计算机科学
COMPUTER SCIENCE

融合三维人脸动态信息和光流信息的人脸表情识别

张华忠, 潘曰凯, 涂晓光, 刘建华, 许罗鹏, 周超

引用本文

张华忠, 潘曰凯, 涂晓光, 刘建华, 许罗鹏, 周超. 融合三维人脸动态信息和光流信息的人脸表情识别[J]. 计算机科学, 2024, 51(6A): 230700210-7.

ZHANG Huazhong, PAN Yuekai, TU Xiaoguang, LIU Jianhua, XU Luopeng, ZHOU Chao. Facial Expression Recognition Integrating 3D Facial Dynamic Information and Optical Flow Information [J]. Computer Science, 2024, 51(6A): 230700210-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700210

融合三维人脸动态信息和光流信息的人脸表情识别

张华忠　潘曰凯　涂晓光　刘建华　许罗鹏　周　超
中国民用航空飞行学院航空电子电气学院　四川 广汉 618300

摘　要　人脸表情识别在静态图像上取得了卓越的成效,但这些方法应用于视频或图像序列时,准确度和鲁棒性往往会受到影响。传统的方法通常无法基于空间信息和光流信息进行人脸表情的识别,然而这些辅助识别信息都是二维信息,没有考虑到人脸的表情变化是一种三维的变化过程。为充分挖掘人脸表情识别的深层语义信息,提出了一种基于三维人脸动态信息和光流信息相结合的融合表情识别方法。该方法构建基于人脸深度图像、光流图像和RGB图像的多流卷积神经网络,通过融合3种模态的信息进行人脸表情识别。所提方法在CAER,RAVDESS数据集上进行了充分验证,实验结果表明,其在表情识别性能上优于目前的主流方法,证明了其有效性。

关键词:表情识别;多流卷积神经网络;三维人脸动态信息;光流信息
中图分类号　TP391.41

Facial Expression Recognition Integrating 3D Facial Dynamic Information and Optical Flow Information

ZHANG Huazhong,PAN Yuekai,TU Xiaoguang,LIU Jianhua,XU Luopeng and ZHOU Chao
Institute of Electronic and Electrical Engineering,Civil Aviation Flight University of China,Guanghan,Sichuan 618300,China

Abstract　Facial expression recognition has achieved excellent results in static images,but when these methods are applied to videos or image sequences,their accuracy and robustness are often affected. Traditional methods cannot usually recognize facial expressions based on spatial information and optical flow information. However,these auxiliary recognition information are all two-dimensional information,without considering that facial expression changes are a three-dimensional change process. In order to fully mine the deep semantic information of facial expression recognition,this paper proposes a fusion expression recognition method based on the combination of 3D facial dynamic information and optical flow information. This method constructs a multi stream convolutional neural network based on facial depth images,optical flow images,and RGB images,and integrates information from three modalities for facial expression recognition. The proposed method has been fully validated on CAER and RAVDESS datasets,and experimental results show that it outperforms current mainstream methods in facial expression recognition performance,which proves its effectiveness.

Keywords　Facial expression recognition,Multi-stream convolutional neural network,3D facial dynamic information,Optical flow information

表情识别(Facial Expression Recognition,FER)是情感识别的一个重要组成部分。面部表情是传递人类情绪信息的重要媒介,在人与人之间进行交流的过程中,人们可以通过控制面部表情来提高沟通的效果。情绪生理学家运用生理、生化方法测量探索情绪的生理机制,发现人的情绪变化直接影响着面部肌肉的变化,进而造成表情的变化。如果能有效地捕捉到人脸面部表情变化,就能有效地判断其情绪,进而推测出其心理活动。使用图像处理、计算机视觉、人工智能领域的算法捕捉人脸面部的表情变化,该技术应用范围较广,包括但不限于安全、汽车、机器人制造、医疗、通信领域等。

根据研究对象以及网络的数据输入形式不同,我们还可以将表情识别分为动态表情识别(Dynamic Expression Recognition,DER)和静态表情识别(Static Expression Recognition,SER)。静态表情识别是对静态的单一视频帧进行研究以判别其情感类别;而动态表情识别的研究对象一般为一段视频或者是一系列的视频帧,动态视频表情识别注重的是表情之间的连贯性,需要把握表情产生的整个过程,而不是只对其中的一帧做出判断。

已有研究者开始利用基于光流的方法从微表情视频或序列中提取运动相关信息,因为光流可以推断出不同帧之间的

基金项目:中国博士后科学基金(2022M722248);中央高校基本科研业务费(J2023-026,ZHMH2022-004);民航飞行技术与飞行安全重点实验室开放项目资助(FZ2022KF06);民航飞行技术与飞行安全重点实验室自主项目(FZ2021ZZ03)
This paper was supported by the China Postdoctoral Science Foundation(2022M722248),Project of Basic Scientific Research of Central Universities of China(J2023-026,ZHMH2022-004),Open Fund of Key Laboratory of Flight Techniques and Flight Safety,CAAC(FZ2022KF06) and Fund of Key Laboratory of Flight Techniques and Flight Safety,CAAC(FZ2021ZZ03).
通信作者:张华忠(zhz_233@yeah.net)

