Engineering Technology and Development, 工程技术与发展 (3)2022,4
ISSN:2661-3506 (Online) 2661-3492(Print)

基于视频的面部表情识别综述

彭　程
澳门科技大学　中国 澳门特别行政区　999078

摘　要：面部表情是反应人类情绪最直观的一种表现，多指眼部肌肉、颜面肌肉和口部肌肉的变化来表