基于深度卷积神经网络的表情识别综述

彭 程
澳门科技大学 中国 澳门 999078

【摘 要】：人脸表情识别可以直接解决电脑和机器人如何理解人类意图的问题。随着深度卷积神经网络的快速发展，很多相关技术也被引入到了表情识别领域。基于此，总结了基于深度卷积神经网络的人脸表情识别技术的研究现状，同时分析了这些技术的优势和不足。最后，对人脸表情识别在实际应用中仍然存在的问题和不足，给出了进一步改进的方向。

【关键词】：深度学习；卷积神经网络；表情识别

引言

面部表情是人类表达自己情绪最直接的方式，所以对于电脑和机器人想要理解人类的意图，进行面部表情识别是最直接的方式。表情识别是指从静态照片或视频序列中首先检测出人脸区域，然后根据人脸区域识别出表情状态，从而判别目标人物的情绪与心理变化。面部表情自动分析在情绪分析、网络教育、智能医疗、人机交互、智能安全、娱乐、网络教育等许多方面有着广泛的研究和应用前景。

1971 年，美国心理学家 Ekman 和 Friesen 定义了人类六种基本表情：包括幸福（Happy）、生气（Angry）、吃惊（Surprise）、害怕（Fear）、厌恶（Disgust）和难过（Sad），并且建立了面部动作编码系统，人脸表情识别的研究由此开始，很多专家学者开始进入表情识别的研究领域，有力的推动了人脸表情识别的落地应用。人脸表情识别的一般步骤如下图 1 所示，拿到图像数据后，先对图像进行面部检测，得到图片中的人脸区域，然后再进行预处理及特征提取，得到的人脸面部区域特征即可送入面部表情分类模块进行人脸表情识别。人脸表情的标记是一个十分困难的事情，需要借助专家的经验知识，所以表情数据库的建立是比较困难的，本文对目前已经发布的各种人脸表情数据库进行了介绍，并分析总结各人脸表情数据库普遍存在的问题，针对一些人脸表情数据库或者不均衡的数据库介绍了基于深度卷积神经网络的数据扩展方法。最后，本文总结了基于深度卷积神经网络的表情识别方法及表情识别相关数据库在研究及实际应用场景中存在的问题，展望了进一步改进的方向。

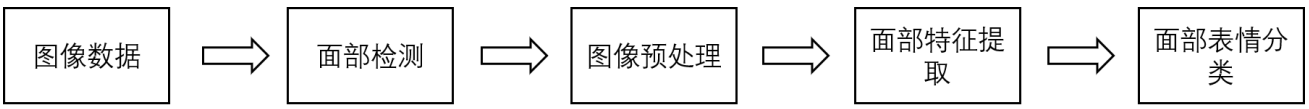

图 1

1 基于深度卷积神经网络表情识别方法

最初，各种人为设计的特征如：局部二值算子、Gabor、VLBP、HOG 和 SIFT 和 ORB 用于人脸表情识别，取得了一定的成绩及效果。深度卷积神经网络在图像识别领域的巨大进步及其获得的显著效果，让很多研究者看到了将这些方法引入到表情识别领域的新思路。

深度卷积神经网络可以同时进行特征提取和面部表情分类，但同时也意味着需要大量的训练数据样本。由于面部表情数据是很难收集的，直接在一个小的表情数据集上训练深度网络模型是很容易造成过拟合的。为了减轻这个问题，一些研究使用额外的相同任务导向的数据在知名预训练模型上先进行微调，然后再去预训练它们的自建网络。基于此方法，Ngetal[1]介绍了他们是如何在一个小的表情数据集上做微调的。他们使用有监督的两步法去微调一个网络，首先使用 FER2013 数据集在预训练好的模型（此模型是基于 ImageNet 数据集的训练的）上微调，然后使用目标数据集的训练集继续微调来使最终模型更加适合目标数据集。

为了把卷积神经网络应用到表情识别领域，Levi et al.提出了一个空前的把图像强度转换到 3D 空间的方法，并且能做到对单调光度变换不变。该方法通过去除输入图像中的混杂因素来简化问题域，减少有效训练深层 CNN 模型所需的数据量，使用 Multi Dimensional Scaling（MDS）转换无序的 LBP 编码值到一个度量空间点。然后就像训练正常图片那样去训练，用有限的情感标记训练数据对每个模型进行微调，得到最终的分类模型。

为了验证深度学习在表情识别领域中可以获得非常好的效果，Jaiswal textitet al.展示了一个在图像中检测面部动作单元的深度学习框架。他们使用卷积神经网络对外观、外形和动态的面部动作单元建模，而且他们只使用一个深度卷积神经网络可以同时学习到所有的关键特征（外观、外形和动态的特征）。该文介绍了一个空前的方法通过使用面部关键点的位置计算二值图像掩模来对面部区域的形状编码，这样就可以使网络学习到相关的形状特征。同时还使用一个结合

了时间窗口卷积神经网络和 Bi-directional 长短期记忆神经网络(BLSTM)来对时态信息建模。BLSTM 是一个能够通过长时间间隔储存信息的循环神经网络架构。脸部区域的预处理是使用[2, 3]的方法来追踪脸部关键点并且使用一个参照外形对齐脸部图片。这个对齐使用的是通过面部眼角和鼻子关键点定义的一个 Procrustes 转换。

