据集预训练的网络模型对不同的目标数据集上进行微调, 从而获得更好的效果. 对于时间流网络, 为了使输入能够接受视频序列中连续多帧的水平和竖直光流信息, 把第一层卷积核的通道数由原来的 3 通道经过复制拓展成 $2L$ 通道, L 为在视频续中连续采样帧的个数, 在经典的双流卷积神经网络中 $L = 10$.

1.1　训练一阶双流卷积神经网络

一阶双流卷积神经网络模型由空间流网络和时间流网络两部分组成, 通常不会直接从参数的重新初始化开始直接独立训练双流网络, 这是由于实验中所使用的行为数据集的大小有限, 容易造成网络训练不收敛或者过拟合, 这样得到的网络效果很差. 一般使用在 ImageNet 数据集[15] 上预训练的网络模型对不同目标数据集进行微调.

经过微调后的一阶空间流和时间流网络, 其输入端的单帧 RGB 图像和单组堆叠光流图像尺寸分别为 $224 \times 224 \times 3$ 和 $224 \times 224 \times 20$, 经过残差单元后, 最后一层输出的卷积特征图尺寸为: $7 \times 7 \times 2\,048$, 其特征描述子数目为 49, 维度为 2 048. 设该输出特征为 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M\}$, 其中 $M = 49$. 经过全局均值池化层

$$\boldsymbol{z} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{x}_i \tag{1}$$

$\boldsymbol{z}$ 的维度为 2 048, 表示单帧 RGB 图像的表达, 即视频局部特征的一阶统计信息. 本文使用 Feichtenhofer 等[8] 单独训练好的空间流模型和时间流模型作为初始化模型提取基于一阶统计信息的视频局部特征.

1.2　建立二阶双流卷积神经网络模型

在一阶双流卷积模型中, 使用的预训练网络是一阶网络模型, 为建立二阶双流卷积神经网络模型, 最直接的办法是利用预训练的二阶卷积神经网络模型. 在 ImageNet 上训练二阶卷积神经网络的工作主要有以下两个: 分别嵌入了 MPN-COV[18] 和 iSQRT-COV[19] 结构层的二阶卷积神经网络, 这两个模块通常位于网络的最后一个卷积层, 其操作均是对卷积层的特征进行协方差池化, 再对协方差矩阵进行幂正规化处理, 将输出作为全连接层的输入. iSQRT-COV 是 MPN-COV 的快速近似算法.

本文以嵌入了 iSQRT-COV 结构层的网络为基础建立二阶双流网络模型. 使用在 ImageNet 上训练好的 iSQRT-COV-ResNet-50-2K 网络在行为识别数据集的 RGB 数据和光流数据上进行微调, 双流网络微调的过程为: 将 iSQRT-COV-ResNet-50-2K 网络中的最后一层分类层的 1 000 个节点替换成目标数据集的类别数, 其中 UCF101 数据集的类别数为 101, HMDB51 数据集的类别数为 51. 随机初始化该层参数, 并以很小的学习率继续训练网络, 网络收敛后即得到二阶空间流网络和二阶时间流网络. 二阶双流模型将作为特征提取器, 提取在视频序列中均匀采样的 N 帧图像空间和时间特征的二阶信息.

经过微调后的二阶空间流和时间流网络, 其输入端的单帧 RGB 图像及单组堆叠光流图像的尺寸分别为: $224 \times 224 \times 3$ 和 $224 \times 224 \times 20$, 经过残差单元, 最后一层输出的卷积特征进入 iSQRT-COV 结构层, 首先经过一个卷积核大小为 1×1, 通道数为 64 的卷积层, 使通道数由 2 048 减少到 64, 相当于特征维度由 2 048 降低到 64, 以降低计算复杂性. 这一卷积层使得正规化协方差形成的特征表达维度降低, 从而减少全连接层的参数, 同时又不损失性能. 将该特征记为 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M\}$, 其中 M 代表特征描述子的数目, 此时 $M = 49$. X 首先经过协方差池化

$$C = X\bar{I}X^{\mathrm{T}} \tag{2}$$

其中, $\bar{I} = \frac{1}{M}(I - \frac{1}{M}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}})$. I 是 $M \times M$ 的单位矩阵, $\boldsymbol{a} = [1, \cdots, 1]^{\mathrm{T}}$ 是 M 维单位向量. 然后, 经过

iSQRT 结构层, 此结构层用于近似求协方差矩阵的平方根. 协方差矩阵平方根的计算方法如下: 样本协方差矩阵是半正定对称矩阵, 其本征分解为

$$C = U\mathrm{diag}\{\lambda_i\}U^{\mathrm{T}},\ \ i = 1, \cdots, M \tag{3}$$

其中, U 是正交矩阵. $\mathrm{diag}(\lambda_i)$ 是 C 的特征值矩阵, 且为对角矩阵. 则矩阵 C 的平方根正规化结果为

$$Q = U\mathrm{diag}\{\lambda_i^{\frac{1}{2}}\}U^{\mathrm{T}},\ \ i = 1, \cdots, M \tag{4}$$

式 (4) 对协方差矩阵进行了平方根正规化. 由于矩阵进行本征分解的过程不能充分利用 GPU 的计算资源, 所以 iSQRT-COV 结构层使用迭代法近似求解协方差矩阵的平方根, 其输出 Q 经过上三角阵的向量化操作后记作 $\boldsymbol{z}$, 是单帧 RGB 图像的表达, 即基于二阶统计信息的视频局部特征, 维度为 2 080 (约为 2 K).