相对运动信息。Xu 等[1]采用基于光流的人脸动态图(Facial Dynamics Model,FDM)方法,在微表情识别方面取得了良好的准确性。Ma 等[2]根据人脸的关键点坐标和人脸编码系统将人脸划分为多个区域,在每个区域内,根据两帧之间的特征进行表情识别,该方法在识别率上优于其他先进的算法。Wang 等[3]提出一种结合人脸关键点与光流特征的微表情识别方法,建立一个人脸关键点和光流的双输入网络模型,并进行了实验,验证了该方法与其他主流算法相比具有更好的识别性能。

Simonyan 等[4]于 2014 年首次提出了一个包含空间和时间网络的双流卷积神经架构,用于视频中的行为识别,相比其他网络获得了更高的准确率。研究还证实了在有限训练数据的情况下,基于多帧密集光流进行训练的卷积神经网络能够取得出色的性能。2016 年,Fernando 等[5]提出了一个新的网络,该网络可以将学习公式化为双层优化问题,以端到端的方式联合估算模型的分类参数。2018 年,Zolfaghari 等[6]提出 ECO 网络,将视频帧首先通过二维卷积,然后再进行三维卷积,同时利用空时信息。Wang 等[7]提出了双流网络行动识别网络,通过设计空间和时间网络的预训练、更小的学习率和更多的数据增强技术等训练方法,在 UCF101 的数据集上验证了双流神经网络集的性能,并达到了 91.4%的识别精度。Aghamaleki 等[8]提出了手工制作的特征和多流结构,作为通过 CNN 提升有限数据性能的一种解决方案,包括三流和单流两种不同的结构,采用三线结构可以提高面部表情分类器的识别率。Zhu 等[9]采用拟合三维人脸来代替直接定位特征点,然后根据三维人脸进行特征点的标记。Lee 等[10]提出了情境感知的表情识别深度网络 CAER-Net,其不仅利用人脸的面部表情,还采用注意力机制来利用其他情境信息,在定性和定量上都比现有的基准测试效果更好。Zhang 等[11]为了能够进一步地提高表情识别的准确率,提出了用于面部表情识别(FER)的多模态学习方法,该方法主要利用了彼此互补的面部图像的纹理和界标模态,并且充分考虑了面部表情的全面性,得到了更全面的情感信息。

Feng 等[12]基于双流的框架提出了一种新的用于动态表情识别的方法,该方法与传统的双流方法不同。在空间网络,两者都是从每个视频选择一帧并放入 CNN 网络中进行训练。但在时间网络中,该方法使用的不是光流图,而是利用 LBP-TOP 特征作为时间网络的输入,并结合帧的特征作为序列通道来训练卷积神经网络。Zhou 等[13]提出了一种用于人脸篡改检测的双流卷积神经网络,训练此网络来检测人脸识别中的篡改伪影,并在新收集的数据集中进行了验证。

在计算机视觉领域,图像的深度信息已经被证明是一种有效信息,可以用来进行辅助识别。Lu 等[14]提出了一种颜色和深度图像融合算法,该算法结合了彩色图像和深度图像的多尺度特征,对比结果表明,该算法的准确性和适应性得到显著提高。Xing 等[15]基于单个深度图的三维手部姿态估计提出了一种具有动态锚点的动态关系网络(DRN)。Jiang 等[16]在没有对噪声统计做出任何明确假设的情况下,提出了一种新的无监督图像去噪框架,该方法建立在深度图像先验(DIP)的基础上,能够实现不同的图像恢复任务。Niu 等[17]提出了一种基于三维人脸数据和二维人脸图像匹配的人脸识别算法,设计了一个基于深度图像和曲率的改进残差神经网络进行人脸识别。在人脸表情识别中使用深度图(Depth Map,DM)的主要优势在于,深度图能够直接反映物体可见表面的几何形状。深度图像可以提供面部表情更细致的三维形状信息,从而更好地反映面部肌肉的形变和位置变化。这些细致的形状信息对于人脸表情识别任务至关重要,可以更好地捕捉面部表情的细微差异。

通过以上分析,本文提出了一种融合三维人脸空间信息和光流信息的表情识别方法。首先对原始视频进行预处理,然后使用 TV-L[1]光流法提取视频中运动信息的水平和垂直方向的光流特征序列。同时使用三维人脸重建中的算法获取人脸深度图,构建多流卷积神经网络来进行信息融合,从而优化动态视频表情识别性能。

