JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
07
VOL.27

ISSN1006-8961
CN11-3758/TB

舰船目标检测 *P2078*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第7期（总第315期）
2022年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年7月 第27卷 第7期（总第315期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

深度学习背景下视觉显著性物体检测综述(第2112页)

提取全局语义信息的场景图生成算法(第2214页)

融合弱监督目标定位的细粒度小样本学习(第2226页)

Volume 27, Number 7
Published July 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Review of deep learning based salient object detection(P2112)

Global semantic information extraction based scene graph generation algorithm(P2214)

Weakly-supervised object localization based fine-grained few-shot learning(P2226)

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)07-2185-14

论文引用格式: Chen T, Xing S, Yang W W and Jin J Q. 2022. Spatio-temporal features based human facial expression recognition. Journal of Image and Graphics,27(07):2185-2198(陈拓,邢帅,杨文武,金剑秋. 2022. 融合时空域特征的人脸表情识别. 中国图象图形学报,27(07):2185-2198)

[DOI:10.11834/jig.200782]

融合时空域特征的人脸表情识别

陈拓,邢帅,杨文武*,金剑秋

浙江工商大学计算机与信息工程学院,杭州　310018

摘　要: **目的**　人脸表情识别是计算机视觉的核心问题之一。一方面,表情的产生对应着面部肌肉的一个连续动态变化过程,另一方面,该运动过程中的表情峰值帧通常包含了能够识别该表情的完整信息。大部分已有的人脸表情识别算法要么基于表情视频序列,要么基于单幅表情峰值图像。为此,提出了一种融合时域和空域特征的深度神经网络来分析和理解视频序列中的表情信息,以提升表情识别的性能。**方法**　该网络包含两个特征提取模块,分别用于学习单幅表情峰值图像中的表情静态"空域特征"和视频序列中的表情动态"时域特征"。首先,提出了一种基于三元组的深度度量融合技术,通过在三元组损失函数中采用不同的阈值,从单幅表情峰值图像中学习得到多个不同的表情特征表示,并将它们组合在一起形成一个鲁棒的且更具辩识能力的表情"空域特征";其次,为了有效利用人脸关键组件的先验知识,准确提取人脸表情在时域上的运动特征,提出了基于人脸关键点轨迹的卷积神经网络,通过分析视频序列中的面部关键点轨迹,学习得到表情的动态"时域特征";最后,提出了一种微调融合策略,取得了最优的时域特征和空域特征融合效果。**结果**　该方法在3个基于视频序列的常用人脸表情数据集CK+(the extended Cohn-Kanade dataset)、MMI(the MMI facial expression database)和Oulu-CASIA(the Oulu-CASIA NIR&VIS facial expression database)上的识别准确率分别为98.46%、82.96%和87.12%,接近或超越了当前同类方法中的表情识别最高性能。**结论**　提出的融合时空特征的人脸表情识别网络鲁棒地分析和理解了视频序列中的面部表情空域和时域信息,有效提升了人脸表情的识别性能。

关键词:人脸表情识别(FER);深度学习;深度度量学习;三元组损失;特征融合

Spatio-temporal features based human facial expression recognition

Chen Tuo, Xing Shuai, Yang Wenwu*, Jin Jianqiu

School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou　310018, China

Abstract: **Objective** Human facial expression recognition (FER) is one of the key issues of computer vision analysis like human-computer interaction, medical care and intelligent driving. FER research has mainly two challenges in related to expression feature extraction and classification recognition. Current methods are mainly design facial expression features artificially, while deep learning based methods can independently learn to obtain semantic facial expression features. The deep learning based FER technology can integrate the two training processes of feature extraction and facial expression recognition. It has strong generalization ability and good recognition accuracy currently. Most of the existing FER algorithms

收稿日期:2020-12-29;**修回日期**:2021-03-22;**预印本日期**:2021-03-29

*通信作者:杨文武　wwyang@zjgsu.edu.cn

基金项目:国家自然科学基金项目(61972353);国家重点研发计划资助(2018YFB1404102);浙江省自然科学基金项目(LY21F020010)

Supported by: National Natural Science Foundation of China (61972353); National Key R&D Program of China (2018YFB1404102); Natural Science Foundation of Zhejiang Province, China (LY21F020010)

are based on expression video sequences or a single peak expression scenario. However, the generation of expression corresponds to a continuous dynamic change process of facial muscles, and the motion-based expression peak frame identifies completed expression information in common. Our method demonstrates a spatio-temporal and features based deep neural network to analyze and understand video sequences derived expression information to improve expression recognition ability. **Method** Our network learn the static "spatial feature" of the expression and its dynamic "temporal feature" based on the video sequence, respectively. First, we illustrate a deep metric fusion network based on triplet loss learning. Our network is composed of two sub-modules like deep convolutional neural network (DCNN) module and *N*-metric module. The DCNN module is derived from a general convolutional neural network (CNN) to extract common detailed CNN facial features. In this module, the Visual Geometry Group 16-layer net (VGG16)-face network model structure is adopted where the output of its final 4 096-dimensional fully connected layer is used as the benched CNN feature. The *N*-metric module contains fully multiple connected layer branches. Each branch uses a triplet loss function to implement the supervised learning to represent different expression semantic multi-features. These dual-features representations are fused through two fully connected layers further. A more robust and spatial feature expression is illustrated. The each two fully connected layers have 256 hidden units and the output of each branch is merged together in a concatenating manner in the DCNN module. In the *N*-metric module, all of fully connected layer branches are shared to the same CNN feature. For example, the output of the final fully connected layer is used as the input of each branch in the DCNN module. In addition, a fixed dimension fully connected layer is used via each branch and it is associated with a certain threshold sampling for learning the corresponding feature embedding. Each branch is supervised and learned by the corresponding triple loss function. Next, facial expressions are essential to facial expression changes in motion because the changes are integrated to the overall facial expression changing. Existing methods are challenged to extract the dynamic expression features in the context of consecutive frames derived time domain through manual design or deep learning methods. But, manual-designed features are constrained of facial image sequence based temporal features extraction. The image sequence related deep neural network is insufficient to employ the prior knowledge of the key features of the face as well due to the non-learning temporal featured expressions. Our landmark-trajectory convolutional neural network analyzes the trajectory in the video sequence and learns the dynamic "temporal features" of the expression sequence consequently, which extracts the accurate motion characteristics of facial expressions in the time domain. Our network consists of four convolutional layers and two fully connected layers. The input of the landmark trajectory CNN (LTCNN) sub-network is a similar feature map constructed based on the trajectory of facial expression in the video. Third, a fine-tuning based fusion strategy is conducted to combine the learned features of two network modules obtained further, which achieves the temporal and spatial features based fusion result optimally. We train the deep metric fusion (DMF) and LTCNN sub-networks each, combine the two sub-networks through feature fusion, and fine-tuning them in an end-to-end manner sequentially. The implemented hyper-parameters are used for fine-tuning training in DMF sub-network optimization. **Result** Our demonstrated FEC algorithm is tested and verified on three public facial expression databases in terms of the extended Cohn-Kanade dataset (CK+), the MMI facial expression database (MMI), and the Oulu-CASIA NIR&VIS facial expression database (Oulu-CASIA). Our method achieves the recognition accuracy of 98.46%, 82.96%, and 87.12% on the databases of CK+, MMI, and Oulu-CASIA, respectively. **Conclusion** For our deep learning based network integrated temporal and the spatial features both to realize video sequences based FER. In the network, our two sub-modules are used to learn the "spatial features" of the facial expression at the peak frame and the "temporal features" of facial expression motion. Finally, a fusion strategy is carried out to achieve better fusion effect of temporal and spatial features based on overall fine-tuning. Our FER method has its potentials to develop further.