2 基于二阶聚合的视频多阶信息融合

为了验证多阶信息融合的有效性, 首先对一阶和二阶双流模型中不同网络流的组合进行融合, 共得到 8 种不同的组合, 分别计算各组合分类器分数的均值得到最终预测. 第 3.2 节在 UCF101 和 HMDB51 数据集上对这种多阶信息融合方式进行了评估, 实验结果表明一阶、二阶空间和时间网络流之间具有一定的互补性, 初步验证了多阶信息的有效性. 在此基础上, 进一步提出了对视频局部特征的一阶和二阶信息分别进行二阶聚合, 并在聚合后形成了高阶视频全局表达, 对于该表达的融合, 采用了表达级融合和分类器分数级融合两种策略.

2.1 多阶信息的二阶聚合

通过从视频序列 V 中均匀采样 N 帧图像, 使用第 1.1 节和第 1.2 节的双流一阶、二阶网络模型获取 N 帧图像空间和时间特征, 是视频局部特征, 获取了视频的多阶信息. 下面将以一种视频局部特征为例阐述二阶聚合的过程. 例如, 如果使用二阶空间流网络提取 N 帧图像的特征, 构成集合 $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N\}$, $\boldsymbol{z}_n \in \mathbf{R}^c$, $\boldsymbol{z}_n$ 是视频中单帧 RGB 图像的表达, 即视频局部空间特征, $c = 2\,080$, Z 为视频局部特征的集合.

N 个视频局部特征经过双线性池化聚合操作

$$B(Z) = \sum_{n\in\mathbf{N}} \boldsymbol{z}_n \boldsymbol{z}_n^{\mathrm{T}} \tag{5}$$

$B(Z)$ 是 N 个视频局部特征的外积, 捕捉了采样帧之间的交互信息, 此高阶视频全局表达可以建模整段视频的空间位置变化信息. $B(Z)$ 是一个 $c\times c$ 的对称矩阵, 取其上三角矩阵并进行向量化操作后维度为 $c(c+1)/2 = 2\,164\,240$, 该视频级表达维度过高, 将会给计算和存储造成较大负担, 以下将说明对其降维的方法.

为了对这些视频表达进行分类, 一般使用线性支持向量机或者逻辑回归等线性核分类器. 对于两类不同人体行为的表达为 $B(Z)$ 和 $B(P)$, 使用线性核比较两类表达

$$\langle B(Z), B(P)\rangle = \sum_{n\in\mathbf{N}}\sum_{u\in\mathbf{N}} \langle \boldsymbol{z}_n, \boldsymbol{p}_u\rangle^2 \tag{6}$$

可以看出, $B(Z)$ 和 $B(P)$ 是基于二项式核的映射函数, 这就相当于对分类器引入了非线性核函数, 对最终分类性能非常有帮助. 将此二项式核表示为 $k(\boldsymbol{z},\boldsymbol{z})$. 如果可以找到低维映射函数 $\Psi(\boldsymbol{z}) \in \mathbf{R}^d$, $d \gg c^2$, 满足 $\langle\Psi(\boldsymbol{z}), \Psi(\boldsymbol{p})\rangle \approx k(\boldsymbol{z},\boldsymbol{p})$, 则式 (6) 可以表示为

$$\langle B(Z), B(P))\rangle \approx \langle O(Z), O(P)\rangle \tag{7}$$

由此可以看出, 可以使用任意多项式核的低维近似, 将高维向量空间向低维向量空间映射, 得到压缩的双线性池化聚合后的视频全局表达 $O(Z) = \sum_{n\in\mathbf{N}} \Psi(\boldsymbol{z}_n)$, 进而解决高维双线性池化表达的计算和存储问题.

对单个视频局部特征 $\boldsymbol{z}$ 进行基于张量速写算法的压缩双线性池化[20] 操作

$$E : \boldsymbol{z} \rightarrow \boldsymbol{y} \tag{8}$$

基于张量速写算法的压缩双线性池化操作具体流程如图 2 所示. 主要经过以下三个步骤:

步骤 1. 随机产生两组参数并将其固定: $h_k \in \mathbf{N}^c$ 和 $s_k \in \{+1, -1\}^c$, $k = 1, 2$. c 是视频局部特征 $\boldsymbol{z}$ 的维度, d 为经过压缩双线性池化编码后表达的维度, $d \ll c^2$, 其中 $h_k(i)$ 服从 $\{1, 2, \cdots, d\}$ 的均匀分布, $s_k(i)$ 服从 $\{+1, -1\}$ 的均匀分布.

步骤 2. 定义张量速写映射函数

$$\Phi(\boldsymbol{z}, h, s) = \{(\mathrm{Q}_{\boldsymbol{z}})_1, \cdots, (\mathrm{Q}_{\boldsymbol{z}})_d\} \tag{9}$$

其中, $(\mathrm{Q}_{\boldsymbol{z}})_j = \sum_{t:h(t)=j} s(t)\boldsymbol{z}_t$. $t : h(t) = j$ 表示求和的取值范围是使等式 $h(t) = j$ 成立的所有 t 值, $j = 1, \cdots, d$, $\boldsymbol{z}_t$ 代表视频局部特征向量 $\boldsymbol{z}$ 中的第 t 个元素, t 的取值范围为 $(1, c)$, 当 $\boldsymbol{z}$ 是一阶特征时, $c = 2\,048$, 当 $\boldsymbol{z}$ 是二阶特征时, $c = 2\,080$.

