# 基于局部约束仿射子空间编码的时空特征聚合卷积网络模型

张冰冰 李培华 孙秋乐
(大连理工大学信息与通信学院 辽宁 大连 116033)

**摘 要** 双流卷积网络能够在视频中建模表观和运动信息，是行为识别中一种重要的基础网络模型. 然而，这种模型只能学习单帧的空间信息和少数几帧的时间信息，无法有效地建模整段视频中的长时信息. 为此，本文提出一种基于局部约束仿射子空间编码的时空特征聚合卷积网络. 该网络的核心是局部约束仿射子空间编码层，能够嵌入到双流卷积网络中用于聚合覆盖整段视频的空间和时间特征，从而获得视频的全局时空表达. 局部约束仿射子空间编码层由权重系数计算和仿射子空间编码组成，其中的参数可与卷积网络中的其他参数进行联合优化从而进行端到端的学习. 同时，本文研究了在代价函数中施加软正交约束、无穷范数约束和谱范数约束三种方法，以保证仿射子空间基的正交性. 在常用的 UCF101、HMDB51 和 Something-V1 数据集上，本文的方法比经典的双流卷积网络识别准确率分别提升 1.7%、8.7%和 4.3%，同时达到或优于当前最先进的方法.

**关键词** 行为识别；双流卷积网络；局部仿射子空间编码；时空特征聚合；正交约束
**中图法分类号** TP391 **DOI 号** 10.11897/SP.J.1016.2020.01589

# Spatial and Temporal Features Aggregation Convolutional Network Model Based on Locality-Constrained Affine Subspace Coding

ZHANG Bing-Bing LI Pei-Hua SUN Qiu-Le
(*School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning* 116033)

**Abstract** Video-based human action recognition is an import task in the field of computer vision. It is widely used in video surveillance, virtual reality and human-computer interaction. This is also a very challenging task because of the large amount of video data and high requirements for computing hardware systems. An effective video representation needs to consider spatial and temporal cues simultaneously. In recent years, researchers are working to develop general network architecture for video classification. 3D spatial-temporal convolutions that potentially learn complicated spatial-temporal dependencies but the large number of parameters in 3D CNNs make it hard to train in practice. Two-stream architectures that decompose the video into motion and appearance streams, and train separate CNNs for each stream, fusing the outputs in the end. Among these successful network architectures, two-steam convolutional network has a great influence in the academic research field. Two-stream convolutional network can model the appearance and motion information in videos, and becomes an important basic network model in action recognition. Two-stream architectures essentially learn a classifier that operates on individual frames or short clip of few frames possibly enforcing consensus of classification scores over different segments of the video. At test time, 25 uniformly sampled frames are classified independently and the classifications scores

收稿日期：2019-08-03；在线发布日期：2020-02-07. 本课题得到国家自然科学基金（No. 61471082）资助. **张冰冰**，博士研究生，主要研究领域为视频中的人体行为识别、图像分类和深度学习. E-mail: icyzhang@mail.dlut.edu.cn. **李培华**（通信作者），博士，教授，中国计算机学会(CCF)会员，主要研究领域为图像/视频识别、目标检测和语义分割. E-mail: peihuali@dlut.edu.cn. **孙秋乐**，博士研究生，主要研究领域为图像识别和语义分割.

are averaged to get the final prediction. However, such architectures mainly focus on learning of spatial information of a single frame and the temporal information of a few frames, thus failing to effectively model the long-term information in the whole video. To overcome the problem, we propose the spatial and temporal features aggregation convolutional network model based on locality-constrained affine subspace coding. LASC coding method has achieved the excellent performance of in image classification and image retrieval tasks. LASC leverages the semantic probabilities of local patches to learn the aggregation weights and construct the semantic affine subspace dictionary, which produces more semantic and discriminative global image representations. Inspired by the classical LASC coding method, we design a LASC-based structure layer to insert into the last layer convolutional layer of spatial stream and temporal stream to acquire more robust high-dimension representation and the two fully-connected layer in two-stream architecture is totally replaced. The core of the network is a locality-constrained affine subspace coding layer, which can be embedded in the two-stream convolution network for aggregating spatial and temporal features of the whole video to obtain the global temporal and spatial video representation. This layer consists of two sub-layers of computing weight coefficients and affine subspace coding, in which parameters of the layer can be optimized jointly with other parameters in convolution network for end-to-end learning. Besides, three regulations, i.e., soft orthogonality regulation, infinite-norm regulation and spectral-norm regulation in cost functions, are further studied to ensure the orthogonality of affine subspace bases during the training process. The proposed method is measured on the commonly used UCF10, HMDB51 and Something-V1 datasets, and the accuracy of our method is 1.7%, 8.7% and 4.3% higher than the classical two-stream convolution network, respectively. At the same time, it achieves superior or competitive performance in comparison to state-of-the-art methods.

**Keywords** action recognition; two-stream convolutional networks; locality-constrained subspace coding; spatial and temporal features aggregation; orthogonality regulation

## 1 引 言

基于视频的人体行为识别是计算机视觉领域的基本问题，其应用范围十分广泛，包括自动驾驶中的视频导航系统、视频剪辑、互联网视频检索和人机交互等. 人体行为识别的方法可分为两种：一种是基于传统手工特征的方法；另一种是基于卷积网络的方法.

当前,卷积网络是得到广泛应用的一种人工神经网络，是首个真正被成功训练的深层神经网络，它在人脸识别[1]、图像分类[2]和图像检索[3]等诸多任务中取得了较大的进展[4]. 卷积网络所取得的巨大成功在很大程度上归功于研究者们建立了大规模图像数据集 ImageNet[5]和许多经典的网络架构的提出[6-9]. 由于卷积网络无法直接建模时序信息，所以如何改进卷积网络使其应用于视频中的行为识别是具有挑战性的研究课题. 目前利用卷积网络建模视频表达的方法主要基于以下两种架构：（1）双流卷积网络架构[10]，在该方法中，以 RGB 为输入的空间流网络和以光流图像为输入的时间流网络，先进行独立地训练，然后通过离线方式对独立网络的分类器预测分数进行融合. 在双流网络的架构的基础上，研究者们使用了多种改进方式，拓展双流网络聚合时空特征的范围，获取全局视频表达. （2）3D 卷积网络[11]，该方法是 2D 卷积网络应用于视频数据上的一种直接拓展，2D 卷积核增加时间维度成为 3D 卷积核. 在 3D 卷积计算过程中，可以同时获得视频中图像的空间信息和时序信息. 以上两种网络模型的发展是同步进行的，但使用 3D 卷积核后，网络参数增多，训练网络时需要庞大的数据，在规模较小的数据库上并不适用. 双流卷积网络可以很容易地利用在 ImageNet 数据集上预训练的网络模型[6-9]，所以近几年来有效的方法也是基于双流卷积网络模型的.

双流卷积网络模型[10]以单帧图像及其对应的光流图像作为输入，通过两路卷积网络独立地学习分类器. 训练时随机选择 1 帧 RGB 图像及其对应光流图像分别作为空间流和时间流网络的输入学

习两个分类器，测试时选取多帧图像及其对应的光流图像，分别进入网络进行独立地预测，最终的识别结果是这些预测的均值. 这种网络训练及测试的方式是以某一片段的时空特征代表整段视频，忽略了视频中的长时信息，不能准确建模视频中动作行为复杂的时空结构，而视频的长时结构信息中包含了对识别有用的关键信息. 为了解决这一不足，本文提出了一种基于局部仿射子空间（LASC）编码的双流卷积网络模型，对卷积特征进行编码并聚合整段视频的时空信息，从而有效地建模视频长时信息.

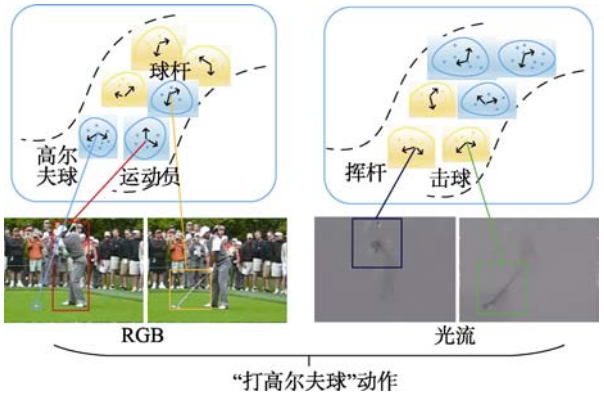



图 1　LASC 的时空特征聚合卷积网络表示行为的物理意义

基于 LASC 的时空特征聚合卷积网络模型与经典的双流网络的不同之处主要体现在以下三个方面：首先，LASC 时空特征聚合网络的输入是在整段视频上进行 RGB 图像及其对应光流图像的采样，这种采样方式得到的多帧图像能够代表整段视频的空间和时间信息；其次，LASC 时空特征聚合卷积网络聚合特征的方式考虑了特征空间的剖分. LASC 的字典由多个视觉词汇组成，视觉词汇周围特征分布的几何结构建模为仿射子空间. 对于空间流，每个视觉词汇表示具有代表性的表观信息，表观相似的物体聚类在一个视觉词汇，例如篮球、足球或者高尔夫球等球类物体通常会聚类在一个视觉词汇所在的仿射子空间中. 对于时间流，每个视觉词汇表示从光流场卷积特征中提取的具有代表性的运动信息，其视觉词汇蕴含的信息比空间流更加丰富，例如跑跳动作中的腿部状态变化或者球类运动中的手臂挥动等不同的运动模式将聚类在不同的视觉词汇中. LASC 嵌入到双流网络中，根据字典中的视觉词汇，可以将行为分别表示成在表观视觉词汇和运动视觉词汇上的 LASC 编码，如图 1 所示，“打高尔夫球”这个行为可以表示为高尔夫球，运动员，球杆，挥杆和击球这 5 个视觉词汇的集合. 最后，双流网络在训练和测试时的采用不同的采样方式，而 LASC 时空聚合网络训练和测试时采用相同的采样方式.