1 方法

1.1 人脸检测

只提取人脸部分可以降低输入数据的复杂性并减少冗余信息,将注意力集中在人脸部分可以提高表情识别算法的准确性。由于人脸表情主要通过面部肌肉的动态变化来表达,将关注点限定在人脸区域可以更好地捕捉和分析这些微小的变化。相比处理整张图像,仅关注人脸部分可以减少计算量,提高表情识别算法的运行效率,从而实现实时或高速的表情识别应用。

本文使用 Dlib 人脸检测器提取 RGB 图像中的人脸,加载预训练的人脸检测器模型并读取待检测的图像,对待检测图像进行预处理来增强检测的准确性,能够快速且准确地对人脸图像进行处理和分析,如图 1 所示。

图 1 基于 RGB 图像的人脸检测

Fig. 1 Face detection based on RGB images

1.2 光流图像

光流(Optical Flow,OF)指在观察成像平面上表示空间运动物体像素运动瞬时速度的一种形式。它反映了图像的变化情况,并提供了目标运动的信息。通过比较视频的当前帧和上一帧,光流可以用来确定目标的运动情况,并计算出相邻帧之间物体的运动信息。光流是一种常用的方法,用于分析和理解视频中物体的动态变化。

亮度恒定的情况下,假设一个像素点在 t 时刻的位置为 (x,y),$I(x,y,t)$ 表示这一帧的光强度。此像素点在 dt 时间内移动了 (dx,dy) 的距离到下一帧。因为是同一个像素点,依据上文提到的假设,该像素在运动前后的光强度是不变的,即:

$$I(x,y,t)=I(x+dx,y+dy,t+dt) \tag{1}$$

对式(1)右侧进行泰勒展开,可得:

$$I(x+dx,y+dy,t+dt)=I(x,y,t)+\frac{\partial I}{\partial x}dx+\frac{\partial I}{\partial y}dy+\frac{\partial I}{\partial t}dt+\varepsilon \tag{2}$$

由式(1)、式(2)可得:

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t}+\frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t}+\frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t}=0 \tag{3}$$

设 m 为水平方向的速度，n 为垂直方向的速度，则：

$$m=\frac{\Delta x}{\Delta t},n=\frac{\Delta y}{\Delta t} \tag{4}$$

则式(3)可以表示为：

$$I_x m+I_y n+I_t=0 \tag{5}$$

其中，I_x，I_y，I_t 可由图像数据求得，使用表情光流估计矢量 (m,n) 描述图像上每个像素的运动大小和方向，实现对表情的分析和识别。

TV-L^1 方法是一种常用的光流计算方法，利用总变差(Total Variation，TV)的概念来求解光流。总变差用于度量函数变化的程度，进而评估图像的平滑性。该方法通过最小化两个图像之间像素强度差异的总变差范数，解决一个最优化问题，从而获得光流图。具体而言，TV-L^1 方法采用了 L^1 范数来约束正则化项，其能量函数可以表示为：

$$E(u)=\int_{\Omega}[\alpha|Du|^2+\beta|u-f|^2+\gamma|Du-w|^2]\mathrm{d}\Omega \tag{6}$$

其中，u 是光流场，f 是当前帧图像，w 是时间平滑的光流场，D 是导数算子，α，β 和 γ 是权重参数。该能量函数包含 3 个项：第一个项为二阶导数关于的项，用于约束光流场的平滑性；第二个项为光流场和当前帧图像的灰度值差异项，用于约束光流场的准确性；第三个项为光流场和平滑后的光流场的差异项，用于约束光流场的一致性。这种能量函数的设计能够综合考虑平滑度、准确性和一致性，以实现更好的光流估计结果。

通过对式(6)能量函数求导，可以得到 TV-L^1 方法的求解公式，即：

$$\frac{\partial u}{\partial t}=-\mathrm{div}\left(\frac{Du}{|Du|+\varepsilon}\right)+\frac{u-f}{\beta}-\gamma(Du-w) \tag{7}$$

其中，div 是散度算子，ε 是一个较小的正数，用于避免分母为 0 的情况。该公式通过迭代求解，不断更新光流场的值，直到能量函数收敛为止。

在人脸表情识别中，使用 TV-L^1 方法获取光流图像能够捕捉到人脸表情在时间维度上的动态信息。这种方法对于每两帧之间的光流通量进行计算，从而得到一系列的光流图像，包括 x 和 y 方向的光流图。通过对这些光流图像进行处理，如密度估计等操作，可以进一步提取人脸表情的特征信息。在实际操作中，我们从视频库中随机选择 $m+1$ 张连续帧，然后利用 TV-L^1 方法计算这些帧之间的光流图像，得到 $2m$ 张光流图，其中包括 x 方向和 y 方向的光流图。最后，将这 $2m$ 张光流图按照时间顺序进行堆叠，形成堆叠光流图，其中通道数为 $2m$。如图 2 所示，这样的处理方式可以更好地利用时间信息，从而提升人脸表情识别的性能。