Key words: facial expression recognition (FER); deep learning; deep metric learning; triplet loss; feature fusion

0 引 言

面部表情提供了丰富的情感信息,是人们内心情感状态最直接和自然的一种传达方式(Li 和 Deng,2020)。人脸表情识别在教育质量监督(Whitehill 等,2014)、医疗应用(Gutierrez,2020)、人机交互(Vinciarelli 等,2009)和自动驾驶等诸多领

域有着广阔的应用前景，因此逐渐成为相关领域的一个研究热点。人脸表情的产生对应着一个连续的面部肌肉运动过程。多数已有的人脸表情识别方法主要针对该运动过程中的表情峰值帧，通过分析和提取该帧人脸图像中的表情空间特征信息来识别其中的面部表情。为了利用面部表情的运动信息，一些方法通过分析人脸表情的视频序列，希望从中提取出的人脸表情特征不仅包含了每帧图像中的表情“空域信息”，并且也包含了连续帧之间的表情“时域信息”，从而实现表情识别性能的有效提升（Zhao 等，2018；Zhang 等，2017；Hasani 和 Mahoor，2017；Kumawat 等，2019）。但是，视频序列邻接帧中的表情空域信息具有一定的连贯性和冗余度，这种冗余性不仅造成了信息浪费，也加大了有效信息的提取和分辨难度（Zhao 等，2018）；此外，面部表情的运动变化可以认为是人脸关键组件（如眉毛、眼睛、鼻子和嘴巴等）的动态变化组合，而直接分析图像序列无法有效利用人脸关键组件的先验知识，因而不利于人脸表情时域信息的提取。

针对上述问题，提出了一种融合时空域特征的深度学习神经网络，以高效鲁棒地分析和理解视频序列中的面部表情空域和时域信息。该网络主要包含两个特征提取模块，分别用于学习单幅表情峰值图像中的表情静态“空域特征”和视频序列中的表情动态“时域特征”。此外，该网络还包含一种微调融合策略，该策略取得了最优的时域特征和空域特征融合效果，有效提升了人脸表情的识别性能。

对于单幅表情峰值图像，个体差异以及光照、遮挡和头部姿势等外在干扰因素都会与其中的表情特征非线性耦合在一起，使得鲁棒提取图像中的表情特征极具挑战性（Liu 等，2017）。基于三元组的深度度量学习技术是一种有效的表情特征学习方法，它可以使得相同表情类别的样本在特征空间中相互靠近，而不同表情类别的样本在该空间中互相远离，最终学习得到能够有效表达表情变化的潜特征（latent features）。在实验中观察到，三元组损失函数中的阈值可以在一个范围内有效变化，并且每个阈值本质上对应着一个不同的类间差异分布，如图 1 所示。因此，在“空域特征”学习模块中，提出了一种基于三元组的深度度量融合技术，通过在三元组损失函数中采用不同的阈值，从单幅表情峰值图像中学习得到多个不同的表情特征表示，并将它们组合在一起，最终形成了一个鲁棒的且更具识别能力的表情特征。

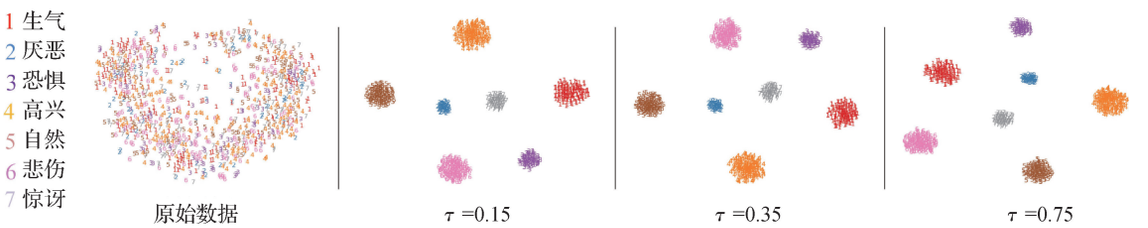


图 1　基于不同三元组阈值学习得到的特征所进行的类间变化分布可视化

Fig. 1　Distributions of inter-class variations with respect to the features learned by the triplet loss with different margins

考虑到面部表情是由一些关键区域肌肉运动产生的，而这些区域的运动可由面部关键点的运动轨迹变化来表示，因此在“时域特征”提取模块中采用了简单的 2 维卷积神经网络（convolutional neural networks，CNN），通过分析视频序列中的面部关键点轨迹，学习得到表情的时序动态变化特征。此外，为了有效提升人脸表情的识别性能，还需要考虑如何有效融合上述两个模块中学习得到的空域特征和时域特征，使得这两个特征在表情识别任务中能够最大化地互为补充。通过大量实验，测试了各种可能的融合策略，最终提出了一种所谓的“微调融合策略”，取得了最优的时空域特征融合效果。主要贡献如下：1）提出了一种融合时空域特征的深度学习神经网络。该网络通过分析单幅表情峰值图像和视频序列中的面部关键点轨迹，有效提取了视频序列中的面部表情空域和时域特征。2）设计了一种基于三元组的深度度量融合技术。不同于传统的三元组度量学习仅使用单个阈值，该技术使用了多个阈值，不仅避免了费时的最优阈值的选取，并且有效提升了提取特征的鲁棒性和可分辨能力。3）提出了一种微调融合策略，取得了最优的时域特征和空域特征融合效果。4）该方法有效提升了人脸表情的

识别性能,在3个公开的基于视频序列的人脸表情数据集CK+(the extended Cohn-Kanade dataset)(Lucey等,2010)、MMI(the MMI facial expression database)(Pantic等,2005)和Oulu-CASIA(the Oulu-CASIA NIR&VIS facial expression database)(Zhao等,2011)上均接近或超越了此前其他各类表情识别方法的性能。

1 相关工作

通常认为生气、高兴、恐惧、厌恶、悲伤和惊讶等6种基本情感在不同文化中具有共通性,因此人脸表情识别研究通常根据这些情感对表情进行分类(Ekman和Friesen,1971)。根据输入特征表示的不同,人脸表情识别方法大致可以分为基于图像的方法和基于视频序列的方法两类(Zeng等,2009)。已有的研究大多属于基于图像的表情识别方法(Liu等,2017;Acharya等,2018;Yang等,2018),主要考虑单幅表情峰值图像中的表情静态"空域特征"。基于视频序列的表情识别方法则进一步考虑了表情生成过程中的面部运动信息(Zhang等,2017;Hasani和Mahoor,2017;Kumawat等,2019),即所谓的表情动态"时域特征",因而通常能够更加有效地完成表情识别任务。

1.1 基于手工设计特征的传统方法

为了在视频序列中提取面部表情的时序特征,研究人员将基于图像的传统手工特征扩展到连续的视频帧特征,提出了LBP-TOP(local binary patterns from three orthogonal planes)(Zhao和Pietikainen,2007)、3D-HOG(3D-histogram of oriented gradients)(Klaser等,2008)以及3D-SIFT(3D-scale-invariant feature transform)(Scovanner等,2007)等方法。Jain等人(2011)使用条件随机场和手工创建的形状外观特征对每个面部形状进行时间建模。Taini等人(2008)则提出了一种纵向地图结构,在Oulu-CASIA数据库上实现了较好的识别性能。Wang等人(2013)通过一种间隔时序贝叶斯网络,捕获了面部肌肉之间复杂的时空关系。Ptucha等人(2011)提出了一种基于流形的稀疏表示,通过使用基于监督的局部保形投影来映射低维流形中的特征,进而实现表情识别。Sikka等人(2016)提出了基于潜序数模型的视频表情识别,使用弱监督分类器将面部关键点的SIFT和LBP特征进行整合,并将表情作为潜变量进行学习。

虽然已有的研究工作设计了各种各样的手工特征来提取表情的时空信息并对其进行分类,但是基于深度卷积神经网络的人脸表情识别方法越来越流行,相比于基于手工设计特征的传统方法,显著提升了表情识别性能。

1.2 基于深度学习的表情识别方法

近年来,深度卷积神经网络逐渐主导了各种计算机视觉任务。例如图像分类(Simonyan和Zisserman,2015)、目标识别(Ren等,2017)和物体分割(Shelhamer等,2017)等。对于视频序列中的人脸表情识别任务,基于深度学习的网络模型也取得了诸多最新研究成果。Jung等人(2015)提出一种使用DTAN(deep temporal appearance network)和DTGN(deep temporal geometry network)两个深度神经网络的方法。DTAN网络是一个简单的3D卷积神经网络,用于从视频序列中捕获表情的时空信息;DTGN网络是一个由全连接层构成的浅层网络,用来捕获面部关键点的时序运动变化。通过对这两个网络进行同时微调,该方法获得了当时最先进的表情识别性能。Zhang等人(2017)进一步改进了Jung等人(2015)的方法,提出了一个空间网络MSCNN(multi-signal convolutional neural network)和一个时间网络PHRNN(part-based hierarchical recurrent neural network),其中MSCNN对应着一个基于单幅表情峰值图像的简单卷积神经网络,用于学习表情的空间信息,而PHRNN则由几层循环神经子网络(recurrent neural network,RNN)构成,用于学习视频序列中的表情时间信息。此外,Zhang等人(2017)还提出了一种排序融合策略,以有效融合这两个网络学习得到的表情时空特征。为了更好地学习视频序列中的表情时空特征,Hasani和Mahoor(2017)将面部关键点和残差单元的输入张量相乘替换原始3D Inception-ResNet中的残差结构。Kumawat等人(2019)提出了一种称为局部二值体的3D卷积层对图像序列上的面部表情进行识别。Deng等人(2019)提出可以同时捕获微观和宏观运动的双流循环网络,以此改善基于视频的情感识别性能。

