一种基于深度度量学习的视频分类方法

智洪欣 于洪涛* 李邵梅 高 超 王艳川
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要：针对视频分类中普遍面临的类内离散度和类间相似性较大而制约分类性能的问题，该文提出一种基于深度度量学习的视频分类方法。该方法设计了一种深度网络，网络包含特征学习、基于深度度量学习的相似性度量，以及分类3个部分。其中相似性度量的工作原理为：首先，计算特征间的欧式距离作为样本之间的语义距离；其次，设计一个间隔分配函数，根据语义距离动态分配语义间隔；最后，根据样本语义间隔计算误差并反向传播，使网络能够学习到样本间语义距离的差异，自动聚焦于难分样本，以充分学习难分样本的特征。该网络在训练过程中采用多任务学习的方法，同时学习相似性度量和分类任务，以达到整体最优。在UCF101和HMDB51上的实验结果表明，与已有方法相比，提出的方法能有效提高视频分类精度。

关键词：视频分类；深度学习；自适应间隔；深度度量学习；多任务学习
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2018)11-2562-08
DOI: 10.11999/JEIT171141

A Deep Metric Learning Based Video Classification Method

ZHI Hongxin YU Hongtao LI Shaomei GAO Chao WANG Yanchuan
(*National Digital Switching System Engineering & Technological Research Center*, *Zhengzhou* 450002, *China*)

Abstract: To solve the common problem of classification performance restriction caused by big intra-class variations and inter-class similarities in video classification domain, this paper proposes a deep metric learning based video classification method. The proposed method designs a deep network which contains three parts: feature learning, deep metric learning based similarity measure as well as classification. The principle of similarity measure is: Firstly, the Euclidean distance between features is calculated as the semantic distance between samples. Secondly, a margin distributing function is designed to dynamically allocate margin in the basis of the semantic distances. Finally, the difference of the sample semantic distance can be learned by calculating the loss and propagating it backwards so as to the network can automatically focus on the hard negative samples and more fully learn the characteristic of them. With a multi-task learning training method in the training stage, the similarity measure and classification can be learned jointly. Experimental results on UCF101 and HMDB51 show that the proposed method can effectively improve the classification precision.

Key words: Video classification; Deep learning; Adaptive margin; Deep metric learning; Multi-task learning

1 引言

随着互联网上视频数量的急速增加，视频分类技术成为计算机视觉领域的研究热点之一。该技术在视频检索、视频监管和人工智能等领域中发挥着重要作用，具有广阔的应用前景。基于深度学习的视频分类方法[1–4]具有能够通过反向传播自动从视频数据中快速学习最合适的特征和能够联合优化特征提取和分类两个步骤，最大程度地发挥两者的协作性，使整个系统的性能达到最优[5]的优势，受到越来越多研究者的青睐，已成为该领域的主流研究方法。

当前，基于深度学习的视频分类方法的研究工作主要集中在特征提取这方面。视频中包含丰富的空域信息和时域信息，充分有效地提取视频特征，特别是提取时域特征历来是视频分类领域中的一个难题。文献[6]提出在视频片段上进行3D卷积来学习时空域特征，把卷积神经网络(Convolutional Neural Network, CNN)扩展到了时域。Simonyan等人[7]提出双流卷积网络(two-stream convolutional network)，把视频帧和相应的光流图分别作为网络的输入，提取视频的空域特征和时域特征。

收稿日期：2017-12-04；改回日期：2018-08-14；网络出版：2018-08-20
*通信作者： 于洪涛 yht_ndsc@139.com
基金项目：国家自然科学基金青年科学基金(61601513)
Foundation Item: The Young Scientists Fund of the National Natural Science Foundation of China (61601513)

文献[8]提出的TS-LSTM网络使用ResNet101[9]提取视频帧特征，然后使用长短时记忆(Long-Short Term Memory, LSTM)网络学习视频的时域特征。最近，Wang等人[10]提出时域分割网络(Temporal Segment Networks, TSN)来捕获长时时域动态信息，取得了较好的分类精度。