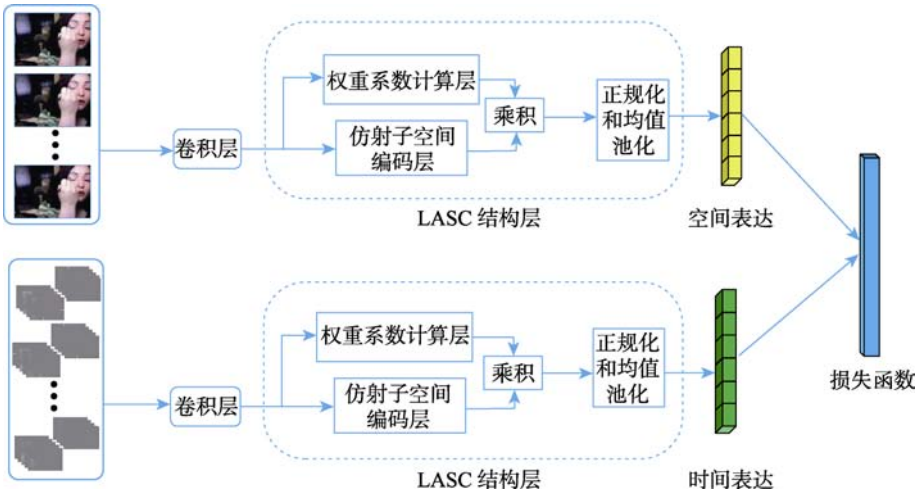



图 2　基于 LASC 的时空特征聚合卷积网络模型结构示意图

基于局部约束仿射子空间编码的时空特征聚合卷积网络模型结构示意图如图 2 所示. 双流网络均由卷积层和 LASC 结构层组成，两路的输入分别为在覆盖整段视频范围内均匀采样的多帧 RGB 图像及其对应的多组光流图像，经过卷积层后得到代表视频长时信息的空间和时间特征，这些特征将作为 LASC 结构层的输入. 根据公式（3），特征的 LASC 编码是特征在每一个子空间上的编码权重和在其对应子空间上编码的乘积，所以本文将 LASC 结构层设计为两个支路，分别是编码权重系数计算层和仿射子空间编码层，第 1 个支路的作用是根据特征到字典中每一个仿射子空间的距离，计算出编码权重系数；第 2 个支路的作用是计算特征描述子在每一个仿射子空间的编码. 两个支路的输出进行相乘即

可得到视频空间特征的 LASC 编码，经过正规化和均值池化，这些编码聚合为视频空间或时间 LASC 表达. 视频空间和时间LASC表达分别经过分类器，得到空间流预测和时间流预测. 空间流网络和时间流网络是独立进行训练的，融合方式为离线的分类器分数融合. 为了验证本文提出网络模型的有效性，在 UCF101、HMDB51 和 Something-V1 这 3 个常用的人体行为识别数据集上进行了实验，本文的方法比经典的双流卷积网络识别准确率分别提升 1.7%、8.7%和 4.3%，达到了与相关方法同样优秀或者超过已有方法的性能，证明了本文方法的有效性.

本文章节安排如下：第 2 节介绍了本文方法的相关工作；第 3 节对如何将 LASC 编码作为结构层嵌入到卷积网络进行了详细的介绍；第 4 节在常用的人体行为识别数据集上进行实验，验证本文方法的有效性；第 5 节为结论部分，总结了本文方法的主要贡献.

## 2 相关工作

本文的相关工作主要体现以下三个方面：（1）基于视觉词袋模型建模视频表达；（2）双流卷积网络的改进方法（3）基于卷积网络的视频长时信息建模方法. 下面将分别介绍研究者们在这三方面的相关工作.

### 2.1 基于视觉词袋模型建模视频表达

视觉词袋模型[13]主要包括以下三个流程：特征提取、生成字典和特征编码. 字典生成的方法较为固定，一般采用无监督聚类的方法如 *k*-means[14]得到字典聚类中心，所以研究者的重点放在特征提取和特征编码这两个过程中. Wang 等人[15]提出了 DT 特征能够很好地建模视频的时空信息，其具体做法是：沿着视频中密集采样点轨迹提取行为局部表观特征和动态特征，如 HOG 特征[16]，HOF 特征[17]和 MBH 特征[17]. Wang 等人[18]在 DT 特征的基础上引入了消除相机抖动的方法，提出了一种改进的密集点轨迹特征（IDT），可以提高建模视频全局信息的鲁棒性. 卷积网络方法被广泛应用于行为识别领域之前，IDT 特征在该领域中处于主导地位. 这些基于密集点轨迹的局部特征通过编特征编码和池化算法聚合形成固定长度的视频表达. 特征编码的过程是通过字典将低维局部特征非线性地映射到高维特征空间，再通过全局池化的方法聚合这些高维特征得到视频全局表达. Peng 等人[19]全面地评估了特征编码方法在人体行为识别领域的应用，并通过实验证明多种编码形成的混合表达性能最佳. 文献[12]中，Li 等人将特征编码方法分为三类：0 阶编码、1 阶编码和高阶编码. 0 阶编码如软分配编码[20]、局部约束线性编码[21]和稀疏编码[22]，通过不同的编码准则为字典中的每个视觉词汇赋一个标量权重；1 阶编码如 VLAD 编码[23]则是指在视觉词汇上的编码是一个向量而不是标量；高阶编码如费舍尔向量[24]则进一步利用了特征分布的高阶信息. 这些方法都在相应的任务上取得了优异的性能提升，所使用的字典通常是由视觉词汇（聚类中心）组成，这种类型的字典一般由 *k*-means 算法生成，是对视觉词汇所在特征空间流形结构粗糙的分段常量近似[14]，没有考虑到字典中视觉词汇周围特征分布的几何结构. 针对这些方法的不足之处，文献[12]提出了 LASC，该方法将低维线性仿射子空间的集合作为字典，每个子空间使用各自的坐标原点及其子空间基向量来刻画特征空间的局部几何结构，这是与其他编码的显著不同之处. 完整的 LASC 具有一阶编码和二阶编码形式，其在多个不同的图像分类任务上取得了优异的性能，具有很强的泛化能力.

### 2.2 双流卷积网络的改进方法

双流卷积网络是行为识别中的主流方法，如引言部分所述，双流网络由独立训练的空间流网络和时间流网络组成，分别用于获取视频中的空间和时间信息. 基于经典的双流卷积网络，研究者们进行了许多改进. Feichtenhofer 等人[25]在双流卷积网络的最后一个卷积层使用 3D 卷积和 3D 池化融合视频中多个采样帧的空间特征和时间特征，该网络拓展了双流卷积网络建模时空信息的范围，但实质也是使用较长的局部时空特征代表全局信息. Wang 等人[26]提出了 TSN 方法，这种方法将整段视频分成多个部分，使用双流卷积网络分别提取每部分中采样帧的时空特征，最后对所有特征进行融合，得到全局视频表达. 由于视频连续多帧的信息有一定的冗余，在视频多个片段中获取局部时空特征然后再融合是较好的选择，但该方法对于不同片段空间和时间特征的融合，仅对不同视频片段的表达进行均值池化融合，从而忽略了不同片段的相关性. 在 TSN 方法的基础上，Rohit 等人[27]提出了一种 ActionVLAD 网络架构，将 VLAD 编码[23]应用于双流卷积网络中，用于聚合视频时空特征，该工作是 NetVLAD[28-29]在视频领域的直接拓展，其主要优点是考虑了特征空间的剖分，其缺点与 VLAD 相同，在聚合时空特征时使用的字典未考虑视觉词汇周围的几何结构. Wang

等人[30]设计了多级时空金字塔策略，使用紧凑的双线性池化方法聚合时空特征，该网络时空特征需要进行 3 次聚合，计算代价较高.

除上述对双流网络的时空特征聚合方式方面的改进之外，改进双流网络的输入数据也是重要的方向，Hakan 等人[31]提出了动态图作为双流网络的输入，对输入数据使用排序池化得到动态图，再构建双流网络，以此构建的双流网络再与经典的双流网络进行融合，该方法需离线计算动态图，在数据计算、存储和读取方面较耗时. Zhang 等人[32]为了使得双流网络在实时视频识别中能够得到应用，提出了使用运动矢量代替光流作为时间流的输入，其速度比 TSN 方法[26]提升 15 倍，但是性能上却比 TSN 方法低 10%左右.

### 2.3　基于卷积网络的视频长时信息建模方法

如何利用卷积网络获取视频长时信息一直以来是研究的热点问题. Ng 等人[33]将视频长时序的卷积特征作为 5 层长短时记忆网络（LSTM）的输入，从而建模视频的长时信息. Varol 等人[34]提出了由 5 个长时卷积层构成的网络，该网络能够建模较长时的视频表达，该方法最多只能建模 60 帧的视频，限制了其应用范围. 在基于 RGB-D 数据的人体行为识别领域，LSTM 及其改进方法有着广泛的应用，Si 等人[35]提出了使用多层 LSTM 构建时序堆栈网络，用于学习时序变化信息. 在此工作基础上，文献[36]使用图卷积和注意力机制改进了 LSTM 内部结构，LSTM 在序列长时建模方面具有一定的潜力，但其缺点在于参数量较大，网络训练代价较高. 同时，文献[37]的实验表明，对于时空特征进行有序的 LSTM 聚合，性能低于对这些特征进行无序的编码聚合，所以对长时信息进行 LSTM 聚合的方法在目前基于 RGB 数据的人行为识别领域并没有得到广泛地应用.

受到以上三个方面相关工作的启发，为了改进经典的双流卷积网络，建模视频长时信息，本文提出将 LASC 编码作为结构层嵌入到双流卷积网络模型，设计了一种时空特征聚合卷积网络模型.