步骤 3. 根据 Count sketch 算法[21], 计算 $\boldsymbol{z}$ 的两组张量速写的循环卷积求取 $\boldsymbol{z}$ 外积的张量速写, 即

$$y = \Psi_{\mathrm{TS}}(\boldsymbol{z}) \equiv \mathrm{FFT}^{-1}(\mathrm{FFT}(\phi(\boldsymbol{z}, h_1, s_1))\cdot$$

$$\mathrm{FFT}(\phi(\boldsymbol{z}, h_2, s_2))) \tag{10}$$

$\boldsymbol{y}$ 是视频局部特征 $\boldsymbol{z}$ 进行压缩双线性池化结果, $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N\}$ 集合中的每个视频局部特征都经过压缩的双线性池化操作, 得到视频局部压缩双线性池化特征集合 $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, 对 Y 进行全局均值池化, 得到多个视频局部特征的二阶聚合结果, 为视频的全局表达. 以上为以视频空间二阶信息为例说明基于多阶信息的视频局部特征的二阶聚合过程. 对于视频局部特征的空间一阶信息、时间一阶信息和时间二阶信息的二阶聚合, 与上述操作相同, 最终可以获得 4 种高阶视频全局表达.

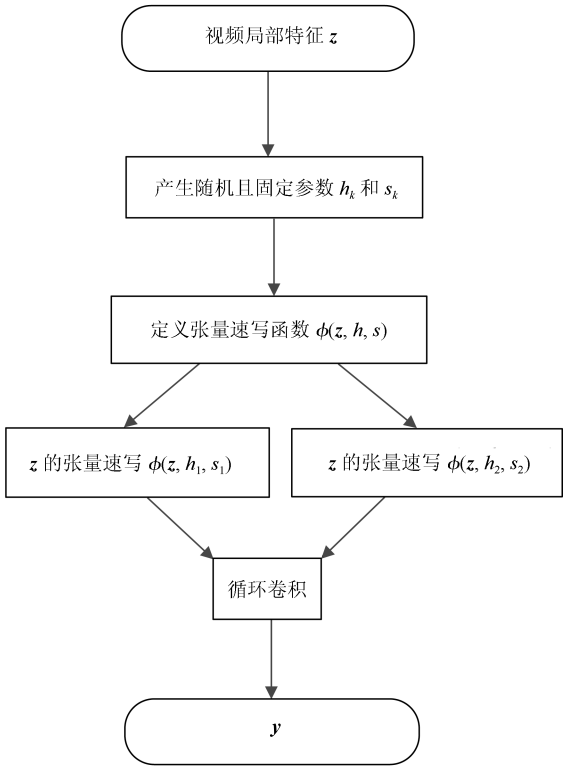


图 2　视频局部特征 $\boldsymbol{z}$ 进行压缩双线性池化操作流程图

Fig. 2　The flow chart of compact bilinear pooling of one local video feature $\boldsymbol{z}$

2.2　视频全局表达的融合

行为识别的方法大多数都是多种特征进行融合的. 在经典的人工设计特征中, 改进的密集点轨迹 (Improved dense trajectory, IDT) 特征[5] 是使用最为广泛的特征, 为了描述轨迹周围的表观结构信息和运动信息, 一般会基于轨迹提取 HOG (Histogram of oriented gradient)、HOF (Histogram of flow) 和 MBH (Motion boundary histogram) 特征, 最终结果是以上几种特征融合的结果. 视频数据中存在多种属性, 使融合视频的特征或者表达成为必然. 本文通过两阶段建模的方式获得了 4 种高阶视频全局表达, 以下介绍融合这 4 种表达的策略并分析其性质.

假设 4 种高阶视频全局表达分别为视频一阶空间信息表达 $\boldsymbol{Y}_{\mathrm{s1}}$、一阶时间信息表达 $\boldsymbol{Y}_{\mathrm{t1}}$、二阶空间信息表达 $\boldsymbol{Y}_{\mathrm{s2}}$ 和二阶时间信息表达 $\boldsymbol{Y}_{\mathrm{t2}}$. 这 4 种表达分别获取了行为视频数据中存在的多种属性. $\boldsymbol{Y}_{\mathrm{s1}}, \boldsymbol{Y}_{\mathrm{s2}}, \boldsymbol{Y}_{\mathrm{t1}}$ 和 $\boldsymbol{Y}_{\mathrm{t1}}$ 分别是视频局部空间特征的一阶信息之间、局部空间特征的二阶信息之间、局部时间特征的一阶信息之间和局部时间特征的二阶信息之间的高阶统计信息.

与上文中提到的一阶、二阶空间和时间网络流的融合方式的组合数量相同, 有 8 种不同的组合形式, 这几种视频全局表达之间也会产生 8 种不同的组合. 在第 3.3 节中将首先对这 8 种不同组合进行评估, 确定互补性最强的组合. 在基于手工特征的方法中, 视觉词袋模型下的特征融合发生在 3 个不同的处理等级: 特征级融合、表达级融合和分类器分数级的融合. 本文方法中多阶信息经二阶聚合后形成了视频级表达, 可采用表达级融合和分类器分数级融合这两种策略, 下面以 4 种表达的组合为例说明两种融合策略及其不同的性质.