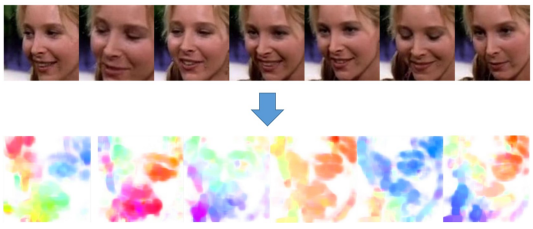

图 2　视频帧生成的光流图像

Fig. 2　Optical flow generated from video frame

1.3　深度图像

三维稠密人脸对齐(3D Dense Face Alignment，3DDFA)可用于 3D 人脸重建和姿态估计的深度学习模型。它可以通过输入一张 RGB 图像，输出对应的 3D 人脸形状、姿态和纹理信息。其中，深度图像是 3DDFA 模型的输出之一。

使用 3DDFA 模型获取深度图，首先需要对输入的图像进行预处理，包括对人脸进行检测和对齐，先选取人脸图片中的 62 个特征点，其中包括 12 个姿势特征点、40 个形状特征点和 10 个表情特征点，这些特征点可用于后续的模型拟合和深度回归网络(Depth Regression Network，DRN)训练。然后，将预处理后的图像输入 3DDFA 模型中，通过深度回归网络，将 2D 人脸图像映射到 3D 人脸模型，获得一个初步的 3D 人脸模型。利用深度回归网络对初步的 3D 人脸模型进行细化和修正，得到更加准确的 3D 人脸模型参数。最后利用人脸模型参数和相机参数(如相机内参和外参)计算出每个 3D 点到相机的距离，即深度值。根据深度值将 3D 人脸模型投影到 2D 平面，得到相应的深度图。在深度图中，像素值表示对应点到相机的距离。

3DDFA 使用的是基于深度图的 3D 可变形模型(3D Morphable Model，3DMM)，其中包含许多 3D 人脸形状，每个 3D 人脸形状可以由基础模型加上一系列形状变化参数来生成。3DMM 模型是由大量的 3D 人脸数据集构建的一个平均模型，包括形状模型和纹理模型。输入是 2D 人脸图像，输出是 3DMM 模型的形状参数和表情参数。形状参数指用于描述 3D 人脸形状变化的参数；表情参数指用于描述表情变化的参数。形状模型是一个线性模型，可以通过主成分分析(Principal Component Analysis，PCA)降维来描述 3D 人脸的形状变化，具体的公式可以表示为：

$$y=\boldsymbol{W}^{T(x-\mu)} \tag{8}$$

$$x=\boldsymbol{W}y+\boldsymbol{\mu} \tag{9}$$

其中，x 为原始数据，y 为降维后的数据，$\boldsymbol{W}$ 为主成分分析得到的转换矩阵，$\boldsymbol{\mu}$ 为均值向量。

$$S=\bar{S}+P_{id}\boldsymbol{\alpha}_{id}+P_{\exp}\boldsymbol{\alpha}_{\exp} \tag{10}$$

其中，S 表示生成的 3D 人脸形状，$\bar{S}$ 表示平均形状，P_{id} 和 $P_{\exp}$ 是系数，$\boldsymbol{\alpha}_{id}$ 和 $\boldsymbol{\alpha}_{\exp}$ 分别是形状和纹理的主成分向量。3D 坐标和 2D 坐标之间的转换计算如下：

$$V(p)=f*Pr*\boldsymbol{R}*(\bar{S}+P_{id}\boldsymbol{\alpha}_{id}+P_{\exp}\boldsymbol{\alpha}_{\exp})+\boldsymbol{t}_{2d} \tag{11}$$

其中，$Pr=\begin{pmatrix}1&0&0\\0&1&0\end{pmatrix}$，$f$ 表示缩放因子，$\boldsymbol{R}$ 表示旋转矩阵，$\boldsymbol{t}_{2d}$ 表示平移向量。

深度图像对光线、阴影等外界因素的影响较小，因为深度图像的像素值是由红外光线反射回来的距离信息得到的。因此，使用深度图像可以提高人脸表情识别的鲁棒性，即使在光线较暗、阴影较强等复杂环境下也能够得到比较稳定的识别结果，如图 3 所示。

图 3　RGB 图像及其深度图像

Fig. 3　RGB images and their depth images

1.4 多流网络设计

用多流卷积神经网络(Multi-Stream Convolutional Neural Network,MSCNN)进行视频表情识别是一种较为有效的方法。该方法将每个视频帧看作一个图像,将多个 CNN 模型分别应用于视频的不同部分,从而可以更全面地提取面部表情特征,以提高表情识别的准确性和鲁棒性。