卷积神经网络可以作为多层感知器（MLP）的一个特殊类型，它通过使用接受域关注像素之间的局部关系。在很多图像识别任务中，卷积神经网络可以获得很高的识别率。所以，Nwosu etal.提出了一个双通道的卷积神经网络方法，该方法可以分为三个阶段：图像预处理，面部特征提取和特征分类。首先使用 violaJones 算法做人脸和面部区域检测，检测到的脸部区域（64 × 64pixels）会被切割出来，然后检测眼睛和嘴巴区域（32 × 64 pixels）并且切割出来作为 CNN 最初的输入数据。这样做可以减少网络学习到的信息并且加快网络的训练。训练阶段包含了使用卷积神经网络抽取特征和分类。在该网络中，面部区域作为第一个卷积层的输入用来抽取眼睛和嘴巴区域，抽取的眼睛区域作为第一个通道的输入，抽取的嘴巴区域作为第二个通道的输入。两个通道的信息会在一个全连接层汇集，这样做可以从这些局部特征中学习全局信息，然后被用作分类。在该方法中，每个 CNN 通道有两个卷积层，2 个 max-pooling 层，一个全连接层，1 个输出层组成。第一个 CNN 层使用了 5×5 的卷积核，然后是一个卷积核为 2×2 的 maxpooling 层，然后接着是第二个卷积层，第二个 CNN 层同样使用 5×5 的卷积核后面跟着一个卷积核为 2×2 的 maxpooling 层。

2 基于生成对比网络和对抗网络的表情识别方法

此外，随着深度卷积神经网络的不断发展，生成对比网络和对抗网络也被使用到了面部表情领域。Kimetal.提出了一个深度生成对比网络来做表情识别。通过估计一个参考表情图片学习一个生成网络这样可以消除表现力这一影响因素。这个估计参考图片被用来通过和原来的表情图片的特征空间对比来估量一个富有表现力的表现。然后，在对比度量学习和一个有监督的重建下可以帮助解决面部表情因素的问题。该方法的框架步骤类似于人脑思考的过程。给予一个表情图片大脑会去与记忆中的表情轮廓相比较。在该网络中，一个给予的表情脸部图片特征与一个被生产网络估计过的参照图片相比较，在此过程中，假设表情因素可以在给予图片与参考图片之间的对比特征中提取出来。

在表情识别的任务中，很难基于少量数据集或者不均衡的数据集来训练一个深层网络，而且在日常生活中，厌恶的表情相比于幸福和难过是很少见的。为了解决这个问题，zhuetal.提出了使用对抗网络来做数据增强，该方法不但能补充和完善数据流形，也能找到两个类别之间更好的边缘，使用了一个 CNN 模型做分类和一个持续循环的对抗网络做生产。

3 数据集

数据库对一个研究领域的重要性是毋庸置疑的，人脸表情数据库是进行人脸表情识别的研究及验证所提算法有效性的不可或缺的工具，所以建立表情数据库对学者探讨和研究表情至关重要。

3.1 美国 Cohn-Kanade 表情数据库

Cohn-Kanade 表情数据库由美国 CMU 机器人研究所和心理学系共同建立，该数据库拥有更大的数据量，它是由不同性别和种族而且年龄在 18 到 50 岁左右的 210 个志愿者的 2105 幅表情图像序列所组成的，每张图像的大小统一 640X490。该数据库主要运用面部运动编码系统将表情划分为六类且每个人的每类表情都来源于同一个表情序列，表情类别为：悲伤、高兴、恐惧、生气、厌恶、惊讶。由于该数据库并未完全开源，所以限制了其对表情识别研究的贡献价值。

3.2 CK+数据库

2010 年发布的 CK+数据库是 CK（Cohn-Kanade 数据集）的扩展。该数据库数据量适中，包含了表情的标记以及运动单元的标记，很多文献都会用到这个数据集来验证自己方法的优劣。CK+数据库一共有 123 个人员参与其中，共采集了 593 个表情图像序列，每个图像序列展示了某个表情的变化过程，帧数范围从 10 到 60 帧不等，是一个比较完善的数据库。

3.3 Yale Face 数据库

该数据库由美国耶鲁大学计算机视觉与控制中心建立，共 165 张图片，数据库里每张图片的大小统一为 100×100。该数据库共采集了 15 位志愿者在不同光照、表情和姿态变化的图片，其中每个人还会有不同的装束。

3.4 FER2013

FER2013 数据集是国际表示学习比赛（ICML2013）中使用的面部表情数据集。该数据集大部分通过 Google 收集，并对面部图像进行标记。该数据集的图像分辨率低，部分图像存在遮挡、平移和旋转等情形，更加接近复杂的真实场景。图像根据标记分为 7 个表情类别，分布如表 1。

表 1 FER2013 类别分布表

类别	生气	厌恶	害怕	幸福	难过	吃惊	自然	共计
数量	4953	547	5121	8989	6077	4002	6198	35887

4 结束语

本文重点总结了表情识别中基于深度卷积神经网络的方法，介绍了基于一般深度神经网络的方法，也介绍了基于生成对比网络的表情识别方法，此外针对表情数据集不足或者不均衡的问题，介绍了对抗网络用于表情数据增强的方法，也对各表情数据集进行了介绍与分析。在深度卷积神经网络与人工智能技术迅速发展的今天，在实际应用场景中，对人脸表情识别的要求也越来越高，其中主要包含识别的准确率和实时性。在提高表情识别准确率时，不但需要克服光照遮挡等复杂环境下的问题，同时也需要解决人脸表情数据样本的不足。在实时性问题上，需要在保证识别准确率的同时，压缩模型大小，可以尝试的方法有蒸馏、搜索、压缩等方法。

参考文献：

[1] Hong Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis,and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In ACM on International Conference on Multimodal Interaction, pages 443–449,2015.

[2] Xuehan Xiong and Fernando De La Torre. Supervised descent method and its applications to face alignment. In IEEE Conference on Computer Vision and Pattern Recognition, pages 532–539,2013.

[3] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild.2014.