对于表达级融合策略, 融合过程发生在得到视频全局表达之后, 先将 4 个表达串联成更长的视频级表达. 图 3 为该种策略的融合过程示意图. 这 4 种表达先经过内部归一化, 即 $\boldsymbol{Y}_{\mathrm{s1}}, \boldsymbol{Y}_{\mathrm{t1}}, \boldsymbol{Y}_{\mathrm{s2}}, \boldsymbol{Y}_{\mathrm{t2}}$ 分别经过指数归一化和 $L2$ 范数归一化处理

$$\boldsymbol{Y}_{\mathrm{final}} = \mathrm{cat}(\boldsymbol{Y}_{\mathrm{s1}}, \boldsymbol{Y}_{\mathrm{t1}}, \boldsymbol{Y}_{\mathrm{s2}}, \boldsymbol{Y}_{\mathrm{t2}}) \tag{11}$$

其中, $\mathrm{cat}(\cdot)$ 表示将 4 种视频全局表达串联. 然后, 对串联后的视频全局表达 $\boldsymbol{Y}_{\mathrm{final}}$ 进行指数归一化和 $L2$ 范数归一化. 最后, 送入支持向量机 (Support vector machine, SVM) 进行识别. 如果每一种视频

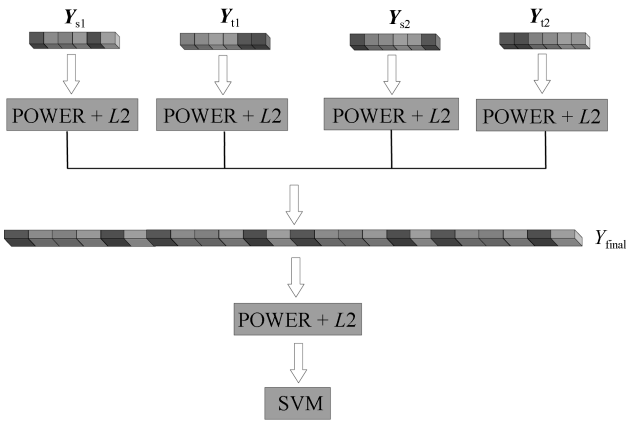


图 3　表达级融合过程示意图

Fig. 3　Fusion at the representation level

级表达的维度为 d, 级联后的表达维度为 $4d$. 此时 $\boldsymbol{Y}_{\text{final}}$ 既包含了视频局部特征多阶信息, 又包含了视频局部特征多阶信息之间的交互信息, 形成了一种更加有效的视频全局时空表达.

对于分类器分数级的融合策略, 融合过程发生在每种视频全局表达独立地送入 SVM 分类器之后, 将所有分类器的得分进行融合, 得分融合策略使用算术平均. 如图 4 所示, 这种融合策略分别对 $\boldsymbol{Y}_{\text{s1}}$, $\boldsymbol{Y}_{\text{t1}}$, $\boldsymbol{Y}_{\text{s2}}$, $\boldsymbol{Y}_{\text{t2}}$ 进行指数归一化和 $L2$ 范数归一化操作, 并分别进行 SVM 分类, 将分类器得分相加, 得到最终的预测结果.

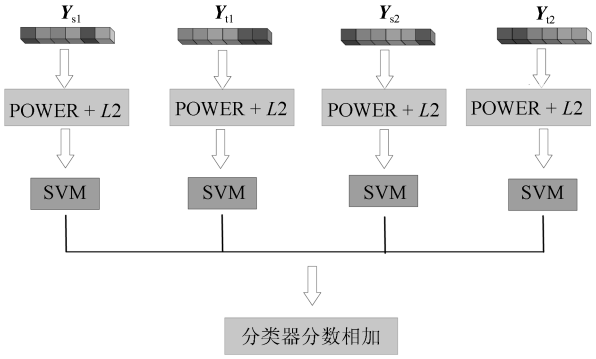


图 4 分类器分数级融合过程示意图

Fig. 4 Fusion at the classifier score level

这两种不同处理级别的融合策略各有优缺点, 具体选择哪种融合策略要研究这 4 种表达在不同处理级别的相关性. 如果 4 种视频全局表达之间相关性很大, 那么选择表达级融合策略比较合理. 否则, 如果表达之间不相关, 将其级联成更长的表达使得分类过程中产生信息丢失的情况. 这些表达相关性比较弱时, 则应该选择分类器分数级的融合. 融合能够提高性能的主要原因是这些表达之间具有一定的互补性, 这种互补性存在于不同的描述层次上.

3 实验

本节使用两个数据集对二阶双流网络模型以及基于二阶聚合的视频多阶信息融合方法进行评估实验, 数据集分别为 UCF101[22] 和 HMDB51[23]. UCF101 数据集包含 101 种人体行为, 共 13 320 个视频序列, 这个数据集里面的大多数行为是关于体育运动的. HMDB51 数据集包含 51 种人体行为, 总共 6 766 个视频序列, 每一类行为至少有 100 个视频样本, 该数据集视频主要来源于网络视频和电影片段, 行为的类内差距非常大, 是目前最难的数据集之一. 这两个数据集使用 3 种方式划分训练集和测试集, 在 3 个划分上的平均准确率作为最终的分类结果.

3.1 实验参数设置

第 1 阶段训练一阶和二阶双流网络的参数设置: 对一阶双流模型和二阶双流模型在 HMDB51 和 UCF101 数据集上进行微调的初始学习率为 0.001, 当验证错误率达到饱和时, 学习率除以 10. 提取视频局部特征时, 一阶视频局部特征的维度 c 为 2 048, 二阶视频局部特征的维度是正规化协方差矩阵 Q 取上三角矩阵并进行向量化操作后的维度 c 为 2 080.

第 2 阶段基于二阶聚合的多阶信息融合的参数设置: 所有指数归一化的操作中指数的取值为 0.45, 视频的表达采用一对多的线性 SVM 进行分类, 其容错参数 $C = 100$.

3.2 一阶、二阶双流模型性能比较

表 1 是二阶网络空间流网络和时间网络在 UCF101 和 HMDB51 上分别与一阶空间流网络和时间流网络的性能比较. 二阶双流模型在测试时采用了与一阶双流模型[6, 9] 相同的标准方式, 在对一个包含多帧的视频进行分类时, 从视频片段中随机选择 25 帧, 每一帧图像中裁剪出 10 个 crop, 对于空间流网络每个 crop 的大小为 $224 \times 224 \times 3$, 而对时间流网络该 crop 的大小为 $224 \times 224 \times 20$, 最后对每个 crop 独立地进行预测, 再计算平均值作为该视频片段的预测值.