## 3　局部约束仿射子空间编码结构层

局部约束仿射子空间编码结构层主要由权重系数计算层和仿射子空间编码层两部分构成. 3.1 节主要介绍 LASC 所要解决的优化问题，该优化目标函数是经典的脊回归问题[38]，具有解析解，该解析解使得 LASC 作为结构层可以嵌入到卷积网络中进行端到端的优化. 3.2 节主要介绍 LASC 结构层的嵌入方法，3.2.1 节介绍该结构层的第 1 个支路，即权重系数计算层，该层可由卷积层和 *softmax* 层实现. 3.2.2 节主要介绍 LASC 结构层中的第 2 个支路，即仿射子空间编码层，该层可以分解为残差层和仿射子空间映射层两个子层. 3.3 节主要介绍对仿射子空间映射层参数施加正交约束的三种方法.

### 3.1　LASC的优化目标函数及其解析解

如第 2 节所述，经典的特征编码方法未考虑到字典中视觉词汇周围特征分布的几何结构. 而在 LASC 中，其字典考虑了视觉词汇所在子空间的几何结构，使用仿射子空间建模每个视觉词汇周围的特征分布，且仿射子空间的维度不大于原始特征的维度. LASC 的字典可以表示为一组附着在聚类中心的低维仿射子空间：

$$\boldsymbol{S}=\{\boldsymbol{S}_i \triangleq \boldsymbol{\mu}_i+\boldsymbol{A}_i\boldsymbol{c}_i, \boldsymbol{\mu}_i\in\mathbb{R}^d\}, i=1,\cdots,M \tag{1}$$

其中，$\boldsymbol{\mu}_i$ 由 *k*-means 算法得到，表示第 $i$ 个仿射子空间的附着点. $\boldsymbol{A}_i\in\mathbb{R}^{d\times p}$ 的列向量构成仿射子空间的基向量，$d$ 为特征的维度，$p$ 为子空间的维度. $\boldsymbol{c}_i\in\mathbb{R}^p$ 是特征在第 $i$ 个仿射子空间的编码向量，$M$ 表示仿射子空间的个数，$\boldsymbol{S}_i$ 定义了一个以 $\boldsymbol{\mu}_i$ 为原点的局部坐标系，可看作对视觉词汇所在特征空间的分段线性近似.

对于单个局部特征 $\boldsymbol{y}$ 进行 LASC 编码，使用近邻的 $k$ 个仿射子空间进行编码，也就是使用临近度量来约束特征 $\boldsymbol{y}$ 在仿射子空间的投影向量. LASC 的目标函数为：

$$\min_{\forall \mathbf{c}_i}\sum_{S_i\in N_k^S(\boldsymbol{y})}\left\|(\boldsymbol{y}-\boldsymbol{\mu}_i)-\boldsymbol{A}_i\boldsymbol{c}_i\right\|_2^2+\lambda\boldsymbol{d}(\boldsymbol{y},\boldsymbol{S}_i)\left\|\boldsymbol{c}_i\right\|_2^2 \tag{2}$$

这里，$\boldsymbol{d}(\boldsymbol{y},\boldsymbol{S}_i)$ 是特征到仿射子空间的距离. 目标函数（2）可以解耦为若干独立的脊回归问题，能够在每个近邻仿射子空间 $\boldsymbol{S}_i\in N_k^S(\boldsymbol{y})$ 分别求解编码向量 $\boldsymbol{c}_i$，目标函数（2）具有解析解：

$$\boldsymbol{c}_i=\alpha_y^i\boldsymbol{A}_i^T(\boldsymbol{y}-\boldsymbol{\mu}_i) \tag{3}$$

$\boldsymbol{A}_i$ 表示仿射子空间的基向量，$\alpha_y^i=(1+\lambda\boldsymbol{d}(\boldsymbol{y},\boldsymbol{S}_i))^{-1}$ 为特征 $\boldsymbol{y}$ 在子空间 $\boldsymbol{S}_i$ 的编码权重.

### 3.2　LASC结构层

根据 3.1 节中 LASC 编码的解析解（3），本文将 LASC 结构层嵌入到双流卷积网络中进行端到端的优化，LASC 结构层的作用是聚合人体行为的空间和时间特征，建模整段视频空间和时间的信息. 假设 $\boldsymbol{y}_{jt}$ 是人体行为视频中第 $t$ 帧（$t\in\{1,\cdots,T\}$）的空间位置 $j$（$j\in\{1,\cdots,N\}$）的 $d$ 维局部特征. LASC 结构层的建模形式如下：

$$\boldsymbol{L}_i=\sum_{t=1}^{T}\sum_{j=1}^{N}\alpha_y^i\boldsymbol{A}_i^T(:,k)(\boldsymbol{y}_{jt}(k)-\boldsymbol{\mu}_i(k)) \tag{4}$$

$\boldsymbol{L}_i\in\mathbb{R}^p$ 是 $\boldsymbol{y}_{jt}$ 在第 $i$ 个子空间聚合时空特征的编码向量. LASC 结构层将 $\boldsymbol{y}_{jt}$ 与 $\boldsymbol{\mu}_i$ 的残差投影到子空间所定义的局部坐标系上，经过 $\alpha_y^i$ 加权后得到在第 $i$ 个仿射子空间上的视频特征编码 $\boldsymbol{L}_i$，级联每一个仿射子空间上的编码 $\boldsymbol{L}_i$ 可以得到全局视频表达.

为了将 LASC 作为结构层嵌入到双流网络中，本文设计了两个支路实现（4）式中的建模形式. 图 3 是以空间流为例的 LASC 结构层示意图，该结构层的输入是视频空间流的卷积特征. 特征首先进行 L2 范数正规化处理，然后进入 LASC 结构层，该结构层由权重系数计算层和仿射子空间编码层两个支路组成，权重系数计算层的主要目的是计算特征在子空间上编码的权重，仿射子空间编码层的主要目的是计算特征在子空间上的编码.

3.2.1 权重系数计算层

如图 3 所示，权重系数计算层是 LASC 结构层的上方分支部分，该层的主要作用是计算特征在每一个仿射子空间编码的权重. 3.1 节计算编码权重的方法是根据特征 $\boldsymbol{y}$ 到仿射子空间的距离找到 $k$ 个近邻仿射子空间，通过计算 $\alpha_y^i=(1+\lambda\boldsymbol{d}(\boldsymbol{y},\boldsymbol{S}_i))^{-1}$ 得到编码权重. 但是在网络训练过程中，特征 $\mathbf{y}$ 和字典子空间的附着点 $\boldsymbol{\mu}_i$ 是同时更新的，需要在每次迭代时计算进行 $\boldsymbol{y}$ 到 $\boldsymbol{\mu}_i$ 的欧式距离，确定特征的 $k$ 近邻仿射子空间，从而获得 $\alpha_y^i$. 这种情况下 $\boldsymbol{y}$ 和 $\boldsymbol{\mu}_i$ 的更新是耦合的，为了简化更新方式和利用 CNN 中固有的模块（例如卷积层和 *softmax* 层），受到 NetVLAD[28-29] 的启发，权重系数计算层通过解耦的方式更新 $\boldsymbol{y}$ 和 $\boldsymbol{\mu}_i$. 此时局部特征 $\boldsymbol{y}$ 在第 $i$ 个字典子空间的编码权重为

$$\alpha_y^i=\frac{\exp(-\gamma\|\boldsymbol{y}-\boldsymbol{\mu}_i\|_2^2)}{\sum_{i'=1}^{M}\exp(-\gamma\|\boldsymbol{y}-\boldsymbol{\mu}_{i'}\|_2^2}} \tag{5}$$

$\gamma$ 是引入的超参数，当 $\gamma$ 取较大值，等价于特征在 $k$ 近邻仿射子空间进行编码. 当 $\gamma$ 无穷大时，等价于特征只在最近邻的一个仿射子空间上进行编码. 公式（5）进一步写成如下形式：

$$\alpha_y^i=\frac{\exp(\boldsymbol{w}_i^T\boldsymbol{y}+\boldsymbol{b}_i)}{\sum_{i'=1}^{M}\exp(\boldsymbol{w}_{i'}^T\boldsymbol{y}+\boldsymbol{b}_{i'})} \tag{6}$$

其中，$\boldsymbol{w}_i=2\gamma\boldsymbol{\mu}_i$ 和 $\boldsymbol{b}_i=-\gamma\|\boldsymbol{\mu}_i\|_2^2$ 是可学习的参数，$\boldsymbol{\mu}_i$ 的初始化通过对训练集样本的采样特征进行 $k$-means 聚类得到. 权重系数计算层可以通过卷积层和 *softmax* 层实现.

3.2.2 仿射子空间编码层

如图 3 所示，仿射子空间编码层是 LASC 结构层下方分支部分，该层由残差层和仿射子空间映射层两个子层组成. 残差层的功能是计算特征 $\boldsymbol{y}$ 和 $\boldsymbol{\mu}_i$ 之间的残差，通过只有偏置 $-\boldsymbol{\mu}_i$ 的 $M$ 个卷积层实现. 仿射子空间映射层的功能是计算残差在其对应子空间上的投影. 仿射子空间映射层的参数 $\boldsymbol{U}_i$ 初始化为仿射子空间的一组正交基，$\boldsymbol{U}_i$ 初始化的具体计算过程是求仿射子空间中所有特征 $\boldsymbol{y}$ 的协方差矩阵，对协方差矩阵进行 SVD 分解的左酉矩阵即是 $\boldsymbol{U}_i$，$\boldsymbol{U}_i$ 是 *Gram* 矩阵为单位阵的标准正交基. 子空间映射层可以由 $M$ 个权重为 $\boldsymbol{U}_i\in\mathbb{R}^{d\times p}$ 且无偏置的卷积层实现.

仿射子空间编码层的输出与权重系数计算层的输出进行相乘运算，再经过两次正规化操作[28-29]，第 1 次对每个仿射子空间的编码进行 L2 范数正规化，第 2 次对级联获得的完整 LASC 编码进行 L2 范数正规化，最后经过均值池化得到空间流 LASC 视频表达，维度为 $Mp$.