视频中由场景和角度的变化、运动速度和方向的变化等带来的较大的类内差异和类间相似性是制约视频分类性能的重要因素。单纯依靠特征提取不能测量样本之间的相似度和差异性。而度量学习能够通过学习样本之间的语义距离对相似度和差异性进行度量，其目标之一是减小类内离散度和类间相似性，提高分类器的判决能力，在计算机视觉相关任务中已得到广泛应用。与神经网络结合的深度度量学习，通过反向传播在学习特征的同时直接从原始数据学习非线性嵌入[11–13]。Bell等人[14]在视觉检索任务中提出使用对比损失函数来训练网络。FaceNet[15]使用三元组损失函数学习人脸确认和识别的特征嵌入。三元组损失方法[13,15]通过学习使异类样本的语义距离比同类样本至少大一个间隔。文献[13–15]提出的方法一次只利用了一个样本对的语义距离信息，Song等人[11]提出的基于三元组的结构化特征嵌入能充分利用一个批量中所有样本的信息。然而这些方法[11–15]都没有考虑不同负向样本之间的差异，在进行度量学习的过程中对所有负向样本赋予同等的重要性。但是，很明显，距离越近的负向样本在分类过程中越容易产生误判，对分类效果的影响越大。因此，在训练网络时应该更加关注这类样本。

为此，本文对原有基于三元组的结构化特征嵌入[11]进行了改进，提出一个自适应间隔的深度度量学习方法。该方法依据样本对之间的语义距离动态分配语义间隔，使网络能够更加关注距离近的负向样本，从而提高分类精度。在此基础上，设计了一个融合时域分割网络、时域池化和相似性度量的网络，在充分提取视频帧空域信息和光流图时域信息的基础上，通过时域池化形成视频的整体表征，并且在训练过程中使用多任务学习的方法，使网络能够同时学习相似性度量和分类，最终得到一个全局最优的网络。

综上所述，本文的主要贡献如下：

(1)提出一个融合多任务学习和深度度量学习的视频分类网络(Network Fusing Multi-task learning and Metric Learning, NFMML)来同时学习视频特征、相似性度量和分类以及它们之间的关系。通过同时学习特征度量和分类，网络达到整体最优。

(2)提出一种基于自适应间隔的深度度量学习方法来学习视频的类内离散度和类间相似性，通过间隔分配函数动态分配间隔使网络更加关注难分样本，进而提高分类精度。本文通过定义间隔分配函数，把语义距离不同带来的影响以损失函数的形式呈现出来，并通过实验验证了该方法的有效性。

2　算法描述

2.1　提出的视频分类网络

在基于深度学习的视频分类方法中，Simonyan等人[7]提出的双流网络能更灵活地对空域特征和时域特征进行进一步处理，本文以此为基础进行网络设计，处理架构如图1所示。从图1可以看到，本文提出的网络包含时域和空域流，分别对空域和时域信息建模。空域流把RGB帧作为输入，时域流把堆叠的光流图作为输入。空域流和时域流的网络结构同通用的CNN网络相同，由卷积、池化和激活层组成。已证明，和其他分类任务类似，采用更深层次的网络结构能够提高双流网络的分类精

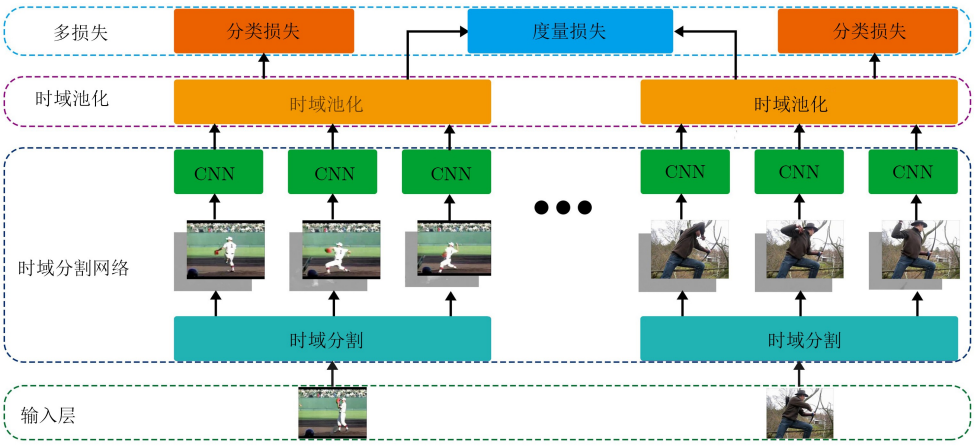


图 1 本文提出的整体网络结构

度[16]。考虑到计算消耗和分类精度的平衡，本文采用BN-Inception[9]网络作为基础CNN网络。空域流采用在ImageNet预训练的模型上用UCF101数据集中的RGB帧进行分类微调。对于时域流，本文遵循常用做法，把连续10帧的光流图作为网络的输入。并且在RGB帧预训练的基础上对时域流进行微调：尽管光流图和视频帧不同，这仍然有利于提高分类精度[10]。