本文方法的基本思想与Zhang等人(2017)方法相似,提出的融合时空域特征的深度学习神经网络主要包含两个特征提取模块,分别用于学习单幅

表情峰值图像中的表情静态“空域特征”和视频序列中的表情动态“时域特征”，但与 Zhang 等人(2017)及其他方法相比，有以下 3 方面的区别：1)一般的表情识别网络均使用 softmax 损失作为训练监督函数，虽然从中提取的 CNN 特征具有一定语义，但是它们与表情含义并没有直接关联，这是因为 softmax 损失函数并没有显式地考虑类内的紧凑和类间的分离。提出的基于三元组的深度度量融合技术不仅能够学习得到有效表达表情变化的语义特征，并且相比于传统的三元组度量学习，这些特征更加鲁棒且更具识别能力。2)循环神经网络一般具有更高的学习和训练难度，因此使用了简单的 2 维卷积神经网络，通过分析视频序列中的面部关键点轨迹，学习得到表情的时序变化信息。3)一般情况会使用特征级别或者决策级别的融合方式来组合多个网络的学习结果，但是不同的网络模型具有不同的学习能力且学习到的特征也不尽相同，简单的融合方式有时不仅无法实现时域特征和空域特征的互补融合，还可能会削弱它们彼此的识别性能。因此，提出了一种微调融合策略，取得了最优的时域特征和空域特征的融合效果。

2　本文算法

如图 2 所示，本文提出的融合时空域特征的深度学习神经网络主要包含空域特征提取模块 DMF(deep metric fusion)和时域特征提取模块 LTCNN(landmark trajectory CNN)两个子网络模块。其中，DMF 子网络使用了本文提出的深度度量融合技术，以视频序列中的单幅表情峰值帧图像为输入，从中提取出表情的静态空间特征。在 LTCNN 子网络中，采用了一个简单的 2 维卷积神经网络结构，利用人脸关键组件中的先验知识，以视频序列中人脸关键点轨迹构成的类特征图作为输入，进而从中提取出连续帧中隐含的表情时序运动特征。在实现中，为了达到网络的最佳训练效率并取得最优性能，首先分别对 DMF 子网络和 LTCNN 子网络进行单独训练，然后将时域和空域两个不同维度上的特征子模块有效融合在一起，以最终提升人脸表情的识别性能。

2.1　基于深度度量融合的空域特征提取

深度度量学习的目标在于学习得到一个特征嵌入函数 $f(\boldsymbol{x})$，它将图像 $\boldsymbol{x}$ 映射到一个特征空间 $\mathbf{R}^d$，使得相同表情类别的样本在特征空间中相互靠近，而不同表情类别的样本在该空间中互相远离。与 Schroff 等人(2015)的方法类似，算法将特征嵌入约束在一个 d 维的超球面上，即 $\|f(\boldsymbol{x})\|_2 = 1$。为了实现该目标，基于三元组的深度度量学习技术从训练数据样本中构造了一系列的三元组。其中，每个三元组包含了两个具有相同表情的图像样本(分别称为锚点和正样本)以及一个具有不同表情的图像样本(称为负样本)，而三元组损失函数则保证在每个三元组中，正样本与锚点之间的特征距离比负样本与锚点的特征距离要小于一个给定的阈值 τ。因此，理想情况下，三元组度量学习技术能够从人脸表情训练数据中学习得到一种隐特征表示，该隐特征能有效地表达数据中的人脸表情变化，而忽略其中的非表情因素干扰。三元组损失函数中的阈值可以在一个范围内变化，并且每个阈值本质上对应着一个不同的类间差异分布。因此，提出的深度度量融合技术的核心思想是通过采样多个阈值 $\{\tau_1,\tau_2,\tau_3,\cdots,\tau_N\}$ 来构建不同的特征嵌入函数 $\{f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),f_3(\boldsymbol{x}),\cdots,f_N(\boldsymbol{x})\}$，进而学习得到不同的表情特征表示。

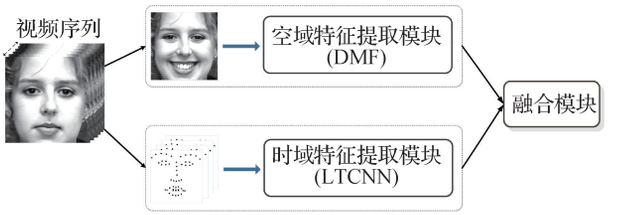


图 2　整体网络结构

Fig. 2　The proposed network structure

如图 3 所示，DMF 子网络中的深度度量融合网络结构主要由 DCNN(deep convolutional neural network)模块和 N-metric(N-metric network)模块两个子模块组成。其中，DCNN 模块对应着一个通用卷积神经网络(CNN)，用于提取人脸图像中普遍的细节特征，即 CNN 特征。在该模块中，采用了 VGG16-Face(Visual Geometry Group 16-layer net)网络模型结构(Parkhi 等，2015)，以其最后一个 4 096 维全连接层的输出作为所需的 CNN 特征，N-Metric 模块则包含多条全连接层分支，每条分支通过一个三元组损失函数来进行监督学习，从而得到多个不同的表情语义特征表示。这些特征表示再通过后续的两个

全连接层进一步融合在一起,最终得到一个更加鲁棒且更具分辨能力的表情特征,即前述的表情静态“空域特征”。实现中,上述的后续两个全连接层每层具有256个隐藏单元,DCNN模块中每条分支的全连接层输出通过级联方式合并在一起,以作为后续全连接层的输入。在N-Metric模块中,所有全连接层分支共享相同的CNN特征,即DCNN模块中最后一个全连接层的输出作为每条分支的输入。对于每一条分支,采用固定维数大小的全连接层,并将它关联到某个阈值采样$\tau_i, i \in \{1,2,\cdots,N\}$,以学习得到对应的特征嵌入$f_i(\boldsymbol{x})$。在训练过程中,每条分支由对应的三元组损失函数进行监督学习,令这些损失函数分别为$loss_i, i \in \{1,2,\cdots,N\}$。给定一个三元组集合,其中每个三元组包含一个锚点$\boldsymbol{x}_i^{\mathrm{a}}$,一个正样本$\boldsymbol{x}_i^{\mathrm{p}}$和一个负样本$\boldsymbol{x}_i^{\mathrm{n}}$。三元组损失函数的目标是保证正样本与锚点之间的特征距离比负样本与锚点之间的特征距离要小于一个给定的阈值τ_i,即

$$\|f(\boldsymbol{x}_i^{\mathrm{n}}), f(\boldsymbol{x}_i^{\mathrm{a}})\|_2^2 > \|f(\boldsymbol{x}_i^{\mathrm{p}}), f(\boldsymbol{x}_i^{\mathrm{a}})\|_2^2 + \tau_i \tag{1}$$

因此,三元组损失函数$loss_i$定义为

$$loss_i = \frac{1}{2M}\sum_{i=1}^{M}[\max(0, \|f(\boldsymbol{x}_i^{\mathrm{p}}) - f(\boldsymbol{x}_i^{\mathrm{a}})\|_2^2 - \|f(\boldsymbol{x}_i^{\mathrm{n}}) - f(\boldsymbol{x}_i^{\mathrm{a}})\|_2^2 + \tau_i) + \max(0, \|f(\boldsymbol{x}_i^{\mathrm{a}}) - f(\boldsymbol{x}_i^{\mathrm{p}})\|_2^2 - \|f(\boldsymbol{x}_i^{\mathrm{n}}) - f(\boldsymbol{x}_i^{\mathrm{p}})\|_2^2 + \tau_i)] \tag{2}$$

式中,M为集合中的三元组个数。注意,上述三元组损失函数不仅保证了正样本与锚点之间的特征距离比负样本与锚点之间的特征距离小于给定的阈值τ_i,同时也保证了锚点与正样本之间的特征距离比负样本与正样本之间的特征距离小于该给定的阈值。