下面给出 LASC 结构层的伪代码如算法 1 所示. 在 LASC 结构层中，权重系数计算层的参数量为 $M\times d+d$，仿射子空间编码层的参数量为 $M\times d+$

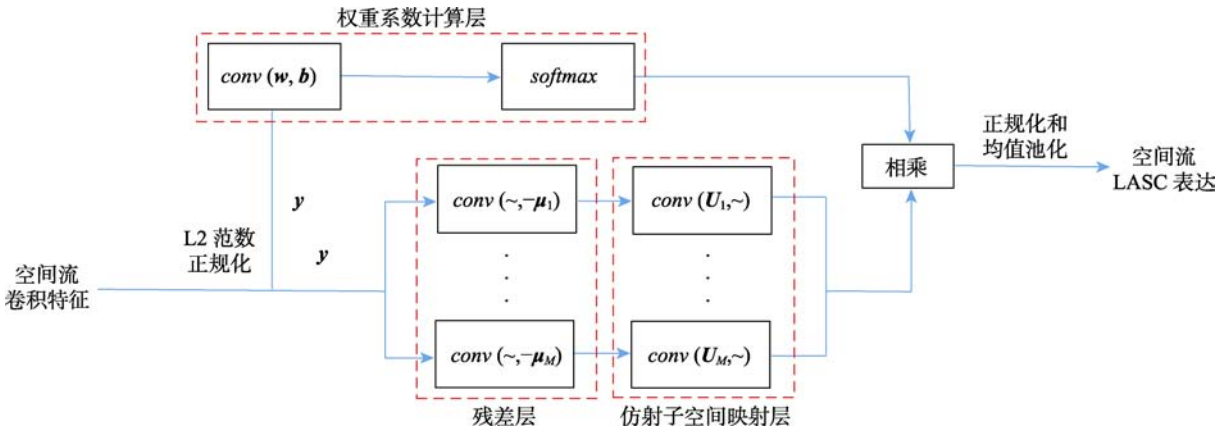



图 3 空间流网络嵌入局部约束仿射子空间编码（LASC）结构层示意图

$M\times d\times p$. 由此可见，LASC 层的参数量主要与仿射子空间的个数 $M$ 和维度 $p$ 有关，在实验部分将对这两个参数对性能的影响进行详细分析.

**算法 1**. LASC 结构层

输入：视频空间或时间特征 $\boldsymbol{y}\in\mathbb{R}^{(T\times N)\times d}$

仿射子空间出初始化参数：附着点 $\boldsymbol{\mu}_i$ 和子空间正交基 $\boldsymbol{U}_i$，其中 $i=1,\cdots,M$.

输出：视频 LASC 表达 $\boldsymbol{L}$

1. 对 $\boldsymbol{y}$ 进行二范数归一化
2. FOR $i$ =1 to $M$
3. 通过权重系数计算层计算 $\boldsymbol{y}$ 在第 $i$ 个子空间的编码权重 $\boldsymbol{\alpha}_i$： $\boldsymbol{\alpha}_i\leftarrow softmax\left(conv_{(\boldsymbol{w}_i,\boldsymbol{b}_i)}(\boldsymbol{y})\right)$
4. 通过仿射子空间编码层计算 $\boldsymbol{y}$ 在第 $i$ 个子空间的编码 $\boldsymbol{l}_i$： $\boldsymbol{l}_i\leftarrow conv_{(\boldsymbol{U}_i,\sim)}(conv_{(\sim,-\boldsymbol{\mu}_i)}(\boldsymbol{y}))$
5. 计算 $\boldsymbol{y}$ 在第 i 个子空间的编码 $\boldsymbol{L}_i$： $\boldsymbol{L}_i\leftarrow\boldsymbol{\alpha}_i\times\boldsymbol{l}_i$
6. $\boldsymbol{L}_i$ 进行二范数归一化
7. END FOR
8. $\boldsymbol{y}$ 在 M 个子空间上的编码级联： $\boldsymbol{L}_{T\times N}\leftarrow(\boldsymbol{L}_1,\cdots,\boldsymbol{L}_M)$
9. $\boldsymbol{L}_{T\times N}$ 进行二范数归一化
10. $\boldsymbol{L}_{T\times N}$ 进行均值池化得到视频 LASC 表达 $\boldsymbol{L}$

### 3.3 仿射子空间映射层参数的正交约束

如 3.2.2 节所述，仿射子空间映射层参数 $\boldsymbol{U}_i$ 初始化时为一组正交基，由于在网络训练过程中会破坏 $\boldsymbol{U}_i$ 的正交特性，所以需要对 $\boldsymbol{U}_i$ 施加正交约束. 受到文献[39]的启发，本文使用软正交约束、无穷范数约束和谱范数约束三种方法使 $\boldsymbol{U}_i$ 在参数更新过程中保持正交. 对参数 $\boldsymbol{U}_i$ 施加正交约束，则网络的损失函数为

$$Loss=\overline{Loss}+R_{\boldsymbol{U}_i} \tag{7}$$

其中 $\overline{Loss}$ 表示真实值和预测值之间的损失函数，$R_{\boldsymbol{U}_i}$ 为正则项，约束 $\boldsymbol{U}_i$ 在优化过程中保持正交. 在以下的 3.3.1 节至 3.3.3 节将详细阐述 $R_{\boldsymbol{U}_i}$ 的三种不同形式. 在网络进行反向传播时，将 $R_{\boldsymbol{U}_i}$ 对于 $\boldsymbol{U}_i$ 的导数附加到 $\boldsymbol{U}_i$ 更新过程中.

#### 3.3.1 软正交约束

对仿射子空间映射层的参数施加软正交约束（SOR, Soft Orthogonality Regulation），该正则项如下所示：

$$R_{\boldsymbol{U}_i}=\lambda\left\|\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}\right\|_F^2 \tag{8}$$

这里，$\lambda$ 为引入的超参数，且 $\lambda>0$. 对目标函数施加式（8）的约束，在反向传播时，其可以看作是权重衰减项，使 $\boldsymbol{U}_i$ 在参数优化过程中保持正交. 参数更新时，$R_{\boldsymbol{U}_i}$ 对于 $\boldsymbol{U}_i$ 的导数为 $4\lambda\boldsymbol{U}_i\left(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}\right)$ 附加到 $\boldsymbol{U}_i$ 更新过程中.

#### 3.3.2 无穷范数约束

对于仿射子空间映射层的参数 $\mathbf{U}_i\in\mathbb{R}^{d\times p}$，其相关系数定义为

$$\mu_w=\max_{m\neq n}\frac{\left\|\langle\boldsymbol{w}_m,\boldsymbol{w}_n\rangle\right\|}{\left\|\boldsymbol{w}_m\right\|\cdot\left\|\boldsymbol{w}_n\right\|} \tag{9}$$

这里，$\boldsymbol{w}_m$ 代表 $\boldsymbol{U}_i$ 的第 $m$ 列，$m=1,...,p$. 相关系数 $\mu_w\in[0,1]$ 用来衡量 $\boldsymbol{U}_i$ 任意两列之间的相关性. 为了使 $\boldsymbol{U}_i$ 列正交或者接近正交，$\mu_w$ 的值应该为 0 或者趋于 0. $\langle\boldsymbol{w}_m,\boldsymbol{w}_n\rangle$ 可以看作是 Gram 矩阵的 $\boldsymbol{U}_i^T\boldsymbol{U}_i$ 的第 $m$ 行 $n$ 列，这里 $m\neq n$.

如果仅考虑非对角元素，则可以通过以下约束矩阵无穷范数的方式达到使得相关系数 $\mu_w$ 最小化的目的，正则项有如下形式：

$$R_{\boldsymbol{U}_i}=\lambda\left\|\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}\right\|_\infty \tag{10}$$

该正则项称为无穷范数约束（INR, Infinity-norm Regularization）.

#### 3.3.3 谱范数约束

根据文献[40]关于矩阵谱有限等距性质的研究，$\boldsymbol{U}_i$ 满足有限等距性质：对于所有向量 $\boldsymbol{z}\in\mathbb{R}^n$，有 $r$ 个非零元素，则存在一个很小的数 $\delta_w\in(0,1)$，满足

$$(1-\delta_w)\leqslant\frac{\left\|\boldsymbol{U}_i\boldsymbol{z}\right\|^2}{\left\|\boldsymbol{z}\right\|^2}\leqslant(1+\delta_w) \tag{11}$$

当 $r=n$ 时，谱有限等距性质可以重写为以下形式：

$$\left|\frac{\|\boldsymbol{U}_i\boldsymbol{z}\|^2}{\|\boldsymbol{z}\|^2}-1\right|\leqslant\delta_w,\forall\boldsymbol{z}\in\mathbb{R}^n \tag{12}$$

令 $\sigma(\boldsymbol{U}_i)=sup\left(\frac{\|\boldsymbol{U}_i\boldsymbol{z}\|}{\boldsymbol{z}},\boldsymbol{z}\in\mathbb{R}^n,\boldsymbol{z}\neq0\right)$ 是 $\boldsymbol{U}_i$ 的谱范数，$sup$ 代表上确界. 容易得到

$$\sigma(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})=\sup_{\boldsymbol{z}\in\mathbb{R}^n,\boldsymbol{z}\neq0}\frac{\left\|\boldsymbol{U}_i\boldsymbol{z}\right\|}{\boldsymbol{z}} \tag{13}$$

从谱有限等距性质角度约束 $\boldsymbol{U}_i$ 正交，则等价于在 $r=n$ 的特殊情况下，最小化有限等距常量 $\delta_w$，即最小化 $\sigma(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})$. 那么，正则项有如下形式：

$$R_{\boldsymbol{U}_i}=\lambda\cdot\sigma(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}) \tag{14}$$

该正则项称之为谱范数约束（SNR, Spectral-norm Regulations）. 进行网络前向计算时，由于计算式（14）需要对 $\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}$ 进行本征分解，计算代价较大. 为了减少计算代价，本文使用能量迭代法[41-42]计算 $\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}$ 的谱范数. 首先随机初始化一个向量 $\boldsymbol{u}\in\mathbb{R}^n$，通过迭代

$$\boldsymbol{v}\leftarrow\frac{(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})^T\boldsymbol{u}}{\left\|(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})^T\right\|},\boldsymbol{u}\leftarrow\frac{(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})\boldsymbol{v}}{\left\|(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})\boldsymbol{v}\right\|} \tag{15}$$

得到 $\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I}$ 的谱范数：

$$\sigma(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})=\boldsymbol{u}^T(\boldsymbol{U}_i^T\boldsymbol{U}_i-\boldsymbol{I})\boldsymbol{v} \tag{16}$$

能量迭代法近似谱范数的计算复杂度为 $O(mn^2)$，而使用本征分解法的计算复杂度 $O(n^3)$，能量迭代法显著降低了计算复杂度. 在本文中，迭代次数设置为 15. 谱范数近似计算的具体推导过程见附录 1.