视频的时域动态信息包含用于指明视频类别的重要信息。然而，传统的双流网络只能捕获短时动作信息，忽略了长时动作信息以及外观、物体和场景等动态演化信息。事实上，视频中连续帧之间有较大的冗余信息，而时域分割网络[10]通过随机分段、稀疏采样的方法能够捕获长时时域信息。首先把视频随机分为K个子视频，从每个子视频中随机采样一个视频段，每个视频段由1个视频帧和10张相应的连续光流图组成。然后把采样的视频帧和光流图分别输入到空域和时域流中得到帧特征。最后采用时域池化把所有子视频的帧特征聚合为视频特征。

时域分割网络能够捕获长时时域动态信息，而时域池化能够使网络用一个特征来表征整个帧序列，前者是作用在通道上而后者是作用在特征图上，进一步增强了特征的表征能力。常用的时域池化方法有平均池化、最大池化和卷积池化[17]等。在实践中，这3种池化方法的性能相近，平均池化略好[17]，因此，本文采用平均池化的方法。

为了使网络能够学习类内离散度和类间相似性，本文在网络中引入深度度量学习结构。与此同时，在训练阶段使用多任务学习的方法，在网络中同时进行度量学习和分类任务，避免网络只专注于度量学习而给分类网络带来偏置。在测试阶段，只使用分类分支，得到空域流和时域流的分类得分后，通过平均融合得到视频的分类结果。

在多任务学习训练方法的设计理念下，NFM-ML总的损失函数为

$$J = \lambda_1 J_c + \lambda_2 J_m \tag{1}$$

其中J是一个批量中所有样本的总体损失，J_c和J_m分别是分类损失和度量损失。λ_1和λ_2分别是分类损失和度量损失的权重。通过调整λ_1和λ_2，可以调整分类损失和度量损失在网络学习中的重要性。本文将在实验部分探讨λ_1和λ_2对分类性能的影响。

式(1)中的分类损失J_c由softmax分类器得到，其损失函数为

$$\left.\begin{aligned} J_c &= -\frac{1}{M}\left[\sum_{i=1}^{M}\sum_{j=1}^{C}\mathbb{I}\left[y^i=j\right]\lg(p_{ij})\right] \\ p_{ij} &= \frac{\exp(z_i)}{\sum_{l=1}^{C}\exp(z_l)} \end{aligned}\right\} \tag{2}$$

其中，M是一个批量中的样本数量，C是数据的类别数量，y^i是样本x_i的标签。$\mathbb{I}[\cdot]$是指示函数，当表达式为真时输出为1否则输出为0。z是分类分支的正则化输出。p_{ij}是样本x_i属于类j的概率。

softmax损失函数的梯度如式(3)：

$$\left.\begin{aligned} \frac{\partial \tilde{J}_i}{\partial z_i} &= p_{ij} - \mathbb{I}\left[y^i=j\right] \\ p_{ij} &= \frac{\exp(z_i)}{\sum_{l=1}^{C}\exp(z_l)} \end{aligned}\right\} \tag{3}$$

下面将详细介绍式(1)中的度量损失J_m。

2.2 基于自适应间隔的深度度量学习

在深度度量学习的研究中，文献[11]提出通过学习一个基于三元组的结构化特征嵌入来利用一个批量中所有样本的语义距离信息来训练网络，其损失函数为

$$\left.\begin{aligned} J_m &= \frac{1}{2\left|P\right|}\sum_{(i,j)\in P}\max(0,\tilde{J}_{i,j})^2 \\ \tilde{J}_{i,j} &= \lg\left(\sum_{(i,k)\in N}\exp\left\{\alpha - D_{i,k}\right\} + \sum_{(j,l)\in N}\exp\left\{\alpha - D_{j,l}\right\}\right) + D_{i,j} \end{aligned}\right\} \tag{4}$$

其中，P和N分别表示一个批量中的正向样本对集合和负向样本对集合。$D_{i,j}$表示样本对(x_i, x_j)之间的语义距离：$D_{i,j} = \|f(x_i) - f(x_j)\|_2$，$f(\cdot)$表示度量学习分支学习到的嵌入特征，$\alpha$是语义间隔。式(4)的含义是负向样本到正向样本的语义距离应该大于正向样本对之间的距离加上α之和。该方法不仅能够挖掘难分样本，还能充分利用该批量中所有样本的语义距离信息。但是文献[11]提出的损失函数中间隔α是常量，即对所有负向样本分配同样的间隔。如前所述，语义距离越小的负向样本越容易造成误判，如果忽略语义距离的差异，而对所有负向样本分配统一的常量间隔，显然不能充分利用这些信息。为此，本文引入自适应间隔，使网络更加关注这些负向样本，从而提高模型的分类效果。