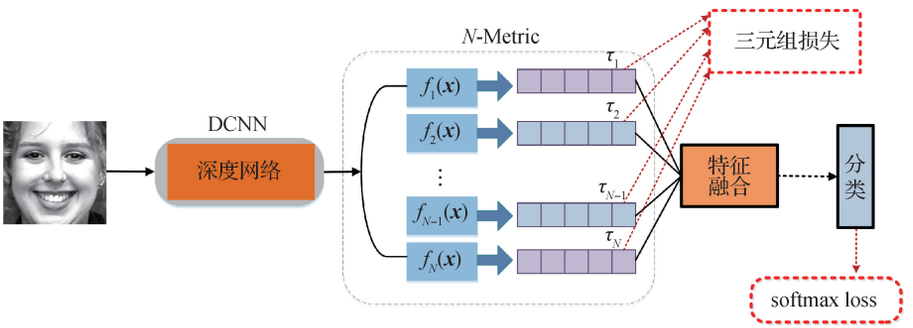


图3 DMF子网络:深度度量融合网络结构

Fig. 3 DMF sub-network: the structure of deep metric fusion

2.2 基于人脸关键点轨迹的时域特征提取

考虑到卷积神经网络(CNN)出色的特征表示学习能力,同时为了避免3D CNN的高计算量,可以使用2D CNN学习视频序列上的时域表情运动特征。因此,提出了基于人脸关键点轨迹的卷积神经网络(LTCNN),通过分析视频序列中人脸关键点的运动变化来提取其中蕴含的表情时域特征。如图4所示,LTCNN子网络对应一个简单的2D卷积神经网络,由4个卷积层和2个全连接层组成。LTCNN子网络输入的是由视频中人脸关键点轨迹构建而成的类特征图。给定一个人脸表情视频序列,首先从视频中均匀采样到一个固定帧数的图像序列。在实现中,均匀采样了11帧。然后,针对每个采样帧,可以在人脸的双眼、眉毛、鼻子和嘴巴等4个关键部位上检测出51个关键点,如图4所示。所有采样帧中关键点的位置变化即对应着视频中人脸关键点的运动轨迹。最后,将所有采样帧中关键点的坐标组合在一起,即得到输入到LTCNN子网络的类特征图。此外,受图像RGB三通道表示的启发,基于关键点的序列数据,在实现中采用两种方式构造LTCNN子网络的输入特征图。

1)将每帧中51个关键点的x、y坐标依次组合在一起,形成一个102维的特征向量$(x_1, y_1, x_2, y_2, \cdots, x_{51}, y_{51})$。然后将所有采样帧对应的特征向量组合在一起,即得到一个11×102×1大小的向量,该向量可以看做是带1个通道而大小为11×102的特征图,并称以该特征图作为输入的LTCNN子网络为LTCNN-1CL。

2)将每帧中51个关键点的x、y坐标分别组合在一起,形成两个51维的特征向量$(x_1, x_2, \cdots, x_{51})$和$(y_1, y_2, \cdots, y_{51})$。然后分别将所有采样帧对应的$\boldsymbol{x}$或$\boldsymbol{y}$特征向量组合在一起,即得到一个11×51×2

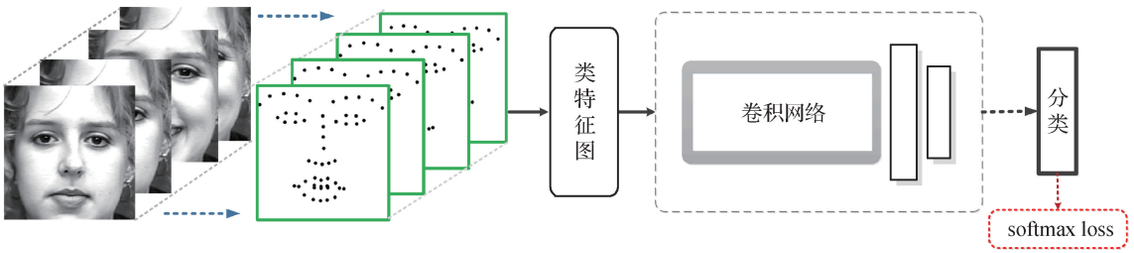


图 4　LTCNN 子网络：基于人脸关键点轨迹的卷积神经网络结构

Fig. 4　LTCNN sub-network: the structure of landmark trajectory convolutional neural network

大小的向量，该向量可以看做是带 2 个通道而大小为 11 × 51 的特征图，并称以该特征图作为输入的 LTCNN 子网络为 LTCNN-2CL。

2.3　DMF 与 LTCNN 子网络的最优融合

提出的融合时空域特征的深度学习神经网络通过将提取空域信息的 DMF 子网络和提取时域信息的 LTCNN 子网络融合在一起，实现了人脸识别性能的有效提升。一般地，通常可以采用特征融合策略或者决策融合策略。

2.3.1　基于决策融合的后期融合策略

多数人脸表情识别方法通过决策融合来提高算法性能。如图 5 所示，该融合策略首先单独训练 DMF 和 LTCNN 子网络，每个子网络得到一个分类结果，然后将所有子网络的分类结果通过某种数学方式进行汇总，汇总结果即为最终的分类结果。一般可以使用简单的加权平均来汇总分类结果，也可以采用稍微复杂的汇总方式，例如决策排序融合（Zhang 等，2017）。在决策融合策略中，因为两个子网络是单独训练，因而无法考虑它们之间的互补性。

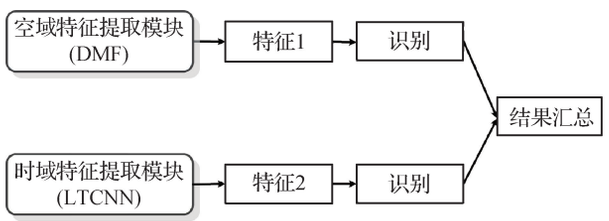


图 5　基于决策融合的后期融合策略

Fig. 5　Late-fusion strategy based on decision fusion

2.3.2　基于特征融合的前期融合策略

在该融合策略中，DMF 子网络的特征输出（即最后一个全连接层的输出）与 LTCNN 子网络的特征输出（即最后一个全连接层的输出）通过后续的全连接层融合在一起，以得到一个更具分辨能力的表情特征，如图 6 所示。在实现过程中，使用了一个 256 大小的全连接层来融合 DMF 和 LTCNN 子网络的输出特征，并结合 softmax 表情分类层对整个网络通过一种端到端的方式进行训练。但是，由于 DMF 和 LTCNN 子网络在学习过程中的收敛速度可能不同，因而以统一的学习率对它们进行端到端的训练无法充分照顾它们不同的收敛特性。

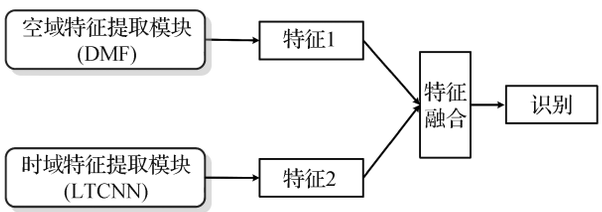


图 6　基于特征融合的前期融合策略

Fig. 6　Early-fusion strategy based on feature fusion

2.3.3　基于微调的特征融合策略

针对前期特征融合策略下 DMF 和 LTCNN 子网络可能存在不同训练下收敛速度不一致以及后期决策融合策略下两个子网络因单独训练而没有考虑结果互补性的问题，提出了第 3 种融合策略，即基于微调的特征融合策略。其思想简单，先对 DMF 和 LTCNN 子网络分别进行训练，然后通过特征融合的方式将这两个子网络结合在一起，并以端到端的方式进行统一微调。在实现中，采用 DMF 子网络优化时所用的超参数进行微调训练，并考虑了 4 种微调方案。1）局部微调。固定两个子网络参数，只微调后面新加的全连接融合层和 softmax 分类层。2）固定 DMF 的微调。固定 DMF 子网络参数，联合微调 LTCNN 子网络以及后面新加的全连接融合层和 softmax 分类层。3）固定 LTCNN 的微调。固定 LTCNN 子网络参数，联合微调 DMF 子网络以及后面新加的全连接融合层和 softmax 分类层。4）整体微调。