## 4 实 验

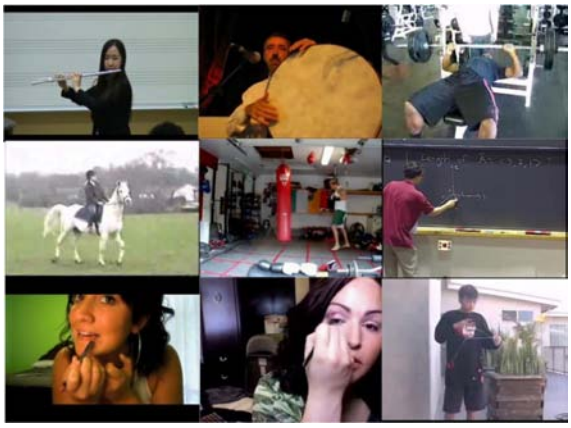
### 4.1 数据集介绍

本节使用 3 个常用的人体行为识别数据集 UCF101[43]、HMDB51[44]和 Something-V1①对基于局部约束仿射子空间编码的时空特征聚合卷积网络进行性能评估. 其中所有消融实验是在 HMDB51 数据集的第 1 个划分上进行的.

如图 4（a）（b）（c）所示，分别为 UCF101、HMDB51 和 Something-V1 数据集中的视频经过拆帧后的 RGB 图像示例. UCF101 包含 101 种人体行为，共计 13 320 个视频序列，该数据集由体育运动行为组成. HMDB51 包含 51 种人体行为，共计 6 766 个视频序列，每一类行为至少拥有 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，行为的类内差异大，背景复杂，是目前难度较大的数据集之一. 这两个数据集提供 3 种不同的划分训练集和测试集的方式，3 个划分上的平均识别准确率作为最终的分类结果. Something-V1 数据集包含 174 种人体行为，共计 108 499 个视频序列，均是人与物品的交互行为. 该数据集只有 1 种划分训练集和测试集的方式，评价指标为识别准确率.

### 4.2 实验设置

实验主机配置：CPU 为 Intel Core i7-4770K，3.50GHz，64GB 内存，GPU 为两块 NVIDIA GTX1080ti. 本文算法使用 TensorFlow②深度学习工具包实现. 空间流初始化模型为在 ImageNet 数据库[5]上预训练的 VGG-16 模型[7]，时间流的初始化模型是双流卷积网络的时间流模型在 UCF101 和 HMDB51 数据集上微调后的 VGG-16 模型. 训练时，每个视频片段中均匀采样 25 帧 RGB 图像及其对应的光流图像分别作为空间流网络和时间流网络的输入. 均匀采样的含义是对于一段视频，已知视频的帧数，同时设定采样帧数，据此可以确定采样间隔，然后以此采样间隔对视频进行采样. 对 RGB 图像及其对应的光流图像进行随机裁剪，处理后空间流输入图像大小为 $224\times224\times3$，时间流输入图像大小为 $224\times224\times20$. 权重系数计算层中的超参数 $\gamma$ 设置为 1000.



（a）UCF101 数据集 RGB 图像示例

（b）HMDB51 数据集 RGB 图像示例

（c）Something-V1 数据集 RGB 图像示例

图 4 数据集 RGB 图像示例

网络优化使用自适应矩估计算法[45]，其中超参数设置为 $\varepsilon=1\mathrm{e}-4,\beta_1=0.9,\beta_2=0.999$. 网络训练过程采用两阶段方式：第一阶段仅训练分类器，学习率设为 1e-2，迭代 5000 次后降为 1e-3，再经过 5000

① https://20bn.com/datasets/something-something/v1

② https://tensorflow.google.cn

次迭代. 第二个阶段训练分类器，LASC 结构层和 conv5 层. 学习率设为 1e-4，迭代 5000 次后降为 1e-5，再经过 3000 次迭代. 空间流网络和时间流网络是独立训练的，双流网络的融合过程是在测试过程中进行的. 测试时，在对一个包含多帧的视频进行分类时，从视频片段中均匀采样 25 帧，每一帧图像的左上角，右上角，左下角，右下角和中间裁剪出 5 个图像块，对于空间流网络图像块的大小为 224×224×3，共 5 组，每组中 25 帧 RGB 图像可以聚合为一个空间表达，对于时间流网络图像块的大小为 224×224×20，共 5 组，每组中的 25 帧光流图像可以聚合为一个时间表达. 每组图像块作为输入独立地进行预测，这 10 组图像块的预测均值将作为该视频片段的最终预测.

### 4.3　参数敏感性分析

首先，本小节实验在 HMDB5 数据集的第 1 个划分上评估了 LASC 结构层的字典子空间的数量 $M$ 和字典子空间的维度 $p$. 然后，评估了三种不同正交约束方法中的超参数 $\lambda$.

（1）评估仿射子空间的数量 $M$. 实验设置为：使用谱范数正则约束仿射子空间映射层的参数 $\boldsymbol{U}_i$，字典子空间的维度 $p=512$，超参数 $\lambda$=0.5. 实验结果如图 5 所示，在 $M$ 从 4 逐渐增加至 64 的过程中，分类准确率分别为 45.9%、48.2%、50.1%、50.3%、50.8%和 52.2%，随着仿射子空间数量的增加分类准确率不断提升. 在 $M=80$ 时，识别准确率有所下降，所以在以下的实验中设置仿射子空间数量 $M=64$.

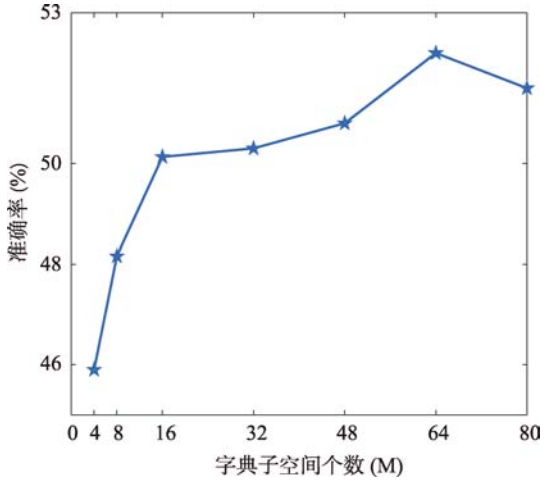



图 5　HMDB51 第 1 个划分上评估字典子空间数量（$M$）

（2）评估仿射子空间的维度 $p$. 此时实验设置为：使用谱范数正则约束仿射子空间映射层的参数，超参数 $\lambda$=0.5. 如图 6 所示，当 $p$ 分别取 64，128，256，384, 512 时，分类准确率呈现明显的上升趋势，当 $p$=512 时，分类准确率达到 52.2%，此时子空间的维度和特征的维度一致，说明高维度可以更好地刻画视觉词汇所在仿射子空间的几何结构. 字典子空间正交基的维度 $p$ 小于等于特征的维度，$p$ 最大取值为 512. 根据以上的参数评估实验，在以下的实验中，固定字典子空间的数目为 $M=64$，字典子空间的维度 $p$=512.

（3）评估三种不同正交约束方法中的超参数 $\lambda$. 实验设置为：$M=64$，$p$=512. 图 6 为 $\lambda$ 取 0.01 到 0.9 的不同数值情况下，软正交约束，无穷范数约束和谱范数约束这三种不同约束的分类准确率情况.

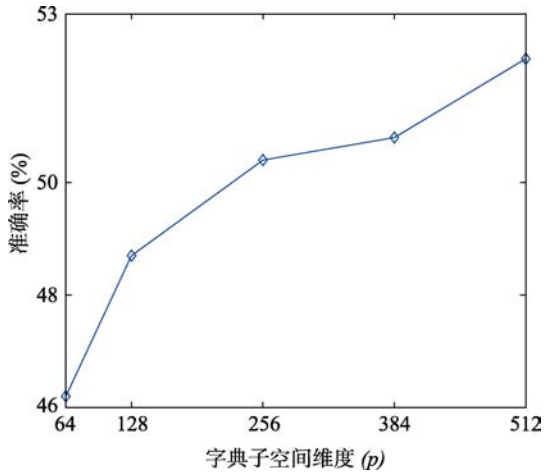



图 6　HMDB51 第 1 个划分上评估字典子空间维度（$p$）

如图 7 所示，三种正交约束的方法达到最佳性能时，$\lambda$ 取值有所不同；同时，识别准确率随正交约束项中的超参数 $\lambda$ 的变化趋势并没有遵循一定的规律，这说明 $\lambda$ 是需要精细调试的参数. 在评估 $\lambda$ 的同时，由图 7 的三条曲线可知在三种正交约束方法中，谱范数约束的性能较好，软正交约束的性能

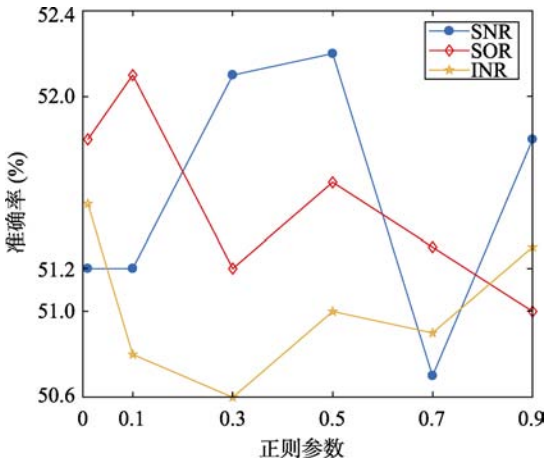



图 7　HMDB51 数据集第 1 个划分上评估三种不同正交约束的超参数 $\lambda$

优于无穷范数约束. 在评估 $\lambda$ 的同时，对三种正交约束的性能也相当于对不同正交约束方法进行了评估.