网络的输入是一个视频序列,包括若干个帧,每个帧包括 RGB 图像、深度图像和 x 方向和 y 方向的光流图像。可以将 RGB 图像、两个方向的光流图像和深度图像分别输入 4 个并行的卷积神经网络中,以提取不同信息的特征,如图 4 所示。

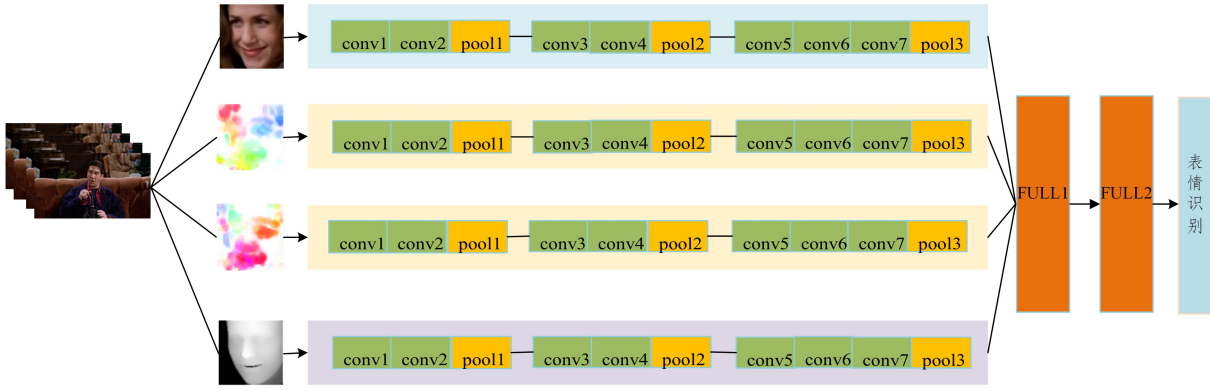

图 4 多流卷积神经网络架构

Fig. 4 Structure of multi-stream convolutional neural network

每个卷积神经网络由若干个卷积层和池化层组成,最后连接到一个全连接层,用于输出特征向量。将该特征向量映射到表情类别的概率分布上,以实现表情识别任务。每个子模型都包含了多个卷积层和最大池化层,其中卷积层使用不同的卷积核进行特征提取,并使用 ReLU 激活函数进行非线性变换。最大池化层用于降采样,以减小特征图的尺寸和减少参数数量。每个子模型的输出结果被合并到一个特征向量中,该特征向量可以被送入后续的全连接层进行分类、回归等任务。整个模型可以在 GPU 上进行训练和推理。

4 个子网络都输入三通道的图像,每个子模型包含 3 个卷积层和 3 个池化层,每个卷积层都采用 3×3 的卷积核,64 个滤波器,padding=1,stride=1,输入通道数为 3～64,然后是 128,最后是 256。每个池化层都采用 2×2 的最大池化层。

这 3 个子模型都被训练成了特征提取器,它们分别从 RGB 图像、光流图像和深度图像中提取出特征。这些特征分别代表图像的颜色、动态变化和空间结构。这些特征都被展开成一维向量,并拼接在一起,形成一个更加丰富的特征向量。多流卷积神经网络的设计可以根据任务的要求进行调整。在人脸表情识别中,可以采用类似于 Inception-v3 的结构,使用 1×1,3×3 和 5×5 的卷积核进行卷积,以获得不同尺度的特征。

此外,还可以在每个分支中使用 Dropout 和 Batch Normalization 等技术,以减少过拟合和加快训练速度。最后,通过将多个分支的输出连接在一起,可以得到一个全局特征向量。可以在这个向量上添加全连接层和 Softmax 层,以将不同类型的特征信息融合在一起,并进行表情分类。

使用多流卷积神经网络的优势在于可以利用不同类型输入的特定信息,从而提高表情识别的准确性。例如,在使用 RGB 图像进行表情识别时,由于光照和遮挡等因素的影响,准确率可能不够高。但是,如果同时使用深度图像,则可以更好地区分面部表情,提高准确性。同时,使用光流图像可以捕捉到表情的动态变化信息,从而更准确地识别表情。因此,多流卷积神经网络可以结合多种输入类型的信息,更全面地学习和提取表情特征,进而提高表情识别的准确性。

2 实验

2.1 实验设置

CAER 数据集:人脸数据集,总共 11 859 个视频序列,包括 9 222 个训练视频,2 637 个测试视频。每个序列长度约为 90 帧,包含人物姿态表情变化。单张图像可能包含多张人脸,但只标注 1 张。