对网络中所有模块进行联合微调。

实验发现,后 3 种微调方案均能够有效实现 DMF 和 LTCNN 子网络的同步训练以及互补融合。其中,整体微调取得了最高的表情分类精度。

3 实验结果

3.1 3 个表情数据集

为了评估提出的融合时空域特征的深度学习神经网络的性能,选取 3 个公开且广泛使用的基于视频序列的表情数据集 CK + (Lucey 等,2010)、MMI (Pantic 等,2005) 和 Oulu-CASIA (Zhao 等,2011) 进行实验。

CK + (Lucey 等,2010) 是人脸表情识别评估方法中使用最为广泛的实验室环境下数据集,包含来自 118 个主体的 327 个视频序列,每个序列包括 10 ～60 帧不等,表示了中性面部表情到峰值表情的变化过程。每个视频序列有 1 个标签,对应生气 (anger)、蔑视 (contempt)、厌恶 (disgust)、恐惧 (fear)、高兴 (happy)、悲伤 (sadness) 和惊讶 (surprise) 等 7 种基本表情之一。以原始视频中的第 1 帧作为初始帧,表情峰值帧为最后 1 帧,中间均匀采样 11 帧来获得具有固定帧数的样本数据。由于 CK + 没有提供指定的训练集、验证集和测试集,按照已有的协议 (Liu 等,2014),将数据样本以严格的主体独立方式分为 10 折,然后进行 10 折交叉验证。主体独立使得任何两个子集中的主体都是互斥的,最终识别精度为 10 次验证的平均值。

相比于 CK + ,MMI 数据集 (Pantic 等,2005) 中的个体表情差异更大,并且部分存在遮挡 (例如眼镜和胡须等),因此更具挑战性。数据集由来自 31 个主体的 236 个图像序列组成,每个序列对应 6 个基本表情 (没有蔑视) 之一,实验中选择了正面视图拍摄的 208 个序列。每个序列以中性表情开始,在序列中间达到表情峰值,并以中性表情结束。与 CK + 类似,通过均匀采样获得具有固定帧数的样本,并使用严格主体独立的方式进行 10 折交叉验证。

Oulu-CASIA 数据集 (Zhao 等,2011) 在明亮、弱光和黑暗 3 种不同的光照条件下采集,每种光照条件下分别为 80 个主体 (年龄 23 ～58 岁) 采集了 6 种基本面部表情 (没有蔑视),即该数据库在每种光照条件下都有 480 个视频序列。与 CK + 类似,所有序列以中性表情开始,在表情达到峰值时结束。实验中采用明亮光照条件下的数据,并以严格主体独立的方式进行 10 折交叉验证。

3 个表情数据集的部分示例如图 7 所示。其中,MMI 和 Oulu-CASIA 数据集中没有“蔑视”的面部表情。

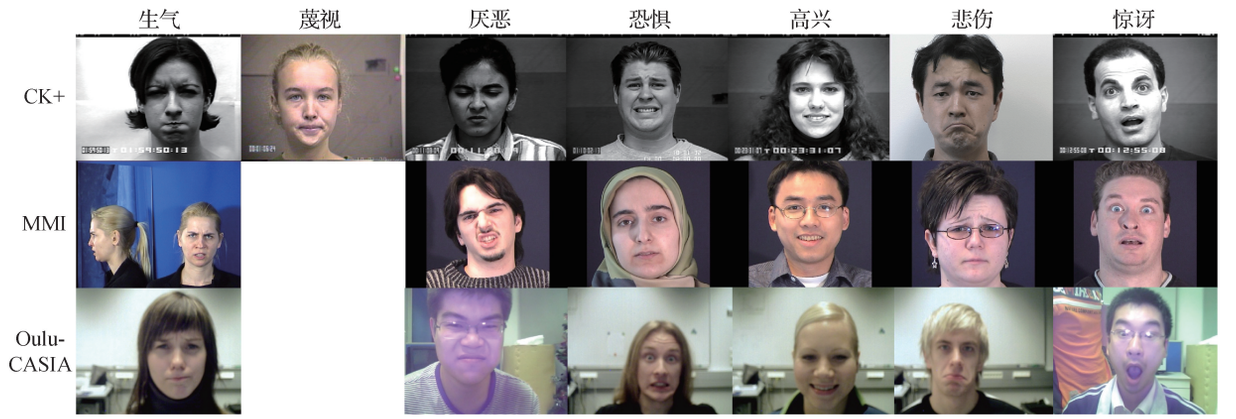


图 7 3 个表情数据集中的部分示例

Fig. 7 Some examples of three expression datasets

3.2 实现细节

3.2.1 DMF 子网络实现细节

1) 数据预处理与数据增强。DMF 子网络以视频序列中的单幅表情峰值帧图像作为输入。首先使用该帧图像中的人脸关键点裁剪出人脸图像并缩放到 236 × 236 像素。对没有提供人脸关键点的峰值帧图像,使用 MTCNN 算法 (Zhang 等,2016) 检测其中的人脸关键点。此外,所有的人脸图像均进行了相应的直方图均衡化和全局对比度归一化处理。进一步,为了防止过拟合,在训练阶段,同时采用了在

线和离线的数据增强方法来扩充数据训练集中的数据样本。在离线增强阶段,分别使用 -10°、-5°、0°、5°、10°等 5 个角度对每幅图像进行旋转。在训练过程中,进一步通过在线增强扩充数据。一方面,从图像的 5 个位置(4 个角和中心)随机裁剪出 224 × 224 像素的图像块作为训练数据样本;另一方面,以 0.5 的置信度对图像进行随机水平翻转。最终,通过离线和在线数据增强处理,可以将原始数据集的大小扩充 50 倍。在测试阶段,仅将从图像中心裁剪出的 224 × 224 像素的一个图像块作为 DMF 子网络的输入。

2)三元组构造。对于 *N*-Metric 模块中计算三元组损失函数所需的三元组样本,通过批次难例挖掘策略(batch hard)构建(Hermans 等,2017),即对训练批次中的每个样本 a,可以找到最难的(与 a 特征距离最大)正样本以及最难的(与 a 特征距离最小)负样本,分别称为锚点、正样本和负样本,并以此来形成一个三元组。

3)DMF 子网络的优化训练。为了对 DMF 子网络进行单独训练,在 DMF 子网络的最后加了一个 softmax 表情分类层。因此,DMF 子网络可以以一种端到端的方式进行单独训练,其整体损失函数定义为

$$loss_{\text{tot}} = \lambda \sum_{i=1}^{N} loss_i + loss_0 \tag{3}$$

式中,$loss_0$ 对应用于表情分类的 softmax 损失函数,λ 用于控制不同种类损失函数之间的贡献权重。在实现中,$\lambda = 0.5/N$。为了训练得到具有较高泛化能力的 DMF 子网络,首先通过在人脸表情数据库 FER-2013(facial expression recognition-2013)(Goodfellow 等,2013)上微调 VGG16-Face 网络模型来预训练子网络中的 DCNN 模块。然后,针对每个实验数据集,对整个 DMF 子网络进行整体微调。训练采用 Adam 优化器,学习率设为 5E-6,批处理大小为 96,全连接层使用了 0.5 权重的 dropout 策略,整个子网络以一种端到端的方式训练 70 个 epoch。

4)*N*-Metric 模块中分支的确定。在 *N*-Metric 模块中,虽然可以在每条分支中使用不同维度的特征嵌入空间,但是在具体实现中使用了 256 维的固定大小。在三元组深度度量学习中,需要将来自不同类别的样本以较大的阈值 τ 间隔开来,但是如果阈值太大,会导致整个学习过程收敛较慢。因此,在实验中,将阈值 τ 的有效范围设置为[0.15,0.75]。然后,在该有效范围内,以一个给定的间隔对阈值 τ 进行采样,最终得到 $\{\tau_i\}_{i=1}^{N}$。对于采样间隔,如果间隔过小,则相邻阈值相差很小,因而这些相邻阈值对应的类间变化分布往往也差异较小,导致相邻阈值学习得到的特征之间缺乏可辨识度。此外,较小的采样间隔对应着更多的分支数,往往需要更多的训练消耗。另外,如果采样间隔过大,极有可能遗漏掉某些待学习的显著性类间变化分布。实验中使用采样间隔为 0.1,这是一个较为合理的采样间隔,它对应着 7 个采样阈值,分别为 0.15、0.25、0.35、0.45、0.55、0.65 和 0.75。其中,每个采样阈值与 *N*-Metric($N = 7$)模块中的一个分支相关联。表 1 给出了 DMF 子网络使用不同采样间隔在 3 个数据库上的性能差异。