基于上述思路，本文提出了自适应间隔策略，

并设计了如下的间隔分配函数：

$$\left.\begin{aligned}&\alpha_{i,k}=\exp(D_{i,j}/D_{i,k})\\&\alpha_{j,l}=\exp(D_{i,j}/D_{j,l})\\&D_{i,k}\neq 0,\ D_{j,l}\neq 0,(x_i,x_j)\in P,\\&\qquad(x_i,x_k)\in N,\ (x_j,x_l)\in N\end{aligned}\right\}\tag{5}$$

其中，$D_{i,j}$是正向样本对(x_i,x_j)之间的语义距离，$D_{i,k}$和$D_{j,l}$分别是负向样本对(x_i,x_k), (x_j,x_l)之间的语义距离。

采用上述间隔分配函数，最后的度量损失函数为

$$\left.\begin{aligned}&J_m=\frac{1}{2\left|P\right|}\sum_{(i,j)\in P}\max(0,\tilde{J}_{i,j})^2\\&\tilde{J}_{i,j}=\lg\left(\sum_{(i,k)\in N}\exp\left\{\alpha_{i,k}-D_{i,k}\right\}\right.\\&\qquad\left.+\sum_{(j,l)\in N}\exp\left\{\alpha_{j,l}-D_{j,l}\right\}\right)+D_{i,j}\\&\alpha_{i,k}=\exp(D_{i,j}/D_{i,k})\\&\alpha_{j,l}=\exp(D_{i,j}/D_{j,l})\\&D_{i,k}\neq 0,\ D_{j,l}\neq 0,(x_i,x_j)\in P,\\&\qquad(x_i,x_k)\in N,\ (x_j,x_l)\in N\end{aligned}\right\}\tag{6}$$

根据式(5)，对于距离较近的负向样本，网络会分配较大的间隔，这样在计算损失时，其计算得到的残差较大，在反向传播过程中，对网络参数学习的贡献越大，即网络更加关注这些难分样本。除了上述指数函数，本文还探讨了对数函数$\alpha_{\lg}=\lg(D_{i,j}/D_{i,k})$和比值函数$\alpha_{\mathrm{r}}=D_{i,j}/D_{i,k}$作为间隔分配函数时的情况，性能比较在实验部分呈现。

上述度量损失函数J_m的梯度为

$$\frac{\partial J_m}{\partial D_{i,j}}=\frac{1}{|P|}\tilde{J}_{i,j}\mathbb{I}\left[J_{i,j}>0\right]\tag{7}$$

$$\frac{\partial J_m}{\partial D_{i,k}}=\frac{1}{|P|}\tilde{J}_{i,j}\mathbb{I}\left[\tilde{J}_{i,j}>0\right]\cdot\frac{-\{(D_{i,j}/D_{i,k}^2)\alpha_{i,k}+1\}\exp\{\alpha_{i,k}-D_{i,k}\}}{\exp\{\tilde{J}_{i,j}-D_{i,j}\}}\tag{8}$$

$$\frac{\partial J_m}{\partial D_{j,l}}=\frac{1}{|P|}\tilde{J}_{i,j}\mathbb{I}\left[\tilde{J}_{i,j}>0\right]\cdot\frac{-\left\{\left(D_{i,j}/D_{j,l}^2\right)\alpha_{j,l}+1\right\}\exp\{\alpha_{j,l}-D_{j,l}\}}{\exp\{\tilde{J}_{i,j}-D_{i,j}\}}\tag{9}$$

式中，$D_{i,k}\neq 0$, $D_{j,l}\neq 0$。

在训练过程中，网络把分类损失和度量损失的梯度相加再反向传播到网络的其余部分。

3　实验结果及分析

3.1　数据集和实验参数设置

本文使用Caffe[18]构建了网络，并使用基于NVIDIA NCCL库的两块NVIDIA M40 GPU并行运算加速训练进程。在数据集UCF101和HMDB51上进行了实验。

UCF101[19]包含13,320个视频短片(共27 h时长)，分为101个动作类，每个类至少包含100个视频短片，每个动作类分为25组，每组包含4~7个视频短片。这些类可以划分为5种：人-物互动，身体动作，人-人互动，弹奏乐器和运动。