为了方便查看不同正交约束方法取得的最佳性能，将图 7 中三种正交约束的最佳性能进一步整理到表 1 中. 如表 1 所示，软正交约束在 $\lambda$=0.1时，分类准确率达到 52.1%，对于无穷范数约束，当 $\lambda$=0.01时，分类准确率达到 51.5%，而对于谱范数约束，当 $\lambda$=0.5时，分类准确率为 52.2%. 可以看出这三种正交约束方式都十分有效，软正交约束的方法与谱范数约束方法性能基本相同. 而最好的识别准确率是在谱范数约束下产生的，所以在以下实验中将采用谱范数约束 $\boldsymbol{U}_i$，且设置超参数 $\lambda$=0.5.

**表 1 HMDB51 数据集 split1 上三种不同正交约束下 LASC 聚合网络性能比较（空间流）**

| 正交约束方法 | $\lambda$ | 准确率（%） |
|---|---|---|
| 软正交约束（SOR） | 0.1 | 52.1 |
| 无穷范数约束（INR） | 0.01 | 51.5 |
| 谱范数约束（SNR） | 0.5 | 52.2 |

### 4.4 本文方法和双流卷积网络模型比较

#### 4.4.1 LASC 结构层分别嵌入空间流和时间流网络与双流卷积网络模型比较

根据 4.2 节和 4.3 节确定的参数，本小节展开了一组消融实验，评估 LASC 结构分别嵌入到空间流和时间流网络的贡献，比较的基线是双流卷积网络模型，识别准确率是 UCF101 和 HMDB51 数据集 3 个划分上的平均准确率. 双流卷积网络的测试方法是从视频片段中随机选择 25 帧，从图像中裁剪出 10 个图像块，对于空间流网络每个图像块的大小为 224×224×3，共计 250 个，而对时间流网络该图像块的大小为 224×224×20，共计 250 个，每个图像块作为输入进入网络中独立地进行预测，这 500 个图像块预测均值作为该视频片段的最终预测. 本文的测试方法已在 4.2 小节中介绍. 两种方法的本质区别在于双流网络仅在测试时进行多帧采样，而本文方法在训练和测试时都对视频进行多帧采样.

如表 2 所示，在 UCF101 数据集上，本文方法的空间流网络、时间流网络和融合后的网络分别比双流卷积网络的方法提升了 3.2%、1.2%和 1.7%. 在 HMDB51 数据集上，该提升分别 8.5%、5.6%和 8.7%. 以上实验结果验证了将 LASC 模块嵌入到双流卷积网络模型对空间和时间流网络性能的提升都起到了作用，在空间流网络上的提升高于时间流. 同时，融合双流网络性能的提升可以证明 LASC 聚合网络产生鲁棒的全局视频表达，提高了基于视频的人体行为识别准确率.

**表 2 消融实验：LASC 分别嵌入空间流和时间流的贡献**

| | 方法 | UCF101 | HMDB51 |
|---|---|---|---|
| 双流卷积网络 | 空间流 | 78.4 | 42.2 |
| | 空间流+LASC | **81.6** | **50.7** |
| | 时间流 | 87.0 | 55.0 |
| | 时间流+LASC | **88.2** | **60.6** |
| | 融合空间流和时间流 | 91.4 | 58.5 |
| | 融合空间流+LASC 和时间流+LASC | **93.1** | **67.2** |

#### 4.4.2 视频表达维度及网络参数量比较

在使用 4.2 节和 4.3 节确定的参数情况下，表 3 对本文方法与双流卷积网络的视频表达维度和参数量进行了比较. 本文方法和双流卷积网络均是对空间流网络和时间流网络进行离线融合的，融合方法是将两个训练好的网络的预测分数加权产生最终的预测分数，那么全局视频表达的维度是空间流网络和时间流全局视频表达维度的加和. 如表 3 所示，双流卷积网络输出生成的视频表达维度为 8K，本文方法的空间流表达和时间流表达都是 32K，最终视频表达维度 64K，高维表达损失更少的视频信息，能够建模覆盖视频的空间和时间信息. 在设计网络架构时，参数量也是需要考虑的一个重要参数，将直接影响网络的训练和收敛速度.

**表 3 本文方法和双流卷积网络在视频表达维度及网络参数量的比较**

| 方法 | 视频表达维度 | 模型参数 |
|---|---|---|
| 双流卷积网络[46-47] | 8K | 268M |
| 本文方法 | 64K | 66M |

在参数量方面，由于 LASC 结构层代替了双流卷积网络中的全连接层，本文方法的模型参数为 66M，仅是双流卷积网络的 1/4，但在性能上比双流网络有了较大地提升，这一现象能够证明，由于 LASC 结构层的嵌入，使得双流网络能够建模高维度且分辨能力更强的非线性视频空间表达和时间表达.

### 4.5 本文方法与其他行为识别方法比较

#### 4.5.1 UCF101 数据集和 HMDB51 数据集

在 4.3 节确定的最优参数的情况下，本小节将本文方法在 UCF101，HMDB51 这 2 个数据集上与其他行为识别方法进行了比较，实验结果如表 4 所示. 表中的结果可以分为两个部分，首先是本文方

法与其他识别方法分别在两个数据集的 3 个划分上平均准确率的比较；其次是本文方法融合 IDT 特征与其他方法融合 IDT 特征的比较.

**表 4　本文方法与其他行为识别方法的准确率（%）比较**

| 方法 | 网络架构 | UCF101 | HMDB51 |
|---|---|---|---|
| 双流卷积网络[46-47] | VGG-16 | 91.4 | 58.5 |
| 双流卷积网络+3D 卷积+3D 池化[25] | VGG-16 | 92.5 | 66.4 |
| C3D [11] | 3D 卷积网络 | 85.2 | – |
| TDD+FV [48] | 卷积特征+编码 | 90.3 | 63.2 |
| TSN [26] | BN-Inception | 94.2 | 69.4 |
| LTC [34] | 长时卷积网络 | 91.7 | 64.8 |
| ST-Pyramid[30] | VGG-16+时空特征金字塔融合 | 93.2 | 66.8 |
| ActionVLAD [27] | VGG-16 | 92.7 | 66.2 |
| 本文方法 | VGG-16 | **93.1** | **67.2** |
| C3D + IDT[11] | 3D 卷积网络 | 90.4 | – |
| LTC + IDT[34] | 长时卷积网络 | 92.7 | 67.2 |
| ActionVLAD + IDT[27] | VGG-16 | 93.6 | 69.8 |
| 本文方法+IDT | VGG-16 | **93.9** | **70.1** |

如表 4 所示，在不与 IDT 特征融合的情况下，本文方法在 UCF101 和 HMDB51 上准确率达到了 93.1%和 67.2%，比双流卷积网络[46-47]方法提升了 1.7%和 8.7%，说明 LASC 结构层分别嵌入到双流网络中，对视频片段的空间和时间特征进行聚合，生成全局视频表达，能够较好地改进双流卷积网络. 本文方法比 3D 卷积+3D 池化方法[25]在两个数据集上分别提升了 0.6%和 0.8%，该方法在聚合空间和时间特征的方法是 3D 卷积和 3D 池化操作，聚合了覆盖 50 帧的视频空间和时间信息，但并没有覆盖到整段视频，存在获取视频运动变化信息不足的现象. 本文方法与 C3D[11]相比，优势十分明显，在 UCF101 数据集上有 7.5%的提升. C3D 的输入是视频中某一短时的连续采样帧，在这样的情况下，该方法也没有建模视频中的长时信息. 本文方法比 TDD+FV[48]在两个数据集上分别提升 1.4%和 2.4%，该方法是传统的分离式框架，使用卷积特征和传统的 FV 编码，没有进行端到端的优化，而本文方法将传统的编码方法设计为卷积网络中的标准模块进行端到端的优化，能够同时发挥两者的优势. LTC[34]网络模型设计了输入为视频中连续较长片段时的长时卷积网络模型，其卷积层采用 3D 卷积核，最长能够建模连续 60 帧的时序范围，也并没有完全覆盖整个视频，该方法在两个数据集上均低于本文方法，这是由于本文方法的输入方式能够获取覆盖整段视频的空间和时间的信息. TSN 方法[26]与本文的本质不同之处在于聚合不同视频片段的方式上，首先，TSN 聚合的主要对象是不同视频片段的空间和时间表达，聚合方法是均值池化的方法，而本文是通过 LASC 编码的方式聚合不同视频片段的空间和时间特征. 需要说明的是本文方法与 TSN 方法的实验设置具有显著不同：（1）二者使用的基础网络不同，本文使用 VGG-16 模型，而 TSN 使用性能更强的 BN-Inception 模型，而根据 TSN[26]中的实验结果，利用 BN-Inception 模型构建的双流卷积网络在 UCF101 数据集第 1 个划分上的识别准确率比 VGG-16 模型高 1.1%.（2）TSN 方法使用了更多模态的信息. 具体地，本文方法使用了 RGB 图像和光流图像两种模态信息，而 TSN 方法则使用了 RGB 图像、光流图像、变形光流图像三种模态信息. 为了公平比较，我们补充了一组新的实验，在相同实验设置基础上对二者进行了比较. 具体地，我们采用在 ImageNet 上预训练的 VGG-16 模型，以及 RGB 图像（空间流）和光流图像（时间流）两种模态信息，在 UCF101 数据集的第 1 个划分上进行实验. TSN 的识别准确率为 92.3%，本文方法的识别准确率为 93.2%，能够证明本文方法的优势. ST-Pyramid[30]在 UCF101 上的结果基本与本文一致，而在 HMDB51 数据集上本文方法更有优势，这是由于 HMDB51 数据集中的视频中物体数量多于 UCF101，而本文方法的仿射子空间字典能够采样到这些基本物体词汇，再通过对这些基本物体词汇进行编码，得到的表达能够建模视频的完整信息. 同时，ST-Pyramid 对空间和时间特征进行了金字塔聚合，聚合方式比本文更加复杂. ActionVLAD[27]将 VLAD 编码嵌入到双流卷积网络中，本文方法在两个数据集上比该方法提升 0.4%和 0.7%，这主要是本文方法的 LASC 结构层所使用的仿射子空间字典考虑了视觉词汇周围的几何结构.

