计算机科学
COMPUTER SCIENCE

基于时序信息对齐的连续手语跨模态知识蒸馏

肖正业，林世铨，万修安，方昱春，倪兰

引用本文

肖正业，林世铨，万修安，方昱春，倪兰. 基于时序信息对齐的连续手语跨模态知识蒸馏[J]. 计算机科学, 2022, 49(11): 156-162.

XIAO Zheng-ye, LIN Shi-quan, WAN Xiu-an, FANGYu-chun, NI Lan. Temporal Relation Guided Knowledge Distillation for Continuous Sign Language Recognition[J]. Computer Science, 2022, 49(11): 156-162.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

多层注意力机制融合的序列到序列中国连续手语识别和翻译

Sequence-to-Sequence Chinese Continuous Sign Language Recognition and Translation with Multi- layer Attention Mechanism Fusion

计算机科学, 2022, 49(9): 155-161. https://doi.org/10.11896/jsjkx.210800026

基于多时间尺度时空图网络的交通流量预测模型

Multi-time Scale Spatial-Temporal Graph Neural Network for Traffic Flow Prediction

计算机科学, 2022, 49(8): 40-48. https://doi.org/10.11896/jsjkx.220100188

基于图卷积网络的专利摘要自动生成研究

Automatic Generation of Patent Summarization Based on Graph Convolution Network

计算机科学, 2022, 49(6A): 172-177. https://doi.org/10.11896/jsjkx.210400117

基于 YOLOv4 的目标检测知识蒸馏算法研究

Study on Knowledge Distillation of Target Detection Algorithm Based on YOLOv4

计算机科学, 2022, 49(6A): 337-344. https://doi.org/10.11896/jsjkx.210600204

基于无标签知识蒸馏的人脸识别模型的压缩算法

Compression Algorithm of Face Recognition Model Based on Unlabeled Knowledge Distillation

计算机科学, 2022, 49(6): 245-253. https://doi.org/10.11896/jsjkx.210400023

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600036

基于时序信息对齐的连续手语跨模态知识蒸馏

肖正业[1]　林世铨[1]　万修安[1]　方昱春[1]　倪　兰[2]
1 上海大学计算机工程与科学学院　上海 200444
2 上海大学文学院　上海 200444
(xiaozy@shu.edu.cn)

摘　要　近年来,连续手语识别的研究工作主要围绕RGB模态的数据展开,并且在现实场景数据集和实验室采集数据集上都取得了显著进展。然而,RGB模态的处理对设备计算能力具有很高的要求,而骨骼关键点模态则由于输入数据复杂度相对低,因此处理速度更快,只是在识别性能上弱于RGB模态。为了综合两种方法的优点,文中提出了一种基于时序关联信息对齐的跨模态知识蒸馏方法(Temporally Related Knowledge Distillation,TRKD)。该方法使用RGB模态的神经网络作为教师网络来指导使用骨骼关键点模态的学生网络,以快速准确地实现连续手语识别。由于教师网络对手语语境的理解能力十分值得学生网络学习,因此提出了具有先验信息以及自适应学习方法的图卷积网络来提取两类模态中的时序关联特征,并通过特征对齐来实现教学。在特征对齐过程中,在教师网络中引入可学习参数会导致教师提供的监督信息丢失。为了解决这个问题,所提出的TRKD方法引入了自监督学习中的对比学习来提供监督信息,从而实现了教师网络与学生网络在时序关联特征上的对齐。文中在Phoenix-2014手语数据集上组织了多项蒸馏任务,以验证所提方法的有效性。

关键词:知识蒸馏;图卷积网络;手语识别
中图法分类号　TP311

Temporal Relation Guided Knowledge Distillation for Continuous Sign Language Recognition

XIAO Zheng-ye[1],LIN Shi-quan[1],WAN Xiu-an[1],FANG Yu-chun[1] and NI Lan[2]
1 School of Computer Engineering and Science,Shanghai University,Shanghai 200444,China
2 College of Liberal Arts,Shanghai University,Shanghai 200444,China

Abstract　Previous researches in continuous sign language recognition mainly focus on the RGB modality and achieve remarkable performance on real-world and laboratory datasets,but they usually require high computation intensity. On the other hand,the skeleton is a modality with small input data and fast computation speed,but poor at the real-world datasets. This paper proposes a cross-modal knowledge distillation method named temporally related knowledge distillation(TRKD) to alleviate the contradiction between RGB and skeleton modality in performance and calculation speed. TRKD utilizes the RGB modality network as a teacher to guide the skeleton modality network for fast and accurate implementation. We notice that the teacher's understanding of sign language context is worth learning by student. It proposes to employ the graph convolutional network(GCN) to learn and align the temporally related features of teacher networks and student networks to achieve this goal. Moreover,since the supervised information from the teacher network is not available for traditional loss functions due to the learnable parameters of GCN in the teacher network,we introduce contrastive learning to provide self-supervised information. Multiple ablation experiments on the Phoenix-2014 dataset demonstrate the effectiveness of the proposed method.

Keywords　Knowledge distillation,Graph convolutional network,Sign language recognition

1　引言

随着社会进步,越来越多的聋人参与到了社交活动中。自动手语识别研究可以拓展手语理解,帮助聋人获得更多的交流方式和更广泛地参与社会生活。连续手语识别是将手语视频识别为手语单词序列的技术,对手语翻译的研究具有重要的推进作用,因此连续手语识别在促进手语理解中具有重要的研究价值。

到稿日期:2022-06-03　返修日期:2022-08-02
基金项目:国家自然科学基金(61976132,61991411,U1811461);上海自然科学基金(19ZR1419200);上海智能计算系统工程研究中心(19DZ2252600)
This work was supported by the National Natural Science Foundation of China(61976132,61991411,U1811461),Natural Science Foundation of Shanghai,China(19ZR1419200)and Shanghai Engineering Research Center of Intelligent Computing System(19DZ2252600).
通信作者:方昱春(ycfang@shu.edu.cn)