RAVDESS 数据集:该数据库由 24 名专业演员组成,共 2 880 个视频序列,包括平静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶的表情,并且这两个数据集都是公开共享的[18]。

本实验在 Windows11 操作系统下使用 python3.9、pytorch1.11.0 框架、CUDA11.7 完成。使用 1.4 节中的多流卷积神经网络模型,设置学习率为 0.001,批量大小为 8,训练轮数为 100。

2.2 结果分析

本文针对表情识别在视频中的应用,进行了消融实验和对比实验,以验证多流卷积神经网络在这一领域的可靠性。首先,在 CAER 和 RAVSESS 数据集上进行了消融实验,通过逐步剔除多流网络中的不同流(如深度图像流、光流图像流和 RGB 图像流),观察表情识别准确率的变化。结果表明,多流卷积神经网络的各个流在视频表情识别中发挥了重要作用,多流网络能够更好地捕捉到视频中的时空信息,充分利用不同类型的图像数据,从而升高表情识别的性能。

2.2.1 基于 CAER 的实验结果

在本次实验中,为验证本文提出的人脸表情识别方法的性能,选用视频数量较多、图像质量较高的 CAER 数据集,先后输入了不同的图像数据。第一种方案只输入空间图像和光流图像,第二种方案输入空间图像和深度图像,第三种方案输入光流图像和深度图像,第四种实验方案输入空间图像、光流图像和深度图像。对不同方案的实验结果进行对比,如表 1 所列。

表 1　CAER 数据集表情识别消融实验结果

Table 1　Ablation experiment results of expression recognition on CAER dataset

（%）

方法	准确度
RGB+OF	78.1
RGB+DM	78.3
OF+DM	79.1
RGB+OF+DM(MSCNN)	81.2

从表 1 可以看出，RGB+OF 和 RGB+DM 这两种方法的准确度较为相近，可以使用的信息较少，不能充分表达出人脸的表情信息；OF+DM 的准确度略高于前两种方法，因为深度图像提供了更多的表情信息；MSCNN 的准确度最高，因为将 RGB、光流和深度图像结合起来可以更好地表达人脸表情的信息，且 MSCNN 可以学习到更好的特征表示。

对于每个表情类别，本实验计算了其对应的真阳性率（TPR）和假阳性率（FPR）以绘制 ROC 曲线，如图 5－图 8 所示。通过绘制各个表情类别的 ROC 曲线，本实验评估了分类器在不同阈值下的性能，并计算了每个表情类别的 AUC 值。可以看出，所有表情类别的 AUC 值均在 0.8 左右，这表明本实验的分类器在不同表情类别上具有良好的区分能力和分类性能。

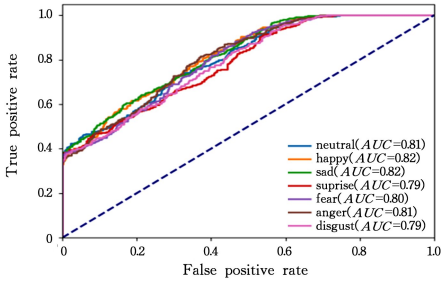

图 5　RGB+OF+DM ROC 曲线

Fig. 5　ROC curve of RGB+OF+DM

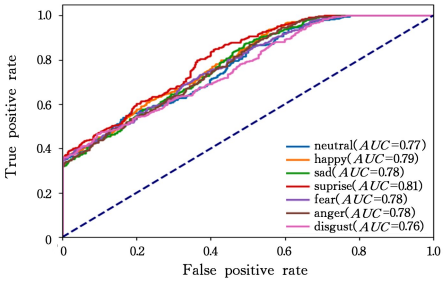

图 6　RGB+OF ROC 曲线

Fig. 6　ROC curve of RGB+OF

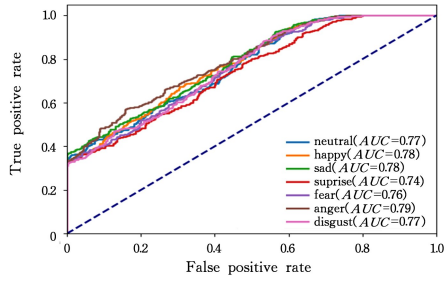

图 7　RGB+DM ROC 曲线

Fig. 7　ROC curve of RGB+DM

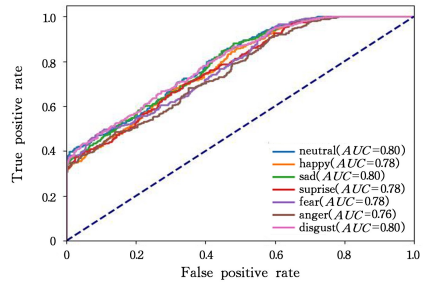

图 8　OF+DM ROC 曲线

Fig. 8　ROC curve of OF+DM

本实验选取了文献[10，19-21]中的部分方法进行了对比，结果如表 2 所列。从表中可以看出 MSCNN 在该数据集上相比 CAER-Net-S 有一定的提高，准确率达到了 81.2%，表现最好，优于其他的识别方法。