HMDB51[20]共有6766个视频短片，这些视频短片主要来自电影和网络视频，包含51个动作类，每类至少包含101个视频。此数据集是由至少两名观察者人工标注的。

网络输入　NFMML的输入为RGB帧和光流图。本文以30 bps速率采样RGB帧，使用基于GPU加速的OpenCV[21]库以TV-L1[22]算法提取相应的光流图。

超参数　本文使用批量随机梯度下降算法(mini-batch Stochastic Gradient Descent, SGD)来优化损失函数学习网络参数。由于受到GPU显存容量的限制，K设置为5，批量大小设置为64。如前所述，空域流使用在ImageNet[23]上预训练的模型。时域流使用在RGB帧上预训练的模型。空域流的初始学习率设置为0.0001，每迭代1000次下降至它的1/10，最大迭代次数为8000次。为了避免梯度爆炸，在训练过程中使用了梯度裁剪法，阈值设置为40。时域流的初始学习率设置为0.001，每迭代5000次下降至它的1/10，最大迭代次数设置为20 k，梯度裁剪阈值同样设置为40。空域和时域流的权重衰减率和冲量都设置为1×10^{-2}和0.9，两个流最后的全连接层都设置为1024维。

数据增量　当训练数据有限时，使用数据增量的方法能够有效提高分类精度。空域和时域流在训练阶段分别从4角和中心把输入图片裁剪为224×224大小，然后进行水平和垂直翻转，增加训练样本。而在测试阶段，不使用数据增量。

时域池化实现　实现时域平均池化有两种方法：使用Caffe中的slice layer和element wise layer实现、使用Caffe中的reshape layer和pooling layer层实现。这两种方法的分类性能相近，但后者运算速度较快，故本文选择后者实现方法。

评价指标　视频分类领域中常用的评价指标是平均精度均值(mean Average Precision, mAP)。mAP综合表征了查准率(precision)和查全率(recall)，

其值越大，表明算法分类性能越好。其定义为

$$\mathrm{mAP}=\frac{\sum_{n=1}^{N}\mathrm{AP}(n)}{N} \tag{10}$$

其中N是总类别数，$\mathrm{AP}(n)$是类别n的平均查准率(Average Precision, AP)[24]。

3.2 网络子结构和参数的影响

(1)时域池化有效性：NFMML使用时域池化聚合视频序列的帧特征，修复视频标签相对于帧分类较粗糙的问题。为了验证时域池化的有效性，本文进行了对比实验，见表1。从表1中可以看到，网络在使用平均池化进行特征聚合以后，相比于原始TSN[10]网络，分类精度有0.3%的提升。网络通过平均池化，在一定程度上能够得到视频特征的总体表征。

表 1 UCF101上时域池化的影响(%)

	原始TSN	TSN+时域池化
RGB	82.3[1)]	83.2
Optical Flow	83.6[1)]	82.9
RGB + Optical Flow	92.5[1)]	92.8

注：1)比原论文中的分类精度低。也许是因为批量大小较小等原因，本文未能复现原论文的实验结果。

(2)λ_1和λ_2的影响：NFMML的损失函数由分类损失和度量损失两部分组成。通过调整λ_1和λ_2，可以调整分类损失和度量损失在网络学习中的相对重要性。由于本文的主要任务是视频分类，并且相对于精细化分类任务，数据集样本之间语义距离的度量损失远大于分类损失。为了使网络能够更加关注分类任务，在训练过程中，λ_1始终设置为1，不断调整λ_2，并且使λ_2始终处于[0, 1)之间，实验结果如表2所示。

从表2中可以看到当λ_2为0.2时，网络的分类精度最好，在后续实验中，λ_2设置为0.2。还可以看到当λ_2为0.3, 0.4时，网络不收敛。事实上，当λ_2大于等于0.3时，网络都是不收敛的。在实验中，度量损失比分类损失大两个数量级，因此即使λ_2设置很小，度量学习对网络仍有较大影响。并且在实验中还发现度量学习中全连接层的初始化方式对网络也有较大影响：当初始化方式为“gaussian”算法时，度量损失计算错误(Nan)，而初始化方式为“xavier”算法时，度量损失正常。从中可以看到，超参数在深度学习起着重要作用，甚至能直接影响到网络是否能够成功训练。

表 2 当λ_1固定为1时，λ_2在数据集UCF101上的影响(%)