表 1　3 个数据库上不同采样间隔的识别精度

Table 1　Recognition accuracy of different sampling intervals on three databases

/%

采样间隔	CK +	MMI	Oulu-CASIA
0.05($N = 13$)	97.54	77.93	81.67
0.1($N = 7$)	**97.86**	**78.09**	**83.54**
0.2($N = 4$)	97.25	75.42	81.88

注:加粗字体表示各列最优结果。

3.2.2　LTCNN 子网络实现细节

1)数据预处理。在实现中,使用 DAN(deep alignment network)算法(Kowalski 等,2017)检测采样图像中的 51 个人脸关键点。为了消除头部姿势及其大小对人脸关键点轨迹分析的影响,对人脸关键点的坐标进行归一化处理。具体方式为:对于每一个视频序列,可以以鼻子中心作为坐标原点,首先将每个关键点的位置坐标减去鼻子中心点的位置坐标,然后将该坐标除以所有采样帧中关键点位置坐标的标准方差。即

$$\bar{x}_i^t = \frac{x_i^t - x_c^t}{\sigma_x}, \bar{y}_i^t = \frac{y_i^t - y_c^t}{\sigma_y} \tag{4}$$

式中,(x_c^t, y_c^t) 为第 t 个采样帧中鼻子中心点的位置坐标,(σ_x, σ_y) 为该视频序列所有采样帧中关键点位置坐标的标准方差。

2)数据增强。为了防止 LTCNN 子网络在训练过程中发生过拟合,对人脸关键点进行随机水平翻转,并在关键点位置坐标中添加随机高斯噪声。即

$$\bar{x}_i^t = \bar{x}_i^t + z_i^t \tag{5}$$

式中，$z_i^t \sim N(0,\sigma_i^2)$ 表示在第 t 个采样帧的第 i 个关键点的 x 坐标上添加的噪声，设置 $\sigma_i^2 = 0.01$ 。

3）LTCNN 子网络的优化训练。与 DMF 子网络类似，为了对 LTCNN 子网络进行单独训练，在 LTCNN 子网络的最后加了一个 softmax 表情分类层。在实现中，LTCNN 子网络前 4 个卷积层的大小分别为 $3\times15\times64$、$3\times11\times96$、$3\times7\times128$ 和 $3\times3\times128$。其中，$3\times15\times64$ 表示使用了 64 个 3×15 大小的 2D 卷积核，其他卷积层大小的含义一样。对于 LTCNN 子网络中的后两个全连接层，分别使用了 512 和 128 个神经元。训练时，使用 Xavier 初始化整个子网络，再采用 Adam 优化器进行优化，设置权重衰减率为 0.000 1，初始学习率、批处理大小以及训练周期分别为 1.0E－4、96 和 70。

3.3 表情识别性能的分析与评估

3.3.1 DMF 子网络中多分支的特征可视化

在 DMF 子网络的 *N*-Metric 模块中，使用了 7 条分支通过基于三元组的深度度量学习来学习得到不同的人脸表情特征。图 8 给出了不同分支上学习特征的可视化结果。其中，第 2—8 列为各分支上的特征，最后 1 列为所有分支融合而成的特征。每个特征通过与其关联的全连接层中的神经元进行可视化，其中 1 个小方格对应着 1 个神经元，且颜色越亮代表值越大。特别说明，对于融合特征，显示了它对应的所有 256 个神经元，而对于各分支的特征，为了清晰显示，仅从其中的 512 个神经元中均匀采样了 64 个神经元进行显示。从图 8 可以看出，1）对于同一幅人脸图像，各个分支上的特征具有各不相同的可分辨特性；2）对于具有相同表情的不同个体图像，每一分支上的表情特征极其相似，而对于同一个体下的不同表情图像，每一分支上的表情特征则相差较大。

综上分析，每条分支显然学习到了不同的特征表示并且对表情具有极强的分辨性。最终，将这 7 条分支上的特征组合在一起，可以得到一个更加鲁棒且更具识别能力的表情“空域特征”。

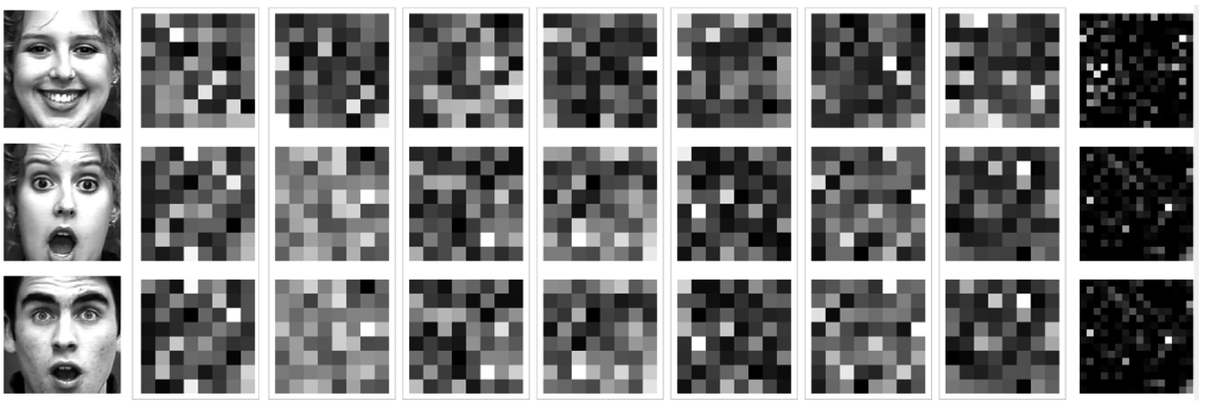


图 8 DMF 子网络中不同分支上的特征可视化结果

Fig. 8 Visualization results of features on different branches in DMF sub-net((a) original images; (b) $\tau_1 = 0.15$; (c) $\tau_2 = 0.25$; (d) $\tau_3 = 0.35$; (e) $\tau_4 = 0.45$; (f) $\tau_5 = 0.55$; (g) $\tau_6 = 0.65$; (h) $\tau_7 = 0.75$; (i) fusion features)

3.3.2 DMF 子网络中单分支与多分支模型的对比

为了进一步验证 DMF 子网络中多分支模型的有效性，仅保留了 DMF 子网络中的一条分支，并分别使用不同的阈值来训练该单分支的 DMF 网络模型。表 2 给出了不同阈值下该单分支 DMF 网络的性能结果。可以看出，模型的识别性能随着阈值的改变发生了相应变化，并且对于不同的数据库，其最佳阈值有所不同，这也验证了前述的观察结果，即通过改变损失函数中的阈值可以学习到不同的表情特征。此外，结果还表明，在 CK＋、MMI 和 Oulu-CASIA 数据库上，即使采用最佳阈值，单阈值方法的性能也比多阈值融合的方法要低，分别低约 1.31%、4.42% 和 2.33%，这充分证明了深度度量融合技术的优势。

3.3.3 两种不同的关键点轨迹特征图

在基于关键点轨迹构造 LTCNN 子网络的输入特征图时，可以采用单通道或双通道的特征图方式，它们分别对应 LTCNN-1CL 和 LTCNN-2CL。表 3 给

表 2　单分支 DMF 网络在不同阈值的识别精度

Table 2　Recognition accuracy of single-branch DMF network at different thresholds

/%

阈值	CK +	MMI	Oulu-CASIA
0.15	96.02	70.65	80.43
0.25	95.11	72.97	80.74
0.35	96.33	73.26	81.15
0.45	95.71	73.57	80.63
0.55	96.33	71.68	**81.21**
0.65	**96.55**	**73.67**	80.56
0.75	95.72	71.56	81.13

注：加粗字体表示各列最优结果。

表 3　LTCNN-1CL 和 LTCNN-2CL 的识别精度

Table 3　Recognition accuracy of LTCNN-1CL and LTCNN-2CL

/%

特征图构造	CK +	MMI	Oulu-CASIA
LTCNN-1CL	94.87	75.11	80.42
LTCNN-2CL	**96.16**	**75.46**	**81.88**

注：加粗字体表示各列最优结果。

出了对应的表情识别性能结果。可以发现，在 3 个数据库上，LTCNN-2CL 均取得了比 LTCNN-1CL 更高的准确率。即 LTCNN-2CL 对应的特征图能够更加准确地提取出关键点轨迹中的运动信息。因此，本文其他所有的相关实验均采用双通道的关键点轨迹特征图作为 LTCNN 子网络的输入。