表 2　CAER 数据集实验结果比较

Table 2　Comparison of experimental results on CAER dataset

（%）

方法	准确度
ImageNet-AlexNet[19]	47.36
ImageNet-VGGNet[20]	49.89
ImageNet-ResNet[21]	57.33
Fine-tuned AlexNet[19]	61.73
Fine-tuned VGGNet[20]	64.85
Fine-tuned ResNet[21]	68.46
CAER-Net-S[10]	73.51
MSCNN(本文方法)	81.20

2.2.2　基于 RAVDESS 的实验结果

为了进一步验证 MSCNN 方法的表情识别性能，本文在 REVDESS 数据集上进行了消融实验。第一种方案只输入空间图像和光流图像，第二种方案输入空间图像和深度图像，第三种方案输入光流图像和深度图像，第四种实验方案输入空间图像、光流图像和深度图像。实验结果如表 3 所列。

表 3　RAVDESS 数据集上的表情识别消融实验结果

Table 3　Ablation experimental results of expression recognition on RAVDESS dataset

（%）

方法	准确度
RGB+OF	77.4
RGB+DM	78.3
OF+DM	78.8
RGB+OF+DM(MSCNN)	85.5

在本次实验中，对 7 种不同的表情也进行了 ROC 曲线的绘制，并计算了每种实验方法在不同表情上的 AUC 值，如图 9－图 12 所示。这些结果进一步证明了本实验在分类性能方面的优越性。通过绘制 ROC 曲线和计算 AUC 值，能够评估分类器在不同表情上的准确性和可靠性，结果显示实验方法在各个表情上都表现出良好的性能。

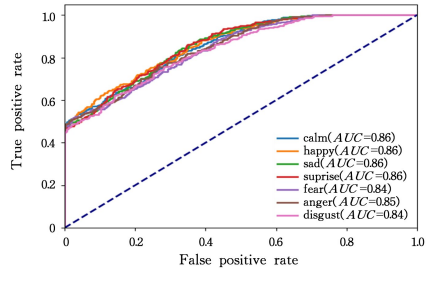

图 9　RGB+OF+DM ROC 曲线

Fig. 9　ROC curve of RGB+OF+DM

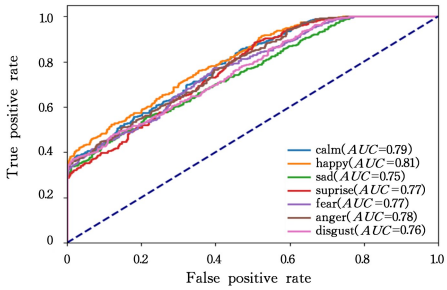

图 10 RGB+OF ROC 曲线

Fig. 10 ROC curve of RGB+OF

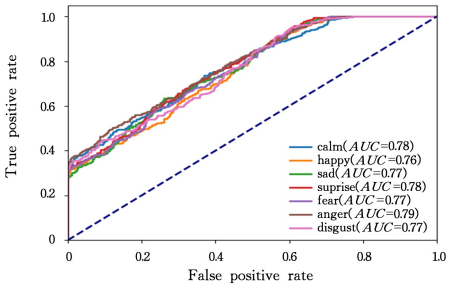

图 11 RGB+DM ROC 曲线

Fig. 11 ROC curve of RGB+DM

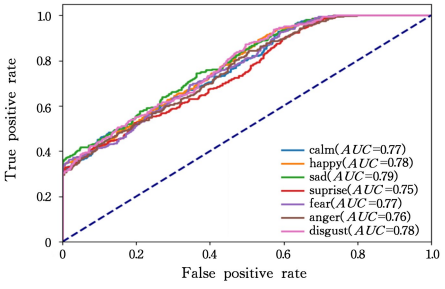

图 12 OF+DM ROC 曲线

Fig. 12 ROC curve of OF+DM

本文选择了文献[22-24]中的一些实验方法进行对比，结果如表 4 所列。由对比实验结果可以发现，本文方法 MSCNN 在准确率方面略高于其他文献中的方法。这进一步表明 MSCNN 方法在处理该问题上具有较好的性能和效果。这一结果也进一步验证了本文方法的可靠性，并为解决相关问题提供了一个可行的方案。

表 4 RAVDESS 数据集实验结果比较

Table 4 Comparison of experimental results on RAVDESS dataset

(%)

方法	准确度
AlexNet(FineTuning)[22]	61.67
CNN-14(Fine-Tuning)[22]	76.58
xlsr-Wav2Vec2.0[23]	81.82
CNN-X[24]	82.99
MSCNN(本文方法)	85.50