λ_2	0	0.1	0.2	0.3	0.4
mAP	92.7	93.1	93.8	不收敛	不收敛

(3)间隔分配函数的影响：本小节对比了3个间隔分配函数在数据集UCF101上的性能：(1) $\alpha_{\mathrm{r}}=D_{i,j}/D_{i,k}$, (2) $\alpha_{\mathrm{lg}}=\lg(D_{i,j}/D_{i,k})$, (3) $\alpha_{\exp}=\exp(D_{i,j}/D_{i,k})$，实验结果如表3所示。

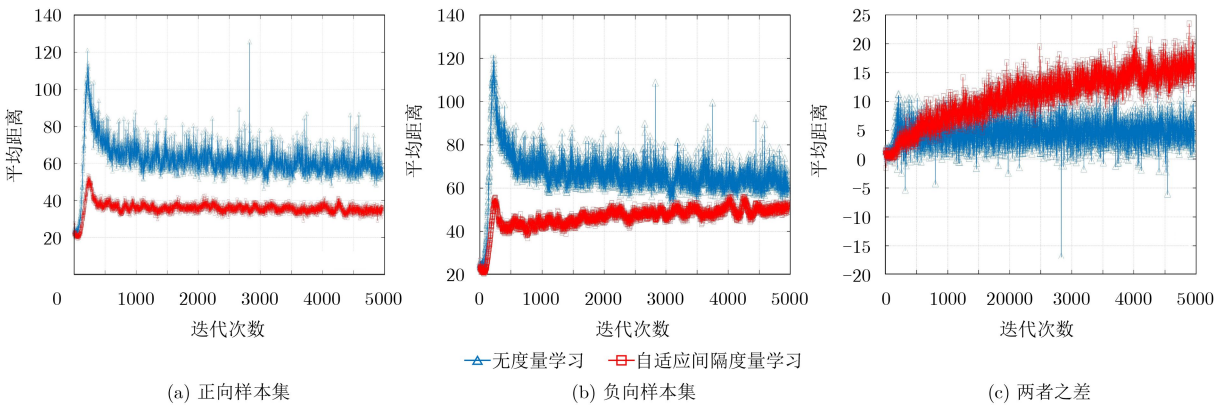
表 3 不同间隔分配函数在UCF101上的性能(%)

函数	α_{r}	α_{lg}	$\alpha_{\exp}$
mAP	92.9	91.9	93.8

从表3可以看到，当使用α_{lg}函数时，NFMML的分类性能比只有分类任务的网络(λ_2为0)时的性能差，而间隔分配函数$\alpha_{\exp}$的分类性能最好。在后续实验中，本文使用$\alpha_{\exp}$函数分配间隔。

为了更加直观地观察提出的深度度量学习方法对视频类内离散度和类间相似性的学习能力，在实验过程中分别记录了在UCF101 split 1上每个批量中正向样本集和负向样本集的平均距离随迭代次数的变化情况，如图2所示。

从图2(a)可以看出，随着训练地进行，每个批量中正向样本集的平均距离在减小，但每个迭代过程中，NFMML比原始方法的值小，这说明自适应



(a) 正向样本集 (b) 负向样本集 (c) 两者之差

图 2 UCF101 split 1上每个批量中样本之间的语义距离随迭代次数变化情况

间隔深度度量学习方法使正向样本集的平均距离更小，使同类样本更加聚集。而在图2(b)每个批量中负向样本集的平均距离随迭代次数变化图中，原有方法的曲线几乎不变甚至有下降趋势，而NFMML方法曲线一直呈上升趋势，这证明自适应间隔深度度量学习方法使负向样本集的平均距离变大，使异类样本相互分离。在图2(c)两者之差图中，NFMML方法曲线明显呈上升趋势，进一步证明了提出的自适应间隔的深度度量学习方法使同类样本聚集而异类样本相互分离。

(4)网络子结构的有效性：这一小节通过实验分析NFMML子结构对分类性能的影响，按照网络正向传播的方向依次添加子结构，实验结果如表4所示：

表 4　NFMML子结构在数据集UCF101上对分类性能的影响(%)