下面我们将本文方法与 IDT 特征进行融合，并与同样融合了 IDT 特征的其他方法进行比较. 本文使用 Peng 等人[19]的方法，在视频中提取 IDT 特征，即沿着视频轨迹提取 HOG，HOF 和 MBH 特征，使用 FV 对 3 种特征进行编码，得到的 3 种编码向量分别送入 SVM 分类器，最后对不同编码向量的预测得分取均值作为 IDT 特征的 SVM 分数. 本文方法与 IDT 特征进行融合时，IDT 特征的 SVM 分数与空间流和时间流的视频表达预测分数进行离线加权融

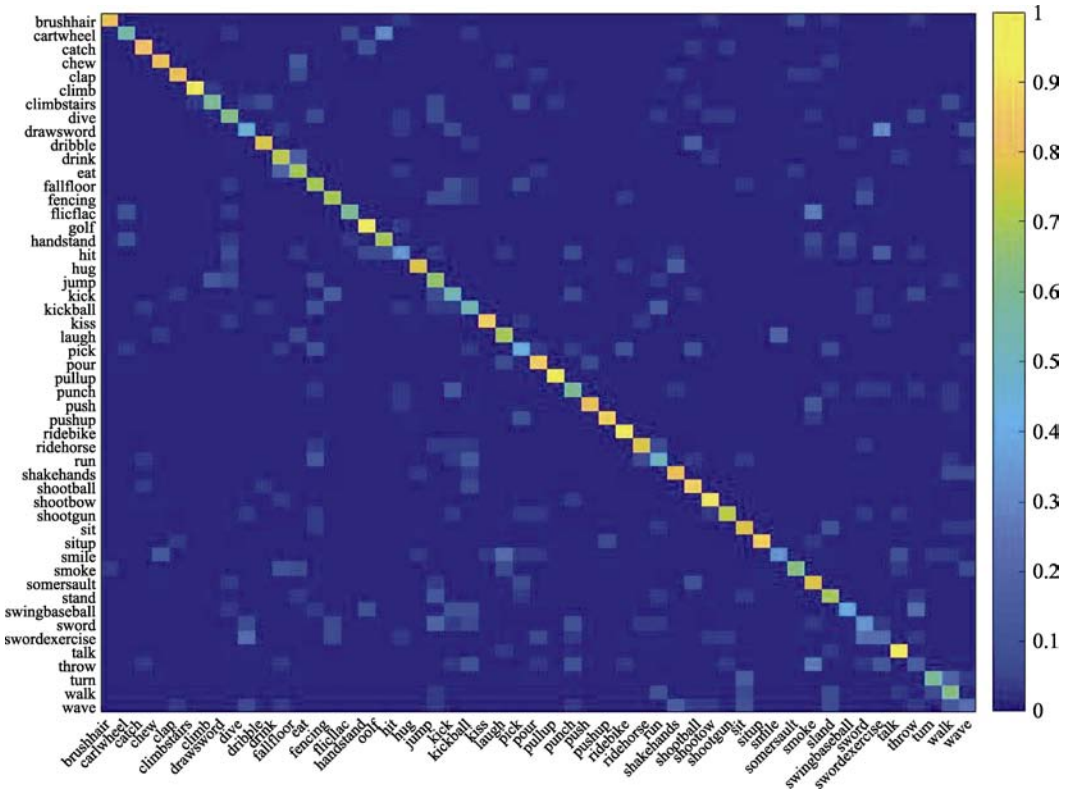



图 8 本文方法在 HMDB51 数据集上的混淆矩阵

合. 在融合 IDT 特征之后，本文方法得到进一步提升，在 UCF101 数据集和 HMDB51 数据集上分别提高 0.8%和 2.9%，说明本文方法与 IDT 特征具有一定的互补性. 与其他融合了 IDT 特征的方法相比，也具有一定优势. 综上所述，相比其他人体行为识别算法，本文方法在 UCF101 和 HMDB51 取得了具有竞争力或者更好的效果.

4.5.2 Something-V1 数据集

本小节主要在 Something-V1 数据集上仅使用 RGB 图像进行实验，即在只评估空间流网络性能的情况下，将本文方法与双流卷积网络和 TSN 方法进行比较. 由于对比算法没有进行在该数据集的实验，表 5 中对比算法的结果均为复现的实验结果.

**表 5 Something-V1 数据集上本文方法与其他行为识别方法的准确率（%）比较**

| 方法 | 模型 | 准确率（%） |
|---|---|---|
| 双流卷积网络[46-47] | VGG-16 | 8.9 |
| TSN [26] | VGG-16 | 11.6 |
| 本文方法 | VGG-16 | 13.2 |

如表 5 所示，双流卷积网络和 TSN 的识别准确率分别为 7.6%和 11.6%，本文方法比双流卷积网络方法提升 4.3%，比 TSN 方法提升 1.6%，能够证明 LASC 结构层在聚合视频局部特征时起到了作用.

### 4.6 本文方法混淆矩阵分析

混淆矩阵能够统计出识别算法在每一类的识别率以及最容易混淆的动作类别. 本节将主要分析本文方法的混淆矩阵，图 8 为本文方法在 HMDB51 的第 1 个划分上的混淆矩阵. 如图 7 所示，本文方法对于“引体向上”、“骑自行车”、“攀登”和“打高尔夫球”这 4 个动作类识别效果较好，识别准确率分别达到了 100%、100%、97%和 97%，然而本文方法在“练剑”、“挥手”和“扔”这 3 个动作类的识别效果比较差，识别准确率仅为 17%、23%和 23%. 最容易有混淆的两组动作分别是“侧手翻”和“单手倒立”以及“练剑”和“拔剑”. 这两组动作有着较强的相似性，识别时发生混淆的原因很大可能是两个动作包含相同的视觉词汇，所以不同的动作表示为相同的视觉词汇组合. 例如练剑和拔剑这两个动作都包含“剑”这类表观视觉词汇，同时也都包括“挥动手臂”这类运动视觉词汇.

## 5 结 论

本文针对经典双流卷积网络无法建模视频中长时信息的不足，提出了基于局部约束仿射子空间编码的时空特征聚合卷积网络模型. 该网络能够聚合

整段视频的时空特征，获得全局视频表达. 本文主要有以下贡献：首先，将图像分类中的 LASC 编码方法设计为结构层嵌入到了双流卷积网络中，实现了能够端到端优化的时空特征聚合网络. 其次，在 LASC 结构层中，使用软正交约束，无穷范数约束和谱范数约束这三种正则方法约束子空间正交基在网络训练过程中保持正交. 另外，在使用谱范数约束时，本文使用能量迭代法近似计算矩阵谱范数，降低了计算复杂度. 本文在 HMDB51 数据库上进行了一系列消融实验，确定了模型参数，在 UCF101，HMDB51 和 Something-V1 数据集的 RGB 图像进行的实验表明，本文提出的方法优于其他行为识别方法，融合 IDT 特征能进一步提升识别准确率. 同时，在双流卷积网络中嵌入 LASC 结构层，网络参数量减少了 4 倍，证明 LASC 结构层取代了双流卷积网络中的全连接层，获得了高维度且更具有分辨能力的全局视频非线性表达.

**致谢**　感谢国家自然科学基金（No. 61471082）的资助. 感谢审稿专家和编辑在百忙之中审阅本文!

## 参　考　文　献

[1] Li Xiao-Xin, Liang Rong-Hua. A review for face recognition with occlusion: From subspace regression to deep learning. Chinese Journal of Computers, 2018, 41(1): 177-207 (in Chinese)
(李小薪, 梁荣华. 有遮挡人脸识别综述:从子空间回归到深度学习. 计算机学报, 2018, 41(1): 177-207)

[2] He Guo-Cai, Liu Xia-Bi. Unsupervised visual representation learning with image triplets mining. Chinese Journal of Computers, 2019, 42(3): 2787-2803 (in Chinese)
(何果财, 刘峡壁. 基于图像三元组挖掘的无监督视觉表示学习. 计算机学报, 2018, 41(12): 2787-2803)

[3] Liu Hao-Miao, Wang Rui-Ping, Shan Shi-Guang, et al. Learning to hash with discrete optimization. Chinese Journal of Computers, 2019, 42(5): 1149-1160 (in Chinese)
(刘昊淼, 王瑞平, 山世光等. 基于离散优化的哈希编码学习方法. 计算机学报, 2019, 42(5): 1149-1160)

[4] Zhang Shun, Gong Yi-Hong, Wang Jin-Jun. The development of deep convolutional neural network an its applications on computer vision. Chinese Journal of Computers, 2019, 42(3): 453-482 (in Chinese)
(张顺，龚怡宏，王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 2019, 42(3): 453-482)

[5] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252

[6] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks//Proceedings of the International Conference on Neural Information Processing Systems. Nevada, USA, 2012: 1097-1105

[7] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets//Proceedings of the British Machine Vision Conference. Nottingham, UK, 2014

[8] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[9] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 448-456

[10] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos//Proceedings of the Conference on Neural Information Processing Systems. Montreal, Canada, 2014. 568-576

[11] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks//Proceedings of the International Conference on Computer Vision. Santiago, USA, 2015: 4489-4497

[12] Li P, Lu X, Wang Q, et al. From dictionary of visual words to subspaces: Locality-constrained affine subspace coding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2348-2357

[13] Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos//Proceedings of the International Conference on Computer Vision. Sydney, Austrilia, 2003: 1470-1477