表 1 一阶、二阶空间和时间流网络在 UCF101 和 HMDB51 上准确率的比较

Table 1 Comparisons of first-order spatial and temporal network with second-order spatial and temporal network on UCF101 and HMDB51

模型	网络类型	UCF101 (%)	HMDB51 (%)
ResNet-50[9]	一阶空间	82.30	48.90
iSQRT-COV-ResNet-50-2K	二阶空间	**85.29**	**49.65**
ResNet-50[9]	一阶时间	87.00	55.80
iSQRT-COV-ResNet-50-2K	二阶时间	**88.07**	**57.64**

由表 1 可知, 无论是空间流网络还是时间流网络, 二阶网络的性能均超过一阶网络. 在 UCF101 上, 二阶空间流网络比一阶空间流网络性能提升 2.99 %, 提升较为显著, 在 HMDB51 上该提升为 0.75 %.

在 UCF101 和 HMDB51 上, 对于一阶时间流和二阶时间流网络的比较, 二阶网络分别比一阶网络提升 1.07 % 和 1.84 %. 初步证明了引入空间和时间特征的二阶统计信息的必要性.

3.3　多阶信息融合有效性的评估

根据第 2.1 节所述, 表 2 列出了一阶和二阶双流模型中网络流进行组合时的 8 种不同情况. 表 2 中的第 1 行是一阶双流模型融合在 UCF101 和 HMDB51 上的识别准确率. 融合方式与经典双流网络相同, 即计算网络 softmax 输出分数的均值最为最终的预测. 如表 2 所示, 在 HMDB51 上, 一阶双流网络融合的准确率为 61.20 %. 除了一阶空间流和二阶时间流及一阶时间流和二阶空间流这两种组合以外, 其他 5 种组合均超过了一阶双流网络融合的识别准确率. 识别率最高的组合为一阶时间流、二阶空间流和二阶时间流融合, 准确率比一阶双流网络融合提升 4.94 %. 在 UCF101 上, 一阶双流网络融合的准确率为 91.70 %, 一阶空间流、二阶空间流和二阶时间流及一阶空间和二阶时间流这两种组合低于一阶双流网络融合的识别结果, 其他 5 种组合均高于一阶双流网络融合. 识别率最高的组合和在 HMDB51 上具有相同的规律, 比一阶双流网络融合提高了 1.26 %. 上述实验初步验证了在行为识别任务中融合多阶信息的必要性.

3.4　参数评估

本小节实验首先评估了基于二阶聚合的视频多阶信息不同组合情况下融合的准确率. 其次, 在 HMDB51 数据集上评估了对视频多阶信息融合有重要影响的两个参数, 即视频中均匀采样帧的数量 N. 最后, 当视频表达维度 d 为 8 K 到 64 K 范围内变化的情况下, 在两个数据库上评估两种不同的视频多阶信息融合策略的性能.

1) 在 UCF101 和 HMDB51 上评估二阶聚合后的视频不同多阶信息进行不同组合时的准确率, 实验设置为: $d = 8\,\mathrm{K}$, $N = 25$, 融合策略为分类器分数级的融合. 从表 3 中可以看出视频的多阶信息融合在 UCF101 和 HMDB51 上表现出相同的规律, 当一阶空间和时间信息及二阶空间和时间信息融合时, 性能达到最优, 在两个数据库上比一阶双流网络信息融合提升了 3.50 % 和 3.91 %. 以上表明在基于二阶聚合的视频多阶信息融合方法中, 4 种多阶信息之间具有较强的互补性. 在以下实验中都基于 4 种多阶信息的融合.

由表 2 和表 3 可以看出, 多阶信息的二阶聚合方法在 UCF101 数据库上的效果不够显著. 其主要原因是: 在 UCF101 上识别性能已经接近饱和, 其 state-of-the-art 性能已经超过了 93 %. 因此, 在 UCF101 上特征分布相对简单, 用一阶统计信息就可以较好地对行为进行分类; HMDB51 的行为类别更为复杂, 类别之间的差异更大. 在 HMDB51 上, 需要用表达能力更强的二阶统计特性 (建模特征之间的相关性) 才能准确地对行为类别进行分类.

2) 评估从视频中均匀采样帧的数量 N. 实验设置为: 二阶聚合后的 4 种视频表达维度 $d = 8\,\mathrm{K}$, 融合策略采用分类器分数级的融合. 实验结果如图 5 所示, 当 N 从 3 逐渐增加至 25 的过程中, 分类准确率随着帧数的增加而不断提高, 当 $N = 25$ 时, 分类准确率达到最高. 当 N 继续增加至 35 帧时, 性能有所下降. 分析造成以上现象的原因, 是由于在

表 2　UCF101 和 HMDB51 上多阶信息融合有效性评估

Table 2　Evaluation of the effectiveness of multi-order information fusion on UCF101 and HMDB51

一阶空间流	一阶时间流	二阶空间流	二阶时间流	UCF101 (%)	HMDB51 (%)
✓	✓			91.70	61.20
	✓	✓		92.90	65.17
✓			✓	91.34	61.63
		✓	✓	92.67	63.50
✓	✓		✓	92.50	65.18
	✓	✓	✓	**92.96**	**66.14**
✓	✓	✓		91.78	60.60
✓		✓	✓	91.12	58.71
✓	✓	✓	✓	92.75	64.74

表 3 UCF101 和 HMDB51 上基于二阶聚合的视频不同多阶信息融合评估
Table 3 Evaluation of fusing different multi-order information of the video based on second-order aggregation on UCF101 and HMDB51