子结构	原始TSN	TSN+时域池化	TSN+时域池化+度量学习
mAP	92.5	92.8	93.8

从表4中可以看到，网络的分类性能左到右依次提高，特别是添加深度度量学习以后，分类性能提升了1%，这证明学习视频的类内离散度和类间相似性能够更好地理解视频。

(5)与现有方法对比：在验证过NFMML中每个子结构的有效性以后，本文在数据集UCF101和HMDB51上进行了实验，并与现有主流方法进行对比，结果如表5所示。

从表5可以看出，本文方法比原始TSN网络在相同实验条件下(本文方法批量大小比其小)在数据集UCF101和HMDB51上的分类精度分别高1.3%, 2.0%，实验结果表明：(1)提出的NFMML网络能够有效对视频建模；(2)提出的自适应间隔深度度量学习方法能够充分学习视频的类内离散度和类间相似性，并有效使同类聚集而异类相互分离。

表 5　与现有主流方法的分类精度对比(%)

方法	UCF101	HMDB51
DT + MVSV[25]	83.5	55.9
iDT + FV[26]	85.9	57.2
iDT + HSV[27]	87.9	61.1
MoFAP[28]	88.3	61.7
Two Stream[7]	88.0	59.4
FSTCN[29]	88.1	59.1
TDD + FV[30]	90.3	63.2
LTC[31]	91.7	64.8
TSN(2 modalities)	92.5	66.7
TS-LSTM	94.1	69.0
NFMML	**93.8**	**68.7**

普遍认为，网络训练及分类的计算复杂度主要与网络层次相关：网络层次越深，复杂度越高；相应地，其识别精度通常越高。在表5中可以看到，本文算法与其它浅层网络算法相比，其精度具有明显优势。而与TS-LSTM算法(该算法采用101层ResNet101网络以及4层LSTM网络)相比，本文算法分类精度略低(在数据集UCF101和HMDB51上分别低0.3%)，但本文网络深度远低于该算法。与本文改进前的同为32层网络的TSN算法相比，本文算法在数据集UCF101和HMDB51上精度分别提升1.3%和2%，在相同的硬件条件下，两种算法的训练时间同为75 h左右，证明本文算法在几乎相同的计算复杂度下，能有效地提高分类精度。

4　结束语

本文提出了一个基于多任务学习和深度度量学习的视频分类网络NFMML。该网络能够同时提取视频空域信息和长时时域信息，聚合帧特征，学习类内离散度和类间相似性和分类任务。提出一个自适应间隔的深度度量学习方法，根据样本对之间的语义距离，自适应分配间隔，使分类网络在训练阶段能够更加关注难分样本，从而提高分类精度。在数据集UCF101和HMDB51上进行了大量实验，验证了NFMML各个子结构的有效性，特别是验证了提出的自适应间隔的深度度量学习方法能够有效使同类样本聚集而异类样本相互分类。实验结果与现有主流方法进行了对比，NFMML有较好的分类精度。但在实验中发现本文提出的自适应间隔的深度度量学习方法对网络超参数较敏感，造成网络收敛性不稳定，下一步工作是增强该方法的鲁棒性和通用性。

参 考 文 献

[1] BRANSON S, VAN HORN G, PERONA P, *et al*. Improved bird species recognition using pose normalized deep convolutional nets[C]. Proceedings of the British Machine Vision Conference, Nottingham, British, 2014: 197–211. doi: 10.5244/C.28.87.

[2] ZHANG Ning, DONAHUE J, GIRSHICK R, *et al*. Part-based R-CNNs for fine-grained category detection[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014, 8689: 834–849. doi: 10.1007/978-3-319-10590-1_54.

[3] KRAUSE J, JIN Hailin, YANG Jianchao, *et al*. Fine-grained recognition without part annotations[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5546–5555.

doi: 10.1109/CVPR.2015.7299194.

[4] LIN Tsungyu, ROYCHOWDHURY A, and MAJI S. Bilinear CNN models for fine-grained visual recognition[C]. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1449–1457.

[5] CUI Yin, ZHOU Feng, LIN Yuanqing, *et al*. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1153–1162.

[6] JI Shuiwang, XU Wei, YANG Ming, *et al*. 3D convolutional neural networks for human action recognition[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, 35(1): 221–231. doi: 10.1109/TPAMI.2012.59.

[7] SIMONYAN K and ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. *Advances in Neural Information Processing Systems*, 2014, 1(4): 568–576.

[8] MA Chihyao, CHEN Minhung, KIRA Z, *et al*. TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition[OL]. arXiv preprint arXiv: 1703. 10667, 2017.

[9] IOFFE S and SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[OL]. arXiv preprint arXiv: 1502. 03167, 2015.

[10] WANG Limin, XIONG Yuanjun, WANG Zhe, *et al*. Temporal segment networks: Towards good practices for deep action recognition[J]. *ACM Transactions on Information Systems*, 2016, 22(1): 20–36. doi: 10.1007/978-3-319-46484-8_2.