[14] Canas G, Poggio T, Rosasco L. Learning manifolds with k-means and k-flats//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2012: 2465–2473

[15] Wang H, Klaser A, Schmid C. Liu C. Action recognition by dense trajectories//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado, USA, 2011: 3169-3176

[16] Dalal N, Triggs B. Histograms of oriented gradients for human detection. computer vision and pattern recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893

[17] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance//Proceedings of the European Conference on Computer Vision. Graz, Austrilia, 2006: 428-441

[18] Wang H, Schmid C. Action recognition with improved trajectories//Proceedings of the International Conference on Computer Vision. Sydney, Austrilia, 2013: 3551-3558

[19] Peng X, Wang L, Wang X, Qiao Y. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding. arXiv preprint arXiv:1706.06905, 2014

[20] Van Gemert J C, Veenman C J, Smeulders A W, et al. Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(7): 1271-1283

[21] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3360-3367

[22] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching

using sparse coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2009: 1794-1801

[23] Jegou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3304-3311

[24] Sanchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision, 2013, 105(3): 222-245

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

[26] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition//Proceedings of the European Conference on Computer Vision. Amsterdam, the Netherlands, 2016: 20-36

[27] Girdhar R, Ramanan D, Gupta A, et al. ActionVLAD: Learning spatial-temporal aggregation for action classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA, 2017: 3165-3174

[28] Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1437-1451

[29] Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5297-5307

[30] Wang Y, Long M, Wang J, et al. Spatiotemporal pyramid network for video action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 1529-1538

[31] Bilen H, Fernando B, Gavves E, et al. Action recognition with dynamic image networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2799-2813

[32] Zhang B, Wang L, Wang Z, et al. Real-time action recognition with deeply-transferred motion vector CNNs. IEEE Transactions on Image Processing, 2018, 27(5): 2326-2339

[33] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: deep networks for video classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4694–4702

[34] Varol G, Laptev I, Schmid C, et al. Long-term temporal convolutions for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1510-1517

[35] Si C, Jing Y, Wang W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning//Proceedings of the European Conference, Munich, Germany, 2018: 106-121

[36] Si C, Chen W, Wang W, et al. An attention enhanced graph convolutional LSTM network for skeleton-Based action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1227-1236

[37] Miech A, Laptev I, Sivic J. Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706.06905, 2017

[38] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, Inference, and Prediction. 2nd Edition. Berlin, Germany: Springer, 2009

[39] Bansal N, Chen X, Wang Z, et al. Can we gain more from orthogonality regularizations in training deep networks//Proceedings of the Conference Neural information processing systems. Montréal, Canada, 2018: 4266-4276

[40] Emmanuel J Candes, Terence Tao. Decoding by linear programming. IEEE Transactions on Information Theory, 2005, 51(12): 4203–4215

[41] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. arXiv preprint arXiv: 1802.05957, 2018

[42] Yoshida Y, Miyato T. Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv: 1705.10941, 2017

[43] Khurram Soomro, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv: 1212.0402, 2012

[44] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition//Proceedings of the International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563

[45] Kingma D, Ba J. Adam: A method for stochastic optimization// Proceedings of the International Conference for Learning Representations. San Diego, USA, 2015

[46] Wang X, Farhadi A, Gupta A. Actions-transformations// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5297-5307

[47] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets. arXiv preprint arXiv: 1507.02159, 2015

[48] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4305-4314

## 附录1. 能量迭代法近似计算矩阵谱范数的推导过程

假设矩阵 $\boldsymbol{B}$ 的谱范数为 $\sigma(\boldsymbol{B})=\sqrt{\lambda_{\max}(\boldsymbol{B}^T\boldsymbol{B})}$，其中 $\lambda_{\max}(\boldsymbol{B}^T\boldsymbol{B})$ 为 $\boldsymbol{B}^T\boldsymbol{B}$ 的最大特征值. 下面推导能量迭代法近似求解矩阵 $\boldsymbol{B}$ 谱范数的过程：

假设 $\boldsymbol{G}=\boldsymbol{B}^T\boldsymbol{B}$，$\boldsymbol{G}$ 的各个特征根 $\lambda_1,...,\lambda_n$ 中，最大的特征根严格大于其余的特征根，那么 $\boldsymbol{G}$ 的特征向量 $\boldsymbol{\eta}_1,...,\boldsymbol{\eta}_n$ 构成完备的基底

$$\boldsymbol{u}^{(0)}=c_1\boldsymbol{\eta}_1+\cdots+c_n\boldsymbol{\eta}_n \tag{1}$$

每次执行迭代过程：

$$\boldsymbol{v}\leftarrow\frac{\boldsymbol{B}^T\boldsymbol{u}}{\left\|\boldsymbol{B}^T\boldsymbol{u}\right\|},\boldsymbol{u}\leftarrow\frac{\boldsymbol{B}\boldsymbol{v}}{\left\|\boldsymbol{B}\boldsymbol{v}\right\|} \tag{2}$$

忽略式（2）中 $\frac{\boldsymbol{B}^T\boldsymbol{u}}{\left\|\boldsymbol{B}^T\boldsymbol{u}\right\|}$ 和 $\frac{\boldsymbol{B}\boldsymbol{v}}{\left\|\boldsymbol{B}\boldsymbol{v}\right\|}$ 的归一化的过程，迭代过程相当于式（1）的两端进行矩阵 $\boldsymbol{G}$ 的重复左乘，即

$$\boldsymbol{G}^\tau\boldsymbol{u}^{(0)}=c_1\lambda_1^\tau\boldsymbol{\eta}_1+\cdots+c_n\lambda_n^\tau\boldsymbol{\eta}_n \tag{3}$$

其中，$\tau$ 为迭代次数，$\lambda_1$ 为矩阵 $\boldsymbol{G}$ 的最大特征根，式（3）两端均除以 $\lambda_1^\tau$：

$$\frac{\boldsymbol{G}^\tau\boldsymbol{u}^{(0)}}{\lambda_1^\tau}=c_1\boldsymbol{\eta}_1+c_2\frac{\lambda_2^\tau}{\lambda_1^\tau}\boldsymbol{\eta}_2+\cdots+c_n\frac{\lambda_n^\tau}{\lambda_1^\tau}\boldsymbol{\eta}_n \tag{4}$$

根据假设，$\boldsymbol{G}$ 最大的特征根大于其余的特征根，则式（4）中的 $\lambda_i/\lambda_1(i\in[2,n])$ 均小于 1，当 $\tau\to\infty$ 时，式（4）中的 $\lambda_i^\tau/\lambda_1^\tau(i\in[2,n])$ 均趋于零，上式可以重写为

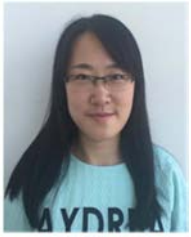
$$\frac{\mathbf{G}^\tau\boldsymbol{u}^{(0)}}{\lambda_1^\tau}\approx c_1\boldsymbol{\eta}_1 \tag{5}$$

忽略模长，式（5）表明，$\tau$ 足够大时，$\boldsymbol{G}^\tau\boldsymbol{u}^{(0)}$ 提供了最大特征值对应特征向量的近似方向，每一步的归一化是为了防止溢出．则 $\boldsymbol{u}=\boldsymbol{B}^\tau\boldsymbol{u}^{(0)}/\left\|\boldsymbol{B}^\tau\boldsymbol{u}^{(0)}\right\|$ 就是对应的单位特征向量，即

$$\boldsymbol{G}\boldsymbol{u}=\lambda_1\boldsymbol{u} \tag{6}$$

$$\boldsymbol{u}^T\boldsymbol{G}\boldsymbol{u}=\lambda_1\boldsymbol{u}^T\boldsymbol{u}=\lambda_1 \tag{7}$$

即可求出矩阵 $\boldsymbol{B}$ 的谱范数的平方.



**ZHANG Bing-Bing,** Ph.D. candidate. Her current research interests include human action recognition, image classification and deep learning.

**LI Pei-Hua**, Ph. D., professor, Ph. D. supervisor. His research interests include image/video recognition, object detection and semantic segmentation.

**SUN Qiu-Le**, Ph.D. candidate. His research interests include image recognition, semantic segmentation

## Background

Video-based human action recognition has drawn increasing attention considering its potential in a wide range of applications, such as video surveillance, human-computer interaction and social video recommendation. For action recognition in videos, there are two crucial and complementary cues: appearances and temporal dynamics. To model the two cues above, most recent video representations for action recognition are primarily based on two different CNN architectures: (1) 3D spatial-temporal convolutions that potentially learn complicated spatial-temporal dependencies but the large amount of parameters in 3D CNNs make it hard to train in practice; (2) Two-stream architectures that decompose the video into motion and appearance streams, and train separate CNNs for each stream, fusing the outputs in the end.

Two-stream architecture-based methods have been widely-used in academic community due to its less computation complex than 3D CNNs. However, two-stream architectures largely disregard the long-term temporal structure of the video and essentially learn a classifier that operates on individual frames or short clip of few (up to 10) frames, possibly enforcing consensus of classification scores over different segments of the video. At test time, T (typically 25) uniformly sampled frames (with their motion descriptors) are classified independently and the classifications scores are averaged to get the final prediction.

Based on two-stream architecture, we present a structure layer to aggregate the long-term temporal information covered the whole video. The aggregation method is inspired by our previous work on the IEEE Conference on Computer Vision and Pattern Recognition of 2015, which is called as “From dictionary of visual words to subspaces: Locality-constrained affine subspace coding”. We extend LASC to a structure layer with learnable parameters that can plug into the CNNs to realize end-to-end parameter optimization. The structure layer is inserted to the last layer convolutional layer of spatial stream and temporal stream to acquire more robust high-dimension representation and the two fully-connected layer in two-stream architecture is totally replaced. It is noted that the parameter of the prosed network is 4 times less than the original two-stream model. As the experiment results shown, our method achieves the better performance compared to the other human action recognition algorithm with the comparison of relative fairness.

This work is supported by the National Natural Science Foundation of China (NO. 61471082).