3.3.4　不同融合策略的性能对比

针对 DMF 和 LTCNN 子网络的融合，表 4 给出了不同策略融合下的表情识别性能。可见，基于整体微调的特征融合方法有效实现了 DMF 和 LTCNN 子网络的互补融合，取得了最高的表情分类精度。

此外，表 4 给出了单独 DMF 子网络和单独 LTCNN 子网络的人脸表情识别精度。显然，通过充分结合表情的时域和空域特征信息，融合时空域特征的人脸表情识别方法取得了表情识别性能的显著提升。需要注意的是，在表 4 中，一般特征融合策略取得的识别精度甚至低于单独使用 DMF 或 LTCNN 子网络的识别精度。这是因为 DMF 和 LTCNN 子网络在学习过程中的收敛速度不同，而以统一的学习率对它们进行端到端的训练无法充分照顾它们的不同收敛特性。

表 4　不同融合策略的识别精度

Table 4　Recognition accuracy of different fusion strategies

/%

方法	CK +	MMI	Oulu-CASIA
DMF	97.86	78.09	83.54
LTCNN	96.16	75.46	81.88
决策融合(加权平均)	97.71	79.75	83.96
决策融合(决策排序)	98.12	80.87	85.42
特征融合	94.71	76.25	79.28
微调特征融合(局部微调)	98.06	78.71	82.63
微调特征融合(固定 DMF)	98.15	82.74	86.04
微调特征融合(固定 LTCNN)	98.03	82.22	84.15
微调特征融合(整体微调)	**98.46**	**82.96**	**87.12**

注：加粗字体表示各列最优结果。

3.3.5　与之前方法的性能比较

表 5 给出了本文方法与其他已有方法的性能对比。在这些已有方法中，DTAGN (deep temporal appearance-geometry network) 通过局部微调的融合方式集成两个子网络学习到的时序外观特征和时序几何特征(Jung 等，2015)。PHRNN-MSCNN 通过决策排序融合的方式集成不同网络学习到的表情时空信息(Zhang 等，2017)。从表 5 可以看出，通过整体微调，本文提出的融合时空域特征的人脸表情识方法取得了较好的性能提升。表 5 进一步给出了 PHRNN-MSCNN 中时域和空域特征子网络各自的表情识别性能。可以看出，1) 相比于 MSCNN 子网络，提出的 DMF 空域特征子网络在 3 个数据库上均取得了明显的性能提升；2) 提出的 LTCNN 时域特征子网络取得了与 PHRNN 子网络较接近的识别性能，但是提出的基于 CNN 的网络结构避免了 RNN 网络结构可能带来的网络训练难度。最近，LBVC-NN(local binary volume convolutional neural network) 通过局部二值体卷积神经网络可以从视频序列的 3 个正交面同时学习其中的时空局部纹理信息(Kumawat 等，2019)，与之相比，本文提出的时空融合网络用专门的子网络分别专注于学习时域信息和空域信息，然后再进行互补融合，取得了更高的表情识别性能。

表 5　不同方法的识别精度

Table 5　Recognition accuracy of different methods

/%

方法	CK +	MMI	Oulu-CASIA
3DCNN(Liu 等, 2014)	85.90	53.20	N/A
3DCNN-DAP(Liu 等, 2014)	92.4	63.40	N/A
DTAGN(Jung 等, 2015)	97.25	70.24	81.46
Enhanced-3DCNN (Hasani 和 Mahoor, 2017)	95.53	79.26	N/A
MSCNN(Zhang 等, 2017)	95.54	77.07	77.67
PHRNN(Zhang 等, 2017)	96.36	76.17	78.96
PHRNN-MSCNN (Zhang 等, 2017)	**98.50**	81.18	86.25
L2-sparseness(Xie 等, 2019)	97.59	78.54	82.92
G2-VER(Tanguy 等, 2019)	97.4	N/A	N/A
LBVCNN(Kumawat 等, 2019)	97.38	N/A	82.41
DMF	97.86	78.09	83.54
LTCNN	96.16	75.46	81.88
整体微调融合	98.46	**82.96**	**87.12**

注:加粗字体表示各列最优结果,N/A 表示无对应数据。

表 6—表 8 分别显示了基于整体微调融合的时空网络在 3 个表情数据集上的混淆矩阵。可以看出,在 CK + 数据集上,本文方法对于每个类别均具有较好的识别性能。对于更具挑战性的 MMI 数据集,由于恐惧与惊讶两种表情较为相似,它们对应的面部关键点的轨迹运动差别较为细微,使得较多数量的恐惧表情错误地识别为惊讶,最终造成恐惧类别的识别率较低。对于 Oulu-CASIA 数据集,本文方法在所有类别上取得了较为均衡的识别性能,并且在生气和惊讶两种表情上取得了最高的识别率。

表 6　本文方法在 CK + 数据集上的混淆矩阵

Table 6　Confusion matrix of this method on CK + dataset

/%

	生气	藐视	厌恶	恐惧	高兴	悲伤	惊讶
生气	**98.13**	1.87	0	0	0	0	0
藐视	0	**100**	0	0	0	0	0
厌恶	0	0	**100**	0	0	0	0
恐惧	0	0	0	**95**	2	0	3
高兴	0	0	0	0	**100**	0	0
悲伤	3.48	0	0	0	0	**96.52**	0
惊讶	0	1.98	0	0	0	0	**98.02**

注:加粗字体为本文方法在各类表情类别中的最高识别精度。

表 7　本文方法在 MMI 数据集上的混淆矩阵

Table 7　Confusion matrix of this method on MMI dataset

/%

	生气	厌恶	恐惧	高兴	悲伤	惊讶
生气	**76.16**	7.05	0	6.22	8.32	2.25
厌恶	13.54	**77**	0	9.46	0	0
恐惧	7.03	0	**50.73**	13.67	10.25	18.32
高兴	0	2	2	**96**	0	0
悲伤	4.12	7.51	5	4.05	**79.32**	0
惊讶	0	0	2	2	0	**96**

注:加粗字体为本文方法在各类表情类别中的最高识别精度。

表 8　本文方法在 Oulu-CASIA 数据集上的混淆矩阵

Table 8　Confusion matrix of this method on Oulu-CASIA dataset

/%

	生气	厌恶	恐惧	高兴	悲伤	惊讶
生气	**92.5**	1.25	1.25	2.75	2.25	0
厌恶	10.75	**80.25**	4.5	4.5	0	0
恐惧	5	1.25	**86**	1.25	4	2.5
高兴	0	0	6	**89**	5	0
悲伤	6.75	7.5	4.5	0	**81.25**	0
惊讶	0	0	8	0	0	**92**

注:加粗字体为本文方法在各类表情类别中的最高识别精度。

4　结　论

针对基于视频序列的人脸表情识别,本文提出了一种融合时空域特征的深度学习神经网络。首先,提出了一种基于三元组的深度度量融合技术,通过采用不同的三元组阈值,从单幅表情峰值图像中学习得到多个不同的表情特征表示,并将它们组合在一起最终形成了一个鲁棒的且更具识别能力的表情“空域特征”。然后,基于视频序列中的人脸关键点轨迹特征图,使用简单的 2 维卷积神经网络,学习得到描述表情运动信息的表情“时域特征”。最后,提出一种基于整体微调的网络融合策略,取得了最优的时域特征和空域特征的融合效果。

在 3 个公开且广泛使用的表情数据集 CK + 、MMI 和 Oulu-CASIA 上验证了本文算法的有效性。实验结果表明,本文方法取得了显著的性能提升,在 3 个数据集上均接近或超越了当前最高的人脸表情识别性能。但本文方法仍有一些不足之处,未来可以通过以下几方面进一步研究:1)提出的方法仅考

虑了视频和图像两种模态下的人脸表情识别,未来可以融合更多模态的特征,例如主体的身份信息、场景描述信息和语音信息等,以进一步增强表情识别算法的鲁棒性。此外,未来还计划将三元组深度度量融合技术推广到其他相关应用,例如图像分类、图像搜索以及可视对象识别等。2)本文方法只探究了几种模型融合策略来结合时序和空间特征。未来可以尝试其他融合方法,更好地利用各个子网络中的互补信息。也可以对最新提出的 3D 卷积进行改进,在利用 3D 卷积联合学习时空特征优势的同时,降低 3D 卷积网络的复杂性。3)许多研究通常在特定的数据库上评估算法性能,但是一些跨数据库实验表明,由于数据的采集方式和环境不同,数据库之间普遍存在数据偏差和注释不一致的问题,这将大幅降低在未知数据上的泛化性能。深度域适应和知识蒸馏是解决数据偏差的可行解决方案。未来可以将研究扩展到跨数据库的人脸表情识别问题上。