一阶空间信息	一阶时间信息	二阶空间信息	二阶时间信息	UCF101 (%)	HMDB51 (%)
✓	✓			89.28	64.24
	✓	✓		87.57	59.56
✓			✓	92.58	65.93
		✓	✓	92.07	64.10
✓	✓		✓	92.68	68.02
	✓	✓	✓	92.60	67.45
✓	✓	✓		88.64	61.44
✓		✓	✓	92.55	64.88
✓	✓	✓	✓	**92.98**	**68.15**

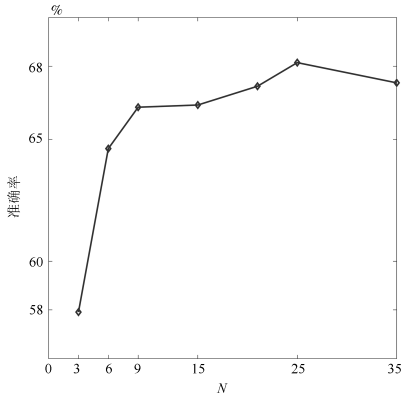
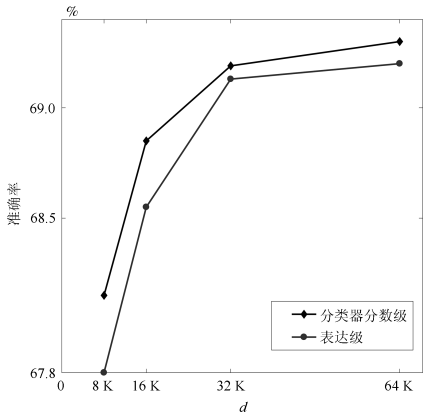


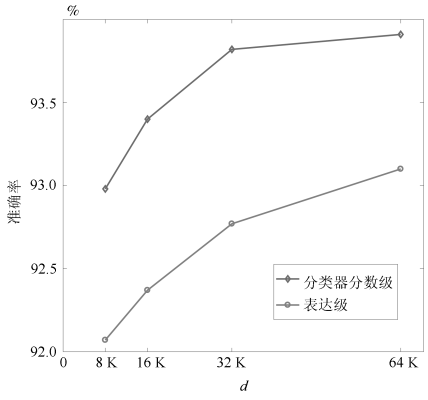
图 5 HMDB51 上对视频序列中均匀采样帧数目 N 的评估
Fig. 5 Evaluation of the number N of the frames uniformly sampled from the video on HMDB51

HMDB51 数据集中, 每段视频序列的帧数大多在 100 到 200 帧之间, 对于空间流取 25 帧时, 在时间流网络则是获取了 250 帧的运动变化信息, 能够建模整段视频的运动变化情况, 而继续增加采样帧数则带来了信息冗余以及噪声. 本文以下实验中, 视频采样帧数 N 均为 25.

3) 评估两种不同的融合策略, 分别为表达级融合和分类器分数级融合. 实验参数设置: $N = 25$, 在不同的视频表达维度 d 的情况下进行评估. 实验结果如图 6 (a) 所示, 在 HMDB51 数据库上, 表达级融合和分类器分数级融合的性能基本一致, 这说明 4 种表达之间具有一定的相关性, 同时具有互补性, 使得两种策略的差别较小. 在视频表达维度为 8 K 时, 分类器分数级的融合比表达级的融合准确率提高 0.35 %, 而随着表达维度的增高, 两种策略的性能差异逐渐减小. 分类器分数级的融合方式性能相对较好. 如图 6 (b) 所示, 在 UCF101 数据库上, 在 8 K 到 64 K 变化过程中, 表达层融合的性能比分类器分



(a) HMDB51 数据集上的评估
(a) Evaluation on HMDB51 dataset



(b) UCF101 数据集上的评估
(b) Evaluation on UCF101 dataset

图 6 HMDB51 和 UCF101 数据集在不同视频全局表达维度 d 下对表达级融合策略和分类器分数级融合策略的评估
Fig. 6 Evaluation of representation level fusion strategy and class score level fusion strategy under the different dimension of the video representation on HMDB51 and UCF101 dataset

数级融合的性能低 1% 左右. 综上所述, 在 HMDB51 数据库和 UCF101 数据库上, 分类器分数级融合的策略优势较为明显. 不论在 HMDB51 还是 UCF101 数据库上, 随着视频表达维度的增加, 性能逐渐提升, 综合计算代价和性能两方面因素考虑, 以下实验中 $d = 32$ K.

3.5 融合算法测试时间比较

实验主机配置情况: CPU 为 Intel Core i7-4770K, 3.50 GHz, 32 GB 内存, GPU 为 GTX1070. 本文算法是在 Windows10 系统, MATLAB 2017b 环境下实现的, 使用了 MatConvNet[24] 工具包. 融合算法时间代价的测试是在 HMDB51 数据集的第 1 个划分上进行的, 处理每段视频的时间为测试集所有视频测试时间的均值. 多阶信息聚合阶段的参数设置是: $N = 25$, $d = 32$ K.