[11] SONG Hyunoh, XIANG Yu, JEGELKA S, *et al*. Deep metric learning via lifted structured feature embedding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 4004–4012. doi: 10.1109/cvpr.2016.434.

[12] YI Dong, LEI Zhen, LIAO Shengcai, *et al*. Deep metric learning for person re-identification[C]. International Conference on Pattern Recognition, Stockholm, Sweden, 2014: 34–39.

[13] CHEN Xingyu, LAN Xuguang, LIANG Guoqiang, *et al*. Pose-and-illumination-invariant face representation via a triplet-loss trained deep reconstruction model[J]. *Multimedia Tools & Applications*, 2017(7): 1–16. doi: 10.1007/s11042-017-4782-y.

[14] BELL S and BALA K. Learning visual similarity for product design with convolutional neural networks[J]. *ACM Transactions on Graphics*, 2015, 34(4): 98–99. doi: 10.1145/2766959.

[15] SCHROFF F, KALENICHENKO D, and PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 815–823.

[16] NG Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, *et al*. Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 4694–4702. doi: 10.1109/cvpr.2015.7299101.

[17] MCLAUGHLIN N, RINCON J M D, and MILLER P. Recurrent convolutional network for video-based person re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1325–1334.

[18] JIA Yangqing, SHELHAMER E, DONAHUE J, *et al*. Caffe: Convolutional architecture for fast feature embedding[C]. ACM International Conference on Multimedia, Orlando, USA, 2014: 675–678.

[19] SOOMRO K, ZAMIR A R, and SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[OL]. arXiv preprint arXiv: 1212. 0402, 2012.

[20] KUEHNE H, JHUANG H, STIEFELHAGEN R, *et al*. HMDB51: A Large Video Database for Human Motion Recognition[M]. Heidelberg, Berlin: Springer, 2013: 2556–2563.

[21] BRADSKI G. The opencv library[J]. *Doctor Dobbs Journal*, 2000, 25(11): 384–386.

[22] ZACH C, POCK T, and BISCHOF H. A Duality Based Approach for Realtime TV-L1 Optical Flow[M]. Heidelberg, Berlin: Springer, 2007: 214–223.

[23] DENG Jia, DONG Wei, SOCHER R, *et al*. ImageNet: A large-scale hierarchical image database[C]. Proceedings of Computer Vision and Pattern Recognition, Miami, USA, 2009: 248–255.

[24] EVERINGHAM M, GOOL LV, WILLIAMS CKI, *et al*. The pascal visual object classes (VOC) challenge[J]. *International Journal of Computer Vision*, 2010, 88(2): 303–338. doi: 10.1007/s11263-009-0275-4.

[25] CAI Zhuowei, WANG Limin, PENG Xiaojiang, *et al*. Multi-view super vector for action recognition[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 596–603.

[26] PENG Xiaojiang, WANG Limin, WANG Xingxing, *et al*. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice[J]. *Computer Vision & Image Understanding*, 2016, 150(C): 109–125. doi: 10.1016/j.cviu.2016.03.013.

[27] WANG Heng and SCHMID C. Lear-inria submission for the thumos workshop[C]. ICCV Workshop on Action Recognition with a Large Number of Classes, Sydney, Australia, 2013: 39–47.

[28] WANG Limin, QIAO Yu, and TANG Xiaoou. MoFAP: A multi-level representation for action recognition[J]. *International Journal of Computer Vision*, 2016, 119(3): 254–271. doi: 10.1007/s11263-015-0859-0.

[29] SUN Lin, JIA Kui, YEUNG D Y, *et al*. Human action recognition using factorized spatio-temporal convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4597–4605.

[30] WANG Limin, QIAO Yu, and TANG Xiaoou. Action recognition with trajectory-pooled deep-convolutional descriptors[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 2015: 4305–4314.

[31] VAROL G, LAPTEV I, and SCHMID C. Long-term temporal convolutions for action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(6): 1510–1517. doi: 10.1109/tpami.2017.2712608.

智洪欣：男，1987年生，博士生，研究方向为计算机视觉.

于洪涛：男，1970年生，研究员，研究方向为大数据和计算机视觉.

李邵梅：女，1982年生，讲师，研究方向为大数据和计算机视觉.

高　超：男，1982年生，讲师，研究方向为大数据和计算机视觉.

王艳川：男，1987年生，硕士生，研究方向为计算机视觉.