参考文献(References)

Acharya D, Huang Z W, Paudel D P and van Gool L. 2018. Covariance pooling for facial expression recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, USA: IEEE: 480-487 [DOI: 10.1109/CVPRW.2018.00077]

Deng D D, Chen Z K, Zhou Y Q and Shi B. 2019. MIMAMO Net: integrating micro-and macro-motion for video emotion recognition [EB/OL]. [2020-12-14]. https://arxiv. org/pdf/1911. 09784. pdf

Ekman P and Friesen W V. 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129 [DOI: 10.1037/h0030377]

Goodfellow I J, Erhan D, Carrier P L, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y C, Thaler D, Lee D H, Zhou Y B, Ramaiah C, Feng F X, Li R F, Wang X J, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie J J, Romaszko L, Xu B, Chuang Z and Bengio Y. 2013. Challenges in representation learning: a report on three machine learning contests//Proceedings of the 20th International Conference on Neural Information Processing. Daegu, Korea (South): Springer: 117-124 [DOI: 10.1007/978-3-642-42051-1_16]

Gutierrez G. 2020. Artificial intelligence in the intensive care unit. Critical Care, 24(1): #101 [DOI: 10.1186/s13054-020-2785-y]

Hasani B and Mahoor M H. 2017. Facial expression recognition using enhanced deep 3D convolutional neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE: 2278-2288 [DOI: 10.1109/CVPRW.2017.282]

Hermans A, Beyer L and Leibe B. 2017. In defense of the triplet loss for personre-identification [EB/OL]. [2020-12-14]. https://arxiv.org/pdf/1703.07737.pdf

Jain S, Hu C B and Aggarwal J K. 2011. Facial expression recognition with temporal modeling of shapes//Proceedings of 2011 IEEE International Conference on Computer Vision Workshops. Barcelona, Spain: IEEE: 1642-1649 [DOI: 10.1109/iccvw.2011.6130446]

Jung H, Lee S, Yim J, Park S and Kim J. 2015. Joint fine-tuning in deep neural networks for facial expression recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2983-2991 [DOI: 10.1109/ICCV.2015.341]

Klaser A, Marszalek M and Schmid C. 2008. A spatio-temporal descriptor based on 3D-gradients//Proceedings of the British Machine Conference. [s. l.]: BMVC: #99 [DOI: 10.5244/C.22.99]

Kowalski M, Naruniec J and Trzcinski T. 2017. Deep alignment network: a convolutional neural network for robust face alignment//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 2034-2043 [DOI: 10.1109/CVPRW.2017.254]

Kumawat S, Verma M and Raman S. 2019. LBVCNN: local binary volume convolutional neural network for facial expression recognition from image sequences//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE: 207-216 [DOI: 10.1109/cvprw.2019.00030]

Li S and Deng W H. 2020. Deep facial expression recognition: a survey. IEEE Transactions on Affective Computing: #2981446 [DOI: 10.1109/TAFFC.2020.2981446]

Liu M Y, Li S X, Shan S G, Wang R P and Chen X L. 2014. Deeply learning deformable facial action parts model for dynamic expression analysis//Proceedings of the 12th Asian Conference on Computer Vision. Singapore, Singapore: Springer: 143-157 [DOI: 10.1007/978-3-319-16817-3_10]

Liu X F, Kumar B V K V, You J and Jia P. 2017. Adaptive deep metric learning for identity-aware facial expression recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 522-531 [DOI: 10.1109/cvprw.2017.79]

Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z and Matthews I. 2010. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, USA: IEEE: 94-101 [DOI: 10.1109/cvprw.2010.5543262]

Pantic M, Valstar M, Rademaker R and Maat L. 2005. Web-based database for facial expression analysis//Proceedings of 2005 IEEE International Conference on Multimedia and Expo. Amsterdam, the Netherlands: IEEE: 317-321 [DOI: 10.1109/icme.2005.1521424]

Parkhi O M, Vedaldi A and Zisserman A. 2015. Deep face recognition//Proceedings of the British Machine Vision Conference. Swansea,

UK: BMVA Press: #41

Ptucha R, Tsagkatakis G and Savakis A. 2011. Manifold based sparse representation for robust expression recognition without neutral subtraction//Proceedings of 2011 IEEE International Conference on Computer Vision Workshops. Barcelona, Spain: IEEE: 2136-2143 [DOI: 10.1109/iccvw.2011.6130512]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/tpami.2016.2577031]

Schroff F, Kalenichenko D and Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 815-823 [DOI: 10.1109/CVPR.2015.7298682]

Scovanner P, Ali S and Shah M. 2007. A 3-dimensional sift descriptor and its application to action recognition//Proceedings of the 15th ACM International Conference on Multimedia. Augsburg, Germany: ACM: 357-360 [DOI: 10.1145/1291233.1291311]

Shelhamer E, Long J and Darrell T. 2017. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651 [DOI: 10.1109/TPAMI.2016.2572683]

Sikka K, Sharma G and Bartlett M. 2016. LOMo: latent ordinal model for facial analysis in videos//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5580-5589 [DOI: 10.1109/cvpr.2016.602]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. [EB/OL]. [2020-12-14]. https://arxiv.org/pdf/1409.1556.pdf

Taini M, Zhao G Y, Li S Z and Pietikainen M. 2008. Facial expression recognition from near-infrared video sequences//Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA: IEEE: 1-4 [DOI: 10.1109/icpr.2008.4761697]

Tanguy A, Mandana F, Saleh B S and Guillaume V. 2019. G2-VER: geometry guided model ensemble for video-based facial expression recognition//Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). [s. l.]: [s. n.]: 1-6 [DOI: 10.1109/FG.2019.8756600]

Vinciarelli A, Pantic M and Bourlard H. 2009. Social signal processing: survey of an emerging domain. Image and Vision Computing, 27(12): 1743-1759 [DOI: 10.1016/j.imavis.2008.11.007]

Wang Z H, Wang S F and Ji Q. 2013. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 3422-3429 [DOI: 10.1109/cvpr.2013.439]

Whitehill J, Serpell Z, Lin Y C, Foster A and Movellan J R. 2014. The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing, 5(1): 86-98 [DOI: 10.1109/TAFFC.2014.2316163]

Xie W C, Jia X, Shen L L and Yang M. 2019. Sparse deep feature learning for facial expression recognition. Pattern Recognition, 96: #106966 [DOI: 10.1016/j.patcog.2019.106966]

Yang H Y, Ciftci U and Yin L J. 2018. Facial expression recognition by de-expression residue learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2168-2177 [DOI: 10.1109/CVPR.2018.00231]

Zeng Z H, Pantic M, Roisman G I and Huang T S. 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1): 39-58 [DOI: 10.1109/TPAMI.2008.52]

Zhang K H, Huang Y Z, Du Y and Wang L. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing, 26(9): 4193-4203 [DOI: 10.1109/TIP.2017.2689999]

Zhang K P, Zhang Z P, Li Z F and Qiao Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10): 1499-1503 [DOI: 10.1109/LSP.2016.2603342]

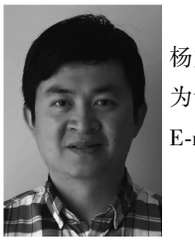
Zhao G Y, Huang X H, Taini M, Li S Z and Pietikäinen M. 2011. Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619 [DOI: 10.1016/j.imavis.2011.07.002]

Zhao G Y and Pietikainen M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6): 915-928 [DOI: 10.1109/tpami.2007.1110]

Zhao J F, Mao X and Zhang J. 2018. Learning deep facial expression features from image and optical flow sequences using 3D CNN. The Visual Computer, 34(10): 1461-1475 [DOI: 10.1007/s00371-018-1477-y]

作者简介

陈拓,1993 年生,男,硕士研究生,主要研究方向为表情识别和人脸识别。
E-mail: 17020100023@pop.zjgsu.edu.cn



杨文武,通信作者,男,教授,主要研究方向为计算机动画、计算机视觉、计算机图形学。
E-mail: wwyang@zjgsu.edu.cn

邢帅,男,硕士研究生,主要研究方向为 3D 人体动作检测与跟踪。E-mail: xs_xingshuai@126.com

金剑秋,男,副教授,主要研究方向为计算机视觉与机器学习。E-mail: jqjin@mail.zjgsu.edu.cn