一阶双流网络的融合方法是预测分数相加, 是离线进行的, 融合时间则是一阶空间流网络和一阶时间流网络测试时间的加和, 一阶时间流和一阶空间流网络均采用标准测试方法 (10-crop)[6]. 如表 4 所示, 一阶双流网络融合算法测试时间为每段视频 9.670 s. 二阶双流网络融合时, 空间流网络和时间流网络同样采用标准测试方法, 二阶双流网络融合测试时间为每段视频 10.459 s. 一阶和二阶双流网络融合测试时间为两个一阶网络和两个二阶网络测试时间的总和, 即每段视频 20.129 s. 本文提出的多阶信息二阶聚合融合方法测试方式是 "1-crop", 从视频片段中随机选择 25 帧, 每一帧图像中随机裁剪出 1 个 crop, 对于空间流网络每个 crop 的大小为 $224 \times 224 \times 3$, 而对于时间流网络该 crop 的大小为 $224 \times 224 \times 20$, 这些裁剪后的图像作为一阶和二阶双流网络的输入, 提取视频局部特征, 这些局部特征进行二阶聚合得到视频全局表达, 最终的预测是 4 种视频全局表达融合的结果. 在这种测试方式下, 本文方法处理每段视频的时间为 6.412 s, 提出的多阶信息聚合方法在融合时处理每段视频时没有带来额外的时间代价, 但本文方法使用 SVM 训练和测试时, 是离线进行的, 这一部分产生额外的时间代价. 所以相比于一阶双流网络融合的方法, 本文提出的多阶信息聚合方法在融合时处理每段视频时没有带来额外的时间代价, 但本文方法使用 SVM 进行训练和测试时, 是离线进行的, 这一部分产生额外的时间代价. 以 HMDB51 第 1 个划分为例, 训练视频 3 570 段, 测试视频 1 530 段, 使用分类器分数级策略进行融合时, 采用分类器分数级策略, 训练 SVM 分类器和测试的时间共计 300 s 左右.

表 4　不同融合方法测试时间比较
Table 4　Test speed comparison of different fusion methods

方法	测试方式	时间 (s/视频)
一阶双流网络融合 (基线)[9]	10-crop	9.670
二阶双流网络融合	10-crop	10.459
一阶 + 二阶双流网络融合	10-crop	20.129
多阶信息二阶聚合	1-crop	6.412

3.6 本文方法与其他行为识别方法比较

为了验证本文方法的优势, 将本文方法与其他基于双流卷积神经网络架构的行为识别方法进行了对比, 各方法的识别结果列入表 5 中. 表 5 中本文

表 5　基于双流卷积神经网络架构的行为识别方法比较
Table 5　Comparison of different human action recognition arthogram based on two-stream convolutional network

方法	网络架构	UCF101 (%)	HMDB51 (%)
Two-stream[6]	VGG-M	88.0	59.4
Two-stream 3D 卷积 + 3D 池化[7]	VGG-16	92.5	66.4
Two-stream[9]	ResNet-50	91.7	61.2
ST-ResNet*[8]	ResNet-50	93.4	66.4
ST-multiplier network[9]	ResNet-50 (空间), ResNet-152 (时间)	94.2	68.9
Two-Stream fusion + IDT[7]	VGG-16	93.5	69.2
ST-ResNet + IDT[8]	ResNet-50	94.6	70.3
ST-multiplier + IDT[9]	ResNet-50 (空间), ResNet-152 (时间)	94.9	72.2
本文方法	ResNet-50	**93.8**	**69.2**
本文方法 + 联合训练[8]	ResNet-50	**94.1**	**70.7**
本文方法 + IDT	ResNet-50	**94.6**	**74.4**

方法的参数设置是: $N = 25$, $d = 32\,\mathrm{K}$, 4 种多阶信息进行分类器分数级的融合. 在许多基于双流卷积神经网络模型的方法中, 都会通过与 IDT 轨迹特征[5] 进行融合来提升性能, 本文方法也进一步融合了 IDT 轨迹特征, 探究其与 IDT 轨迹特征的互补性. 本文使用 Peng 等[25] 公开的代码, 在视频中提取 IDT 特征 (即 HOG, HOF, MBH), 用费舍尔向量 (Fisher vector, FV) 方法对三种 IDT 特征进行编码并分别训练 SVM 分类器. 对 IDT 特征进行融合时, 分别计算三种 IDT 特征的 FV 编码对应的 SVM 分数并取均值, 然后与本文中的 4 种视频高阶全局表达 SVM 分数相加作为最后的预测分数.

由表 5 可知, 本文方法在 UCF101 和 HMDB51 上准确率分别达到了 93.8 % 和 69.2 %, 比经典的 two-stream ResNet-50[9] 方法提升 2.1 % 和 8.0 %. ST-multiplier[9] 方法在 UCF101 上的准确率为 94.2 %, 略高于本文方法, 但该方法所使用的时间流网络是网络层数更深, 性能更强的 ResNet-152 网络. 本文仅采用 ResNet-50 作为基本架构就可以与其达到几乎相同的准确率, 且在 HMDB51 数据集上的准确率高于 ST-mutiplier. 在 UCF101 数据集上, ST-Pyramid 的识别准确率为 93.8 %, 与本文方法一致, 而 ST-pyramid[11] 在网络架构中在特征层面上进行了时空金字塔分层聚合, 网络训练复杂度较高. 在 HMDB51 数据集上, 本文方法比 ST-pyramid 高 2.7 %. 本文方法在难度较大的 HMDB51 数据集上的优势较为明显, 在准确率趋于饱和状态的 UCF101 数据集上也获得了与当前最优算法相同的性能. 使用空间流和时间流联合训练的一阶双流网络架构[8] 作为一阶双流网络的初始化模型, 本文方法的性能进一步提升, 在 HMDB51 数据集上比经典的 two-stream 融合算法提升 9.5 %, 在 UCF101 上该提升为 2.4 %. 本文方法与 IDT 轨迹特征互补性良好, 融合 IDT 特征后识别准确率有所提高, 在 HMDB51 数据集上优势较为明显, 比性能最好的 ST-mutiplier 提高 2.2 %.