目前，对连续手语识别的研究主要侧重于计算机视觉中最常用的 RGB 模态[1-2]。虽然基于 RGB 模态的连续手语识别研究在现实场景数据集和实验室采集数据集上都取得了显著的进展，但是这些方法往往需要消耗大量的计算资源。而骨骼关键点模态由于输入数据量小、计算复杂度低，因此处理速度快。然而，现实场景中的手语视频通常分辨率较低，这使得从中估计出的骨骼关键点存在精度低的问题，因此使用骨骼关键点的识别方法很难在现实场景中达到与使用 RGB 模态的方法一样令人满意的效果。在真实场景数据集上，RGB 模态模型具有识别效果更好的优点，而骨骼关键点模态模型则具有推理速度更快的优点。为了综合这两种模型的优点，我们需要提出一种能够调和这两类模态的识别效果与推理速度的算法，从而在具有良好推理速度的同时取得优秀的识别效果。

随着相机阵列、Kinect 以及各种快速姿态估计算法的发展，实时且廉价地获取骨骼关键点模态已经成为可能。在实际应用中，在训练阶段通过视频 RGB 模态模型对骨骼关键点模态模型进行辅助学习以提升识别效果，在测试阶段使用骨骼关键点模态以确保快速识别，从而调和多模态优势，这是一种具有良好前景的策略。本文提出通过跨模态知识蒸馏实现连续手语识别的多模态间识别精度与推理速度调衡。我们将基于 RGB 模态的神经网络作为教师网络，将基于骨骼关键点模态的神经网络作为学生网络，从而在保持模型轻量计算的同时提升识别效果。

跨模态的知识蒸馏方法虽然在图像和动作识别的研究中已有相关研究，但在连续手语识别任务中尚是一种有待尝试的新方法。其中，跨模态知识蒸馏主要是通过特征或概率分布对齐的方式来实现的[3-4]。在连续手语识别问题中，上下文信息对识别整个序列起到了至关重要的作用，同时 RGB 模态由于具有更为可靠的输入数据，对手语上下文信息的理解更加准确。因此，骨骼关键点模态模型对于手语语境信息理解能力的学习是一个特别需要关注的问题。由此，本文提出了基于时序关联信息对齐的跨模态知识蒸馏方法，使学生网络具备更好的语境信息理解能力。TRKD 使用图卷积网络(Graph Convolution Network，GCN)[5]来从教师网络和学生网络的输出中提取时序关联特征，以表征手语语境信息，同时适用于教师网络和学生网络异构的情况。本文提出的跨模态知识蒸馏 GCN 结合自适应邻接矩阵和先验邻接矩阵，能够从全局以及局部的视角学习到不同的手语语境信息。

教师网络中的图卷积网络是需要训练的，因此存在可学习参数，这导致学习到的时序关联特征不能直接作为监督信息用于计算 KL 散度或者余弦相似度来对齐特征。使用 KL 散度进行训练会导致教师网络的时序关联特征变为非数(NaN)，而余弦相似度会使教师与学生网络的图卷积输出的特征几乎一样，来满足余弦夹角越小越好的要求。为了解决这一问题，本文引入了对比学习方法来实现时序关联特征的对齐。对比学习是自监督学习研究领域常用的方法[6]，通常是将来自同一类别的样本作为正例、来自不同类别的样本作为负例进行对比学习。在连续手语的问题中，同一时间步输出的特征之间互为正例，不同时间步输出的特征之间互为负例，以此通过对比学习损失来对齐时序关联特征。

本文的主要贡献如下：

(1)提出了跨模态知识蒸馏方法 TRKD，用于指导基于骨骼关键点的连续手语识别学生网络从基于 RGB 模态的教师网络中学习到语境信息理解能力，丰富了知识蒸馏的学习内容。

(2)引入了一个先验信息与自适应方法结合的 GCN 来提取时序关联特征，让学习到的特征既具有先验知识提供的局部关联信息，又具有自适应学习到的全局关联信息。

(3)引入自监督的对比学习方法来解决 TRKD 中的时序关联特征对齐问题。学生网络通过与教师网络进行时序关联特征对齐来学习语境信息的理解能力，从而提升自己的识别性能。

2 相关工作

2.1 连续手语识别

目前对连续手语识别的研究工作主要基于 RGB 模态。Koller 等[2]在卷积神经网络(Convolutional Neural Network，CNN)提取特征后使用隐马尔可夫模型(Hidden Markov Model，HMM)对序列进行建模来学习语境信息。Pu 等[7]提出利用动态时间规整算法来实现序列对齐并交互地训练 3D 卷积神经网络(3DCNN)。Cui 等[8]利用 CTC 损失来训练一个 CNN-LSTM 模型，并用概率优化算法来匹配最优伪标签。Zhou 等[9]设计了一个手语理解要素间与手语理解要素内的双分支交互式训练的时空多要素网络。基于骨骼关键点模态的连续手语识别方法大多采用时空图卷积网络模型(ST-GCNs)来提取特征。Wang 等[10]将人体骨骼图分为多个子图，并通过 ST-GCNs 来分步提取人体骨骼关键点的特征。

2.2 跨模态知识蒸馏

目前已经有许多工作在图像和动作识别的跨模态知识蒸馏问题上展开研究。Gracia 等[11]提出利用幻觉网络架起深度模态和 RGB 模态间的桥梁，从而提升基于深度模态模型的性能。Tian 等[4]认为在跨模态知识蒸馏中对特征进行蒸馏尤为重要，他们使用互信息作为替代并利用对比学习损失来提升损失的下界，以此取得了显著成效。近期，跨模态知识蒸馏方法在视频多动作识别问题中也受到了关注。Dai 等[12]认为，在使用跨模态知识蒸馏方法进行多动作识别时，动作之间的关系和动作边界是重要的学习内容，并提出通过对比学习方法和协方差矩阵来在特征层面上限制时序关联关系。