综上所述，实验结果表明，在表情识别任务中，本文所提出的方法在不同表情类别上都表现出了较高的分类性能。研究结果为进一步改进表情识别技术和应用提供了有价值的参考。

结束语 本文提出一种融合三维人脸动态信息和光流信息的表情识别方法，本研究使用多流卷积神经网络，融合深度图、光流图和图像的空间信息，进行动态人脸表情识别任务的研究。通过 CAER 和 RAVDESS 数据集进行定量分析实验，实验结果表明，多流卷积神经网络在动态表情识别任务中表现出色，取得了理想的识别效果。通过对比实验发现，将不同类型的图像信息进行融合可以提高表情识别的准确性，而使用多流卷积神经网络可以有效地实现这一目标。因此，本研究为动态人脸表情识别任务提供了一种有效的解决方案，具有一定的理论和应用价值。

参 考 文 献

[1] XU F, ZHANG J, WANG J Z. Microexpression identification and categorization using a facial dynamics map[J]. IEEE Transactions on Affective Computing, 2017, 8(2): 254-267.

[2] MA H Y, AN G Y, RUAN Q Q. Micro expression recognition described by the average optical flow direction histogram[J]. Journal of Signal Processing, 2018, 34(3): 279-288.

[3] WANG Y, WANG F, JIA H R, et al. Microexpression recognition combined with facial key points and optical flow features [J]. Laser Journal, 2023, 44(5): 72-77.

[4] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27. 568-576.

[5] FERNANDO B, GOULD S. Learning end-to-end video classification with rank-pooling[C]// International Conference on Machine Learning. PMLR, 2016: 1187-1196.

[6] ZOLFAGHARI M, SINGH K, BROXT. Eco: Efficient convolutional network for online video understanding[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 695-712.

[7] WANG L, XIONG Y, WANG Z, et al. Towards good practices for very deep two-stream convnets[J]. arXiv: 1507. 02159, 2015.

[8] AGHAMALEKI J A, ASHKANI CHENARLOGH V. Multi-stream CNN for facial expression recognition in limited training data[J]. Multimedia Tools and Applications, 2019, 78(16): 22861-22882.

[9] ZHU X, LIU X, LEI Z, et al. Face alignment in full pose range: A 3d total solution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 41(1): 78-92.

[10] LEE J, KIM S, KIM S, et al. Context-aware emotion recognition networks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 10143-10152.

[11] ZHANG W, ZHANG Y, MA L, et al. Multimodal learning for facial expression recognition[J]. Pattern Recognition, 2015, 48(10): 3191-3202.

[12] FENG D, REN F. Dynamic Facial Expression Recognition based on Two-Stream-CNN with LBP-TOP[C]// 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems(CCIS). IEEE, 2018.

[13] ZHOU P, HAN X, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017: 1831-1839.

[14] LU B, ZHOU J, WANG Q, et al. Fusion-based color and depth image segmentation method for rocks on conveyor belt[J]. Minerals Engineering, 2023, 199: 108107.

[15] XING H, YANG J, XIAO Y. Learning dynamic relationship be-

tween joints for 3D hand pose estimation from single depth map [J]. Journal of Visual Communication and Image Representation，2023，92：103803.

[16] JIANG H，ZHANG Q，NIE Y，et al. Learning Multi-Scale Deep Image Prior for High-Quality Unsupervised Image Denoising [J]. Computer Graphics Forum. 2022，41(7)：323-334.

[17] NIU W，ZHAO Y，YU Z，et al. Research on a face recognition algorithm based on 3D face data and 2D face image matching [J]. Journal of Visual Communication and Image Representation，2023，91：103757.

[18] LIVINGSTONE S R，RUSSO F A. The Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS)：A dynamic，multimodal set of facial and vocal expressions in North American English[J]. PloS one，2018，13(5)：e0196391.

[19] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[20] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv：1409.1556，20124.

[21] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.

[22] LUNA-JIMÉNEZ C，GRIOL D，CALLEJAS Z，et al. Multimodal emotion recognition on ravdess dataset using transfer learning [J]. Sensors，2021，21(22)：7665.

[23] LUNA-JIMÉNEZ C，KLEINLEIN R，GRIOL D，et al. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset[J]. Applied Sciences，2021，12(1)：327.

[24] KANANI C S，GILL K S，BEHERAS，et al. Shallow over Deep Neural Networks：A Empirical Analysis for Human Emotion Classification Using Audio Data[C]// 5th International Conference on Internet of Things and Connected Technologies (ICIoTCT). 2020. Cham：Springer International Publishing，2021：134-146.

ZHANG Huazhong，born in 1989，associate professor，master's supervisor. His main research interests include flying qualities monitoring，artificial intelligence and image processing.