4 结论

本文针对基于双流卷积神经网络存在的两点不足提出了基于二阶聚合的多阶信息融合方法. 本文的主要贡献在于: 建立了二阶双流网络模型, 获取了空间和时间特征的二阶统计信息, 与经典双流模型获取的一阶统计信息形成了多阶信息. 基于多阶信息的视频局部特征经过二阶聚合后形成了高阶视频全局表达. 实验表明, 二阶双流模型具有更好的性能, 一阶双流模型和二阶双流模型获取多阶信息融合十分有效, 形成的 4 种视频高阶全局表达全部参与融合时互补性最强. 融合后的表达在难度较大的 HMDB51 数据集上优势十分明显, 在 UCF101 上也达到了与当前最好算法相同的性能, 融合 IDT 特征能进一步提高识别准确率.

References

1 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
(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)

2 Su Ben-Yue, Jiang Jing, Tang Qing-Feng, Sheng Min. Human dynamic action recognition based on functional data analysis. *Acta Automatica Sinica*, 2017, **43**(6): 866−876
(苏本跃, 蒋京, 汤庆丰, 盛敏. 基于函数型数据分析方法的人体动态行为识别. 自动化学报, 2017, **43**(6): 866−876)

3 Zhou Feng-Yu, Yin Jian-Qin, Yang Yang, Zhang Hai-Ting, Yuan Xian-Feng. Online recognition of human actions based on temporal deep belief neural network. *Acta Automatica Sinica*, 2016, **43**(6): 1030−1039
(周风余, 尹建芹, 杨阳, 张海婷, 袁宪锋. 基于时序深度置信网络的在线人体动作识别. 自动化学报, 2016, **42**(7): 1030−1039)

4 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: NIPS Foundation, Inc., 2012. 1097−1105

5 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 14th International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 3551−3558

6 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, Inc., 2014. 568−576

7 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 1933−1941

8 Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition. In: Proceedings of the 29th International Conference on Neural Information Processing Systems. Barcelona, ES, Spain: NIPS Foundation, Inc., 2016. 3468−3476

9 Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 7445−7454

10 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770−778

11 Wang Y B, Long M S, Wang J M, Yu S P. Spatiotemporal pyramid network for video action recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 2097−2106

12 Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X D, et al. Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 20−36

13 Hu J, Zheng W, Lai J, Zhang J G. Jointly learning heterogeneous features for RGB-D activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017, **39**(11): 2186−2200

14 Shahroudy A, Ng T, Gong Y H, Wang G. Deep multimodal feature analysis for action recognition in RGB + D videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2018, **40**(5): 1045−1058

15 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2014, **115**(3): 211−252

16 Lin T Y, Roychowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the 15th International Conference on Computer Vision. Santiago, USA: IEEE, 2015. 1449−1457

17 Lin T Y, Roychowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(6): 1309−1322

18 Li P H, Xie J T, Wang Q L, Zuo W M. Is second-order information helpful for large-scale visual recognition? In: Proceedings of the 16th International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2089−2097

19 Li P H, Xie J T, Wang Q L, Gao Z L. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018. 947−955

20 Gao Y, Beijbom O, Zhang N, Darrell T. Compact bilinear pooling. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 317−326

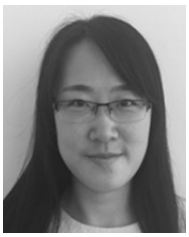
21 Charikar M, Chen K, Farach-Colton M. Finding frequent items in data streams. In: Proceedings of the 2002 International Colloquium on Automata, Languages, and Programming. Malaga, ES, Spain: Springer, 2002. 693−703

22 Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012. 1−7

23 Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: A large video database for human motion recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, ES, Spain: IEEE, 2011. 2556−2563

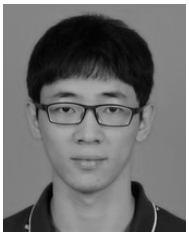
24 MatConvNet: CNNs for MATLAB: Source Code [Onlilne], available: http://www.vlfeat.org/matconvnet, November 7, 2018

25 Peng X J, Wang L M, Wang X X, Qiao Y. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 2016, **150**: 109−125



张冰冰 大连理工大学信息与通信工程学院博士研究生. 2016 年获长春工业大学硕士学位. 主要研究方向为人体行为识别, 图像分类, 深度学习.
E-mail: icyzhang@mail.dlut.edu.cn
(**ZHANG Bing-Bing** Ph. D. candidate at the School of Information and Communication Engineering, Dalian University of Technology. She received her master degree from Changchun University of Technology in 2016. Her research interest covers human action recognition, image classification, and deep learning.)

葛疏雨 大连理工大学信息与通信工程学院硕士研究生. 2016 年获大连理工大学通信工程专业学士学位. 主要研究方向为图像分类, 人体行为识别, 深度学习.
E-mail: gsy@mail.dlut.edu.cn
(**GE Shu-Yu** Master student at the School of Information and Communication Engineering, Dalian University of Technology. He received his bachelor degree from Dalian University of Technology in 2016. His research interest covers image classification, human action recognition, and deep learning.)



王旗龙 博士, 天津大学智能与计算学部副教授. 主要研究方向为图像建模, 视觉数据分类, 深度学习.
E-mail: qlwang@mail.dlut.edu.cn
(**WANG Qi-Long** Ph. D., associate professor at the College of Intelligence and Computing, Tianjin University. His research interest covers image modeling, visual classification, and deep learning.)

李培华 博士, 大连理工大学信息与通信工程学院教授. 主要研究方向为基于信息几何的图像分类与检索. 本文通信作者. E-mail: peihuali@dlut.edu.cn
(**LI Pei-Hua** Ph. D., professor at the School of Information and Communication Engineering, Dalian University of Technology. His research interest covers image classification and search using theoretical and computational methods of information geometry. Corresponding author of this paper.)