3 TRKD 方法

本文提出的 TRKD 方法的总体结构如图 1 所示。我们分别将 RGB 模态以及骨骼关键点模态作为教师网络与学生网络的输入来提取得到各自的特征序列；然后时序关联特征关联模块(Temporal Feature Alignment)使用 GCN 来分别对教师网络和学生网络输出的特征序列进行时序关联特征学习，并通过关联对齐损失 $\mathscr{L}_{\text{rel}}$ 来将其对齐。另外，TRKD 还通过优化传统的概率分布对齐损失 $\mathscr{L}_{\text{kd}}$，输出特征对齐损失 $\mathscr{L}_{\text{repr}}$

以及学生网络自己的分类损失$\mathcal{L}_{ce}$，并将其作为辅助损失来提升连续手语识别的效果。在训练过程中，如图1中反向传播的路径所示，教师网络的图卷积模块以及学生网络的所有模块的参数会参与到训练中。

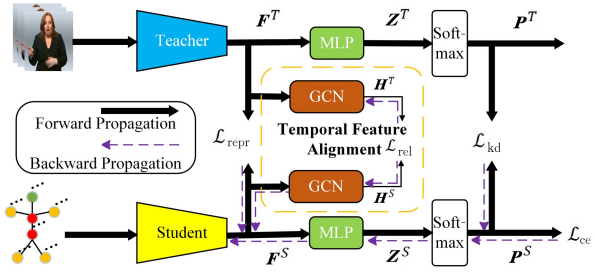


图1 TRKD跨模态知识蒸馏连续手语识别模型框架图

Fig. 1 Schematic diagram of TRKD model for CSLR

3.1 时序关联特征

若要使学生网络的语境信息理解能力更接近教师网络，则需要学习教师网络和学生网络的时序关联特征并将其对齐。不失一般性地将教师网络表示为$G_T(\cdot)$，将学生网络表示为$G_S(\cdot)$。给定教师网络输入模态的序列样本$\boldsymbol{X}^T=\{\boldsymbol{X}_n^T \mid 1\leqslant n\leqslant N\}$以及学生网络输入模态的序列样本$\boldsymbol{X}^S=\{\boldsymbol{X}_n^S \mid 1\leqslant n\leqslant N\}$，其中$N$是序列长度。由于教师网络和学生网络的输入本质是同一段视频中获取的不同模态，因此多模态模型的输入信息具有相同的长度。教师网络以及学生网络对输入样本进行提取特征的过程如式(1)所示：

$$\boldsymbol{F}^T=G_T(\boldsymbol{X}^T),\boldsymbol{F}^S=G_S(\boldsymbol{X}^S) \tag{1}$$

其中，$\boldsymbol{F}^T=\{\boldsymbol{f}_n^T\in\mathbb{R}^{1\times C} \mid 1\leqslant n\leqslant N'\}$是教师网络提取到的特征序列，$\boldsymbol{F}^S=\{\boldsymbol{f}^S\in\mathbb{R}^{1\times C} \mid 1\leqslant n\leqslant N'\}$是学生网络提取到的特征序列，$N'$为特征序列的长度。

对于每一帧的时序关联特征，它应该包含与该帧有语义关联的其他帧的信息以表达其手语上下文的信息。为了得到这样的时序关联特征，可以将序列的每一帧视为一个图节点，使用GCN去建立帧与帧间的关系并学习时序关联特征。在帧与帧间关系，即图节点间的关联性的建模过程中，由于每一个手语视频的每一帧与其他帧的关系是不可简单描述的，并且会随着语义的变化以及手语者动作的变化出现差别，因此我们提出通过自适应地构建邻接矩阵来进行表达。

我们使用$\boldsymbol{A}^r\in\mathbb{R}^{N'\times N'}$，$r\in\{T,S\}$来表达构建的教师网络和学生网络的邻接矩阵，$A_{ij}^r$表示第$i$帧和第$j$帧之间的关联性。$A_{ij}^r$的值通过将$\boldsymbol{f}_i^r$与$\boldsymbol{f}_j^r$拼接后经过多层感知机(Multiple Layer Perceptron，MLP)计算得到，如式(2)所示：

$$A_{ij}^r=\sigma((\boldsymbol{f}_i^r \,\|\, \boldsymbol{f}_j^r)\boldsymbol{W}_0^r)\boldsymbol{W}_1^r \tag{2}$$

其中，$\|$表示拼接操作，$\sigma(\cdot)$表示ReLU激活函数，$\boldsymbol{W}_0^r\in\mathbb{R}^{2C\times C}$和$\boldsymbol{W}_1^r\in\mathbb{R}^{C\times 1}$是可学习参数。

在连续手语识别问题中，迭代训练是一种为手语视频各帧提供伪标签的有效方法。由于每一帧必定与同属一个类别标签的帧有关联，因此根据伪标签生成一个先验邻接关系矩阵$\boldsymbol{A}^p$来引入先验信息。假设$\hat{y}_1^{(1)},\hat{y}_2^{(1)},\hat{y}_3^{(2)},\cdots,\hat{y}_n^{(m)},\cdots,\hat{y}_{N'-2}^{(M-1)},\hat{y}_{N'-1}^{(M)},\hat{y}_{N'}^{(M)}$是具有$M$个词的伪标签序列，那么$\boldsymbol{A}^p$的计算结果如式(3)所示：

$$A_{ij}^p=\begin{bmatrix}1&1&0&&0&0&0\\1&1&0&\cdots&0&0&0\\0&0&1&&0&0&0\\&\vdots&&\ddots&&\vdots&\\0&0&0&&1&0&0\\0&0&0&\cdots&0&1&1\\0&0&0&&0&1&1\end{bmatrix} \tag{3}$$

即具有相同标签的帧与帧之间的关联性为1。

将$\boldsymbol{A}^r$和$\boldsymbol{A}^p$相加后得到$\tilde{\boldsymbol{A}}^r=\eta\boldsymbol{A}^r+\boldsymbol{A}^p$，其中$\eta$是缩放因子，用于平衡自适应邻接关系$\boldsymbol{A}^r$相对于先验邻接关系$\boldsymbol{A}^p$的权重。因此，教师网络和学生网络的邻接矩阵逐行使用softmax归一化后可表示为$\bar{\boldsymbol{A}}^r=\{\bar{A}_{ij}^r\}_{N'\times N'}$，其中$\bar{A}_{ij}^r$的计算过程如式(4)所示：

$$\bar{A}_{ij}^r=\frac{e^{\tilde{A}_{ij}^r}}{\sum_{k=1}^{N'}e^{\tilde{A}_{ik}^r}} \tag{4}$$

根据所构造的邻接矩阵，两层GCN提取时序关联特征的计算过程如式(5)所示：

$$\boldsymbol{H}_l^r=\bar{\boldsymbol{A}}^r\boldsymbol{H}_{l-1}^r\boldsymbol{W}_l^r,l=\{1,2\} \tag{5}$$

其中，$\boldsymbol{H}_l^r\in\mathbb{R}^{N'\times C}$是由GCN更新的特征，并且$\boldsymbol{H}_0^r=\boldsymbol{F}^r$，$\boldsymbol{W}_l^r\in\mathbb{R}^{C\times C}$是可学习参数。记GCN提取的时序关联特征为$\boldsymbol{H}^r$。值得注意的是，学习时序关联特征的过程仅仅在训练阶段，不会对推理阶段的计算速度造成影响。

3.2 关联对齐损失

由于教师网络部分的GCN是可学习的，因此教师网络输出的时序关联特征难以直接作为学生网络时序关联特征的监督信息。对此，我们借鉴了对比学习的思想来解决这个问题。

相同的手语单词出现在句子的不同位置往往具有不同的语义。因此，我们做出了最好的假设，即将同一时刻的帧作为正样本对，而将不同时刻的帧作为负样本对。记$\boldsymbol{h}_i^r$为$\boldsymbol{H}^r$的第i行，对于任意$j\neq i$，我们认为$\boldsymbol{h}_j^T$是$\boldsymbol{h}_j^S$的负样本，而$\boldsymbol{h}_i^T$是$\boldsymbol{h}_i^S$的正样本。由于正样本对之间有关联，我们定义正样本对的采样概率为联合概率分布$p(\boldsymbol{H}^T,\boldsymbol{H}^S)$。而负样本对之间是独立的，因此我们定义其采样概率为边缘分布的乘积$p(\boldsymbol{H}^T)p(\boldsymbol{H}^S)$。定义一个隐变量$D$来指示当前特征对是正样本对还是负样本对，$D=1$表示正样本对，$D=0$表示负样本对，我们可以定义正负样本对的条件概率分布为q，如式(6)所示：

$$\begin{aligned}q(\boldsymbol{H}^T,\boldsymbol{H}^S \mid D=1)&=p(\boldsymbol{H}^T,\boldsymbol{H}^S)\\q(\boldsymbol{H}^T,\boldsymbol{H}^S \mid D=0)&=p(\boldsymbol{H}^T)p(\boldsymbol{H}^S)\end{aligned} \tag{6}$$

此外，由于特征序列长度为N'，并且每个$\boldsymbol{h}_i^S$只有一个对应的正样本以及$N'-1$个负样本。因此，隐变量D的先验概率可以直接计算，如式(7)所示：

$$q(D=1)=\frac{1}{N'},q(D=0)=\frac{N'-1}{N'} \tag{7}$$

由于类条件概率密度和先验概率都已知，因此$D=1$的后验概率的计算过程如式(8)所示：

$$q(D=1 \mid \boldsymbol{H}^T,\boldsymbol{H}^S)=\frac{q(\boldsymbol{H}^T,\boldsymbol{H}^S \mid D=1)q(D=1)}{q(\boldsymbol{H}^T,\boldsymbol{H}^S)} \tag{8}$$

其中，$q(\boldsymbol{H}^T,\boldsymbol{H}^S)$可以通过对$q(\boldsymbol{H}^T,\boldsymbol{H}^S,D)$进行边缘分布求和

得到。对式(8)两边同时求对数，并进一步展开后的结果如式(9)所示：

$$\begin{aligned}&\log q(D=1|\boldsymbol{H}^T,\boldsymbol{H}^S)\\&=\log\frac{p(\boldsymbol{H}^T,\boldsymbol{H}^S)}{p(\boldsymbol{H}^T,\boldsymbol{H}^S)+(N'-1)p(\boldsymbol{H}^T)p(\boldsymbol{H}^S)}\\&=-\log\left(1+(N'-1)\frac{p(\boldsymbol{H}^T)p(\boldsymbol{H}^S)}{p(\boldsymbol{H}^T,\boldsymbol{H}^S)}\right)\end{aligned}\tag{9}$$

对式(9)进行不等式缩放后，对两边求期望的结果如式(10)所示：

$$\begin{aligned}&\mathbb{E}[\log q(D=1|\boldsymbol{H}^T,\boldsymbol{H}^S)]\\&\leqslant-\log(N'-1)+\mathbb{E}\left[\log\frac{p(\boldsymbol{H}^T,\boldsymbol{H}^S)}{p(\boldsymbol{H}^T)p(\boldsymbol{H}^S)}\right]\\&=-\log(N'-1)+I(\boldsymbol{H}^T;\boldsymbol{H}^S)\end{aligned}\tag{10}$$

其中，$I(\boldsymbol{H}^T;\boldsymbol{H}^S)$表示$\boldsymbol{H}^T$和$\boldsymbol{H}^S$的互信息。由此我们可以得到与$I(\boldsymbol{H}^T;\boldsymbol{H}^S)$相关的最终优化的目标函数，如式(11)所示：

$$I(\boldsymbol{H}^T;\boldsymbol{H}^S)\geqslant\log(N'-1)+\mathbb{E}[\log q(D=1|\boldsymbol{H}^T,\boldsymbol{H}^S)]\tag{11}$$

InfoNCE[6]已经被证明是估计概率分布$q(D=1|\boldsymbol{H}^T,\boldsymbol{H}^S)$的一个紧下界，因此我们使用 InfoNCE 作为评价教师网络和学生网络的时序关联特征之间的相似度的损失函数，如式(12)所示：

$$\mathcal{L}_{\text{rel}}=\sum_{i=1}^{N'}\log\frac{\mathrm{e}^{d(\boldsymbol{h}_i^T,\boldsymbol{h}_i^S)}}{\sum_{j=1}^{N'}\mathrm{e}^{d(\boldsymbol{h}_i^T,\boldsymbol{h}_j^S)}}\tag{12}$$

其中，$d(\cdot)$是距离测度，这里使用余弦距离。通过对比学习，学生网络的时序关联特征将尽可能地靠近教师网络中对应帧的时序关联特征，并尽可能地远离其他帧的特征。同时，教师网络也受到反作用，需要提取具有特点的特征以供学生网络学习。

3.3 特征及概率对齐

在传统的知识蒸馏方法中，教师网络和学生网络提取的特征和最终分类决策的概率分布都会被对齐，以蒸馏出特征表示和决策能力。本文也考虑了这两个方面的对齐。

在特征对齐方面，由于教师网络的特征$\boldsymbol{F}^T$和学生网络的特征$\boldsymbol{F}^S$与$\boldsymbol{H}^T$和$\boldsymbol{H}^S$具有相同的维度，并且都是序列特征，因此我们同样使用 InfoNCE 作为评价$\boldsymbol{F}^T$和$\boldsymbol{F}^S$差别的损失函数，如式(13)所示：

$$\mathcal{L}_{\text{repr}}=\sum_{i=1}^{N'}\log\frac{\mathrm{e}^{d(\boldsymbol{f}_i^T,\boldsymbol{f}_i^S)}}{\sum_{j=1}^{N'}\mathrm{e}^{d(\boldsymbol{f}_i^T,\boldsymbol{f}_j^S)}}\tag{13}$$

在概率分布对齐方面，我们使用由 Hinton 等[13]提出的方法。我们首先通过多层感知机(MLP)将$\boldsymbol{F}^T$以及$\boldsymbol{F}^S$变换为 logit 概率$\boldsymbol{Z}^T$以及$\boldsymbol{Z}^S$，然后通过 Softmax 函数得到最终的教师网络和学生网络的分类概率分布，分别记为$\boldsymbol{P}^T$和$\boldsymbol{P}^S$。我们使用 KL 散度来度量概率分布$\boldsymbol{P}^T$和$\boldsymbol{P}^S$间的分布对齐程度，如式(14)所示：

$$\mathcal{L}_{\text{kd}}=KL(\boldsymbol{P}^T,\boldsymbol{P}^S)\tag{14}$$

伪标签可以作为学生网络的监督信息，因此我们使用交叉熵损失来作为学生网络的分类损失。我们将以上的 3 个损失函数与交叉熵损失求和，从而得到本文方法的最终损失函数，如式(15)所示：

$$\mathcal{L}=\mathcal{L}_{\text{ce}}+\alpha\mathcal{L}_{\text{kd}}+\beta\mathcal{L}_{\text{repr}}+\gamma\mathcal{L}_{\text{rel}}\tag{15}$$

其中，$\mathcal{L}_{\text{ce}}$是交叉熵损失，α,β,γ为权重超参数。

4 实验

4.1 数据集与评估方法

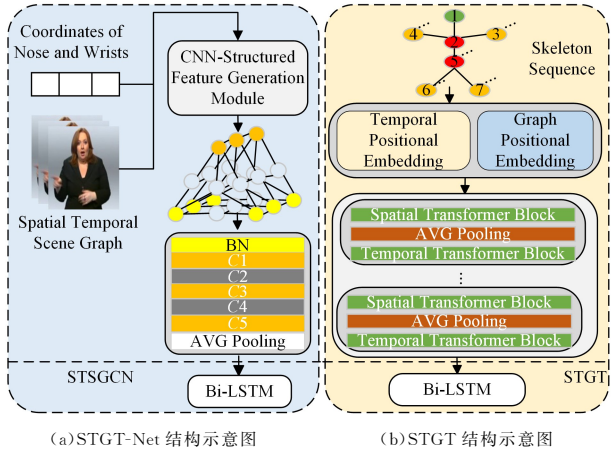
我们在数据集 RWTH-PHOENIX-Weather 2014(PHOENIX-2014)[14]上评估了所提出的 TRKD 方法。PHOENIX-2014 是一个德国手语数据集，被连续手语识别研究领域广泛采用，同时也是连续手语识别任务的基准测试集。这个数据集包含 9 个手语天气播报员的真实天气预报视频，共有 5 672 段视频作为训练集，540 段视频作为验证集以及 629 段视频作为测试集。数据集包含 1 295 个手语词汇，共有 1 232 个词汇出现在训练集中。

词错率(Word Error Rate，WER)是手语识别效果评估常用的指标，可以度量两个序列的相似性。词错率是两个序列相互转化所需要的替换、删除和插入元素的次数，其计算式如式(16)所示：

$$WER=\frac{\text{插入次数}+\text{删除次数}+\text{替换次数}}{\text{真实句子中的单词个数}}\tag{16}$$

4.2 教师模型与学生模型

为了体现 TRKD 方法的优越性，我们使用两种教师网络和两种学生网络来评估所提出的 TRKD 方法。使用的教师网络有 2DCNN＋1DCNN 模型和时空场景图卷积网络(STSG-Net)，用于提取 RGB 帧序列的特征。其中 2DCNN＋1DCNN 是连续手语识别研究领域广泛使用的网络结构，使用 ResNet-18 作为骨干网络来提取空间特征，再使用时序卷积来提取时序特征。STSG-Net 则是结合空间局部特征和时序关系构建出的时空场景图。在 STSG-Net 中，每一帧经过卷积提取出全局特征之后，裁剪出手部和头部等关键部位的特征以构建时空场景图的节点，从而构建出时空场景图来表达手语序列的时空关系。最后使用时空图卷积网络(ST-GCN)[15]来提取时空特征。STSG-Net 的结构如图 2(a)所示。



(a)STGT-Net 结构示意图　(b)STGT 结构示意图

图 2 针对 RGB 模态与骨骼关键点模态分别设计的模型结构示意图

Fig. 2 Schematics of the models for RGB data and skeletons data

我们使用 ST-GCN 和时空图 Transformer(STGT)作为学生网络来提取骨骼关键点序列的特征。ST-GCN 是基于骨骼关键点模态的识别问题中常用的模型，而 STGT 是一种类 Transformer 结构的网络，将时序位置编码和图位置编码分别输入到空间 Transformer 模块和时序 Transformer 模块中

提取特征。STGT 的整体模型如图 2(b)所示。

使用上述模型的组合，我们构造出了 4 组蒸馏实验，教师→学生网络分别为 STSG-Net→STGT，2DCNN＋1DCNN→STGT，STSG-Net→ST-GCN 以及 2DCNN＋1DCNN→ST-GCN。

4.3 实验设置

骨骼关键点模态的数据都使用 OpenPose[16] 对 Phoenix-2014 数据集的视频进行姿态估计生成。在实验中，通过将自适应学习邻接矩阵的权重 η 设为 10 来增强其贡献。在所有实验的训练过程中我们都使用 Adam 优化器，并把初始学习率设置为 0.0003。一共迭代 40 轮，学习率在第 10 轮迭代和第 20 轮迭代衰减为初始学习率的 1/10 和 1/100。此外，我们还设置了正则化参数为 0.0001。蒸馏任务中所有网络的功能都是特征提取。若要完成完整识别任务，还需要把提取的特征输入到一个双向长短期记忆神经网络(BiLSTM)中，并通过 CTC 损失优化。BiLSTM 最终输出手语单词序列，从而能够通过词错率的评价指标来评估算法的有效性。BiLSTM 采用的是 Adam 优化器，学习率设置为 5×10^{-4}，同样在第 10 轮迭代和第 20 轮迭代衰减为原来的 1/10，正则化参数为 1×10^{-4}，共迭代 40 轮。

教师网络通过预训练得到。在学生网络训练期间，仅更新我们提出的图卷积网络的参数，而不涉及教师网络本身的参数。预训练的方式是先通过迭代优化的策略[1]得到伪标签，然后通过伪标签对教师网络进行训练。

4.4 消融实验

我们对自适应邻接矩阵以及先验邻接矩阵进行了消融实验，以验证在图卷积网络中这两种邻接矩阵的必要性。另外，还对式(15)中的 3 个超参数 α,β,γ 进行了消融实验，以分析各个损失的贡献。消融实验在 STSG-Net→STGT 蒸馏任务上进行组织。

4.4.1 邻接矩阵分析

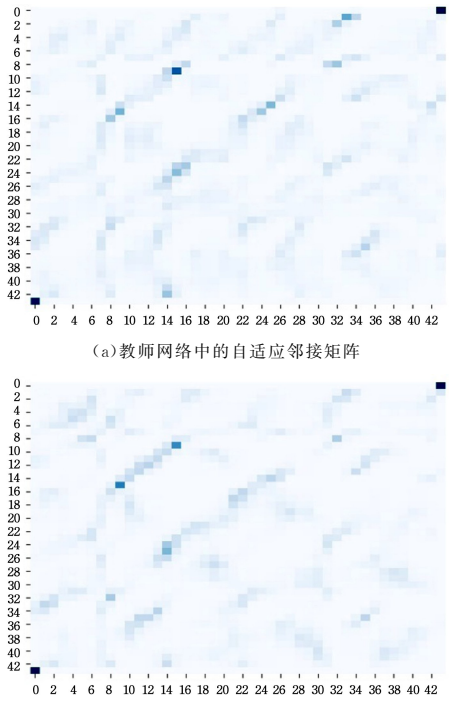
关于邻接矩阵的消融实验结果如表 1 所列。只使用自适应邻接关系矩阵或先验邻接关系矩阵的结果相近，而两者结合的邻接矩阵则达到了最好的效果。由此可见，这两个邻接关系矩阵具有各自的效果并且可以互补。先验邻接矩阵提供了局部的时序关联信息，而自适应邻接矩阵则更加关注全局关联性，将上下文的关联性表现出来。两者相辅相成并在融合后取得了更好的效果。图 3 给出了自适应邻接矩阵的可视化效果，其中图 3(a)和图 3(b)分别对应教师和学生网络各自对应的自适应邻接矩阵。可以看到，这两种模态下都捕捉到了超出先验的时序关联信息，自适应邻接矩阵的结合增强了语境信息的理解能力，从而得到了更好的识别效果。

表 1 邻接矩阵消融实验的结果

Table 1 Results of ablative experiments on adjacency matrices

(单位：%)

Methods	Pseudo-label Accuracy	Validation WER	Test WER
$\boldsymbol{A}^p$	54.9	37.3	35.9
$\boldsymbol{A}^r$	54.7	37.1	36.1
$\eta\boldsymbol{A}^r+\boldsymbol{A}^p$	**56.9**	**35.8**	**34.9**



(a)教师网络中的自适应邻接矩阵

(b)学生网络中的自适应邻接矩阵

图 3 自适应邻接矩阵可视化

Fig. 3 Visualization of adaptive adjacency matrices

4.4.2 损失函数与超参数分析

为了检验式(15)中的 3 个蒸馏损失，我们通过分别设置不同超参数值的组合，进行了各个损失的消融实验。其中，$\alpha\in\{0,1\}$，$\beta\in\{0,1\}$ 和 $\gamma\in\{0,1\}$，也就是考虑 3 个损失函数是否存在对整体模型表现的影响。相关实验结果如表 2 所列。在跨模态知识蒸馏中，我们发现仅仅进行概率分布对齐，即当 $\alpha=1,\beta=0,\gamma=0$ 时，效果并不是很好，这是因为不同模态下的概率分布本来就是多样化的，因此仅仅通过对齐概率分布是不足以实现知识蒸馏的。而单独在特征层面进行对齐，即当 $\alpha=0,\beta=1,\gamma=0$ 时，能取得很大的效果提升。本文使用自监督学习强化的特征对齐，在对齐同一时刻的特征的同时，还区分了不同时刻的特征，减小了它们之间的干扰，因此单独使用可带来最大的提升。仅在提出的时序关联特征层面上进行知识蒸馏，即 $\alpha=0,\beta=0,\gamma=1$ 时，也得到了很大的提升。因此，对语境信息的理解可以显著提升连续手语识别的准确度。

表 2 损失函数消融实验结果

Table 2 Results of ablative experiments on loss function

(单位：%)

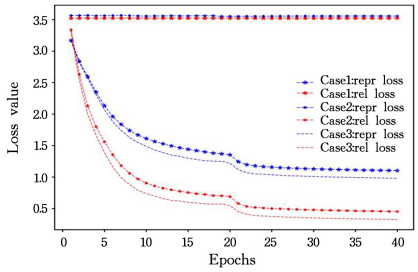
α,β,γ	Pseudo-label Accuracy	Validation WER	Test WER
0,0,0	48.5	46.2	45.0
1,0,0	50.9	41.9	40.4
0,1,0	54.7	37.8	36.3
0,0,1	52.7	40.4	38.0
0,1,1	56.9	35.8	34.9
1,1,1	56.9	35.0	34.7

为了验证时序关联特征的信息是否已经存在于教师网络和学生网络直接提取出的特征中，即是否直接特征对齐就可以达到时序关联特征对齐的效果。我们设置了 $\alpha=0,\beta=1,\gamma=1$ 的实验，结果表明直接特征对齐和时序关联特征对齐

相结合能取得更优的效果。这说明语境信息并不能完全由直接特征表达，时序关联特征与直接特征具有一定的差异性。最后将3种损失设置为同时存在，模型取得了最优的效果，可见概率分布对齐也能使学生网络学到一定的差异化知识，最终提升了性能。

4.4.3 损失变化可视化分析

仅优化特征对齐损失、仅优化关联对齐损失以及同时优化两者的情形如图4所示。在仅优化单一损失时，对应损失能够稳步下降，而另一个损失则几乎保持不变，这表明直接特征和时序关联特征所包含的信息是不同的。因为如果直接特征和时序关联特征是同种特征，那么在优化一种损失的同时另一类损失也会受到影响。另外，可以看到，同时优化两种损失相比优化单一损失能够更好地降低损失值，从而更好地优化模型。因此，学生网络模型的性能得到了提升，表明了所提出的TRKD方法的有效性。



注：Case 1表示只优化特征对齐损失的情形，Case 2表示只优化关联特征对齐损失的情形，Case 3表示同时优化了两者的情形

图4 损失变化曲线

Fig. 4 Loss change curve

4.5 蒸馏实验

我们进行了4组蒸馏实验来验证本文提出的知识蒸馏方法在连续手语识别问题上的有效性，并与其他3种优秀的方法进行了比较，分别是KD[13]，CRD[4]以及AugmentRGB[12]。KD在概率分布层面使用KL散度进行对齐以实现知识蒸馏。CRD是通过自监督学习进行知识蒸馏的优秀方法，它在特征层面使用对比学习来实现知识蒸馏。AugmentRGB是时序动作跨模态知识蒸馏的前沿方法，它利用协方差矩阵来捕捉关联信息。

4.5.1 STSG-Net→STGT蒸馏任务

表3列出了STSG-Net→STGT蒸馏任务中各种方法的实验结果。它们都对学生网络的表现有所提升，而TRKD方法表现出了最好的效果。与Augment RGB方法相比，所提出的TRKD方法通过自适应以及先验信息联合的图卷积学习手语上下文信息，相比协方差矩阵而言，所提出的TRKD方法能够自适应地学习全局关联关系，因此有更好的表现。与CRD方法相比，我们强化了在连续手语问题中手语上下文信息的学习，增加了额外的时序关联特征对齐，因此表现出了优越性。与直接使用交叉熵损失训练的STGT网络相比，我们在测试集上减少了10.3%的词错率，达到了使用RGB模态的I3D[17]模型的水平。然而，所提STGT的计算速度比I3D模型要快很多，其提升已经很明显，同时保持着计算速度上的优势。

表3 STSG-Net→STGT蒸馏任务的实验结果

Table 3 Results of distillation experiments from STGT-Net to STGT

（单位：%）

Methods	Pseudo-label Accuracy	Validation WER	Test WER
STGT w/o distillation	48.5	46.2	45.0
KD[13]	50.9	41.9	40.4
CRD[4]	54.4	37.8	36.3
AugmentRGB[12]	55.4	36.7	36.4
TRKD	56.9	35.0	34.7
STSG-Net(Teacher)	—	26.3	25.9
I3D(RGB)[17]	—	35.2	35.3
2DCNN+1DCNN(RGB)	—	33.6	32.5

4.5.2 2DCNN+1DCNN→STGT蒸馏任务

2DCNN+1DCNN→STGT蒸馏任务的实验结果如表4所列。相比使用STSG-Net作为教师网络，2DCNN+1DCNN的基础识别性能要弱一些，词错率高出7.3%。尽管如此，经过知识蒸馏之后学生网络的性能相差不超过2%，这说明合适的蒸馏算法可能比教师网络更重要。相比其他优秀的算法，TRKD以最好的表现展示出了时序关联特征对齐对学习语境信息理解能力的重要性。

表4 2DCNN+1DCNN→STGT蒸馏任务的实验结果

Table 4 Results of distillation experiments from 2DCNN+1DCNN to STGT

（单位：%）

Methods	Pseudo-label Accuracy	Validation WER	Test WER
STGT w/o distillation	48.5	46.2	45.0
KD[13]	50.9	41.5	40.9
CRD[4]	53.5	40.2	38.7
AugmentRGB[12]	54.6	39.0	37.4
TRKD	55.3	37.6	36.6
2DCNN+1DCNN(Teacher)	—	33.6	32.5

4.5.3 STSG-Net→ST-GCN蒸馏任务

STSG-Net→ST-GCN蒸馏任务的结果如表5所列。相比使用STGT作为学生网络，ST-GCN的表现要稍微差一些。这表明学生网络结构对知识蒸馏的效果有重要影响，所有的知识蒸馏方法在这个组合上的效果都略有下降。尽管在使用知识蒸馏方法之前STGT和ST-GCN的表现相近，但经过知识蒸馏后STGT可以更好地从教师网络学习时序关联特征来取得更好的表现。同样在各种知识蒸馏方法中TRKD取得了最好的表现，再次验证了让学生网络从教师网络学习上下文语境十分重要。

表5 STSG-Net→ST-GCN蒸馏任务的实验结果

Table 5 Results of distillation experiments from STSG-Net to ST-GCN

（单位：%）

Methods	Pseudo-label Accuracy	Validation WER	Test WER
ST-GCN w/o distillation	40.3	46.7	45.7
KD[13]	45.2	42.5	41.7
CRD[4]	46.3	39.6	38.5
AugmentRGB[12]	47.4	39.4	39.4
TRKD	46.1	39.2	38.2
STSG-Net(Teacher)	—	26.3	25.9

4.5.4 2DCNN+1DCNN→ST-GCN 蒸馏任务

2DCNN+1DCNN→ST-GCN 蒸馏任务的实验结果如表 6 所列。在这组实验中，AugmentRGB 的表现略微超过了 TRKD。这可能是在该蒸馏任务中，协方差矩阵的表现更为稳定，因此取得了最好的效果。

表 6 2DCNN+1DCNN→ST-GCN 蒸馏任务实验结果

Table 6 Results of distillation experiments from 2DCNN+1DCNN to ST-GCN

Methods	Pseudo-label Accuracy	Validation WER	Test WER
ST-GCN w/o distillation	40.3	46.7	45.7
KD[13]	43.3	42.0	41.0
CRD[4]	44.0	41.4	40.8
AugmentRGB[12]	46.5	40.6	39.3
TRKD	45.0	40.9	39.7
2DCNN+1DCNN(Teacher)	—	33.6	32.5

结束语 本文提出了一种新颖的跨模态知识蒸馏方法 TRKD，用于提升基于骨骼关键点模态的连续手语识别的性能，同时保持模型在真实场景中计算速度的优势。该方法通过使用 GCN 学习时序关联特征，以表达各个模态的手语上下文信息，并通过特征对齐的方式使学生网络能够学习到教师网络对手语上下文的理解。GCN 中的邻接矩阵通过先验信息以及可学习的自适应方法得到。在特征对齐中，TRKD 方法使用了对比学习，使学生网络提取的特征接近于同一时刻的教师网络特征而远离其他时刻的特征，从而有效地实现了特征对齐。TRKD 方法在 Phoenix-2014 数据集上进行了消融实验，验证了 GCN 提取到的特征能有效地反映手语上下文信息，其中先验信息以及自适应邻接矩阵可以从局部以及全局的角度来表达帧与帧之间的关联。另外，在多项蒸馏任务上的实验结果也表明了 TRKD 方法能有效地提升骨骼关键点模态模型的识别效果。

参 考 文 献

[1] CUI R P, LIU H, ZHANG C S. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7361-7369.

[2] KOLLER O, CAMGOZ N C, NEY H, et al. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019: 2306-2320.

[3] LUO Z L, HSIEH J T, JIANG L, et al. Graph distillation for action detection with privileged modalities[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 166-183.

[4] TIAN Y L, KRISHNAN D, ISOLA P. Contrastive Representation Distillation[C]// International Conference on Learning Representations. 2020.

[5] THOMAS N K, MAX W. Semi-supervised classification with graph convolutional networks[C]// International Conference on Learning Representations. 2017.

[6] OORD A V D, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv: 1807.03748, 2018.

[7] PU J F, ZHOU W G, LI H Q. Iterative alignment network for continuous sign language recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4165-4174.

[8] CUI R P, LIU H, ZHANG C S. A deep neural framework for continuous sign language recognition by iterative training[J]. IEEE Transactions on Multimedia 2019, 21(7): 1880-1891.

[9] ZHOU H, ZHOU W G, ZHOU Y, et al. Spatial-temporal multi-cue network for continuous sign language recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 13009-13016.

[10] WANG Z C, ZHANG J Q. Continuous Sign Language Recognition based on Multi-Part Skeleton Data[C]// 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[11] GARCIA N C, MORERIO P, MURINO V. Modality distillation with multiple stream networks for action recognition[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 103-118.

[12] DAI R, SRIJAN D, BREMOND F. Learning an Augmented RGB Representation with Cross-Modal Knowledge Distillation for Action Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13053-13064.

[13] HINTON G, VINYALS O, DEANY J. Distilling the knowledge in a neural network[J]. arXiv: 1503.02531, 2015.

[14] OSCAR K, JENS F, HERMANN N. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers[J]. Computer Vision and Image Understanding, 2015, 141: 108-125.

[15] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[16] CAO Z, HINDALGO G, SIMONT, et al. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(1): 172-186.

[17] JOAO C, ANDREW Z. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

XIAO Zheng-ye, born in 1996, bachelor. His main research interests include machine learning and computer vision.

FANG Yu-chun, born in 1975, Ph. D, professor. Her main research interests include machine learning, multimedia, pattern recognition and image processing.

(责任编辑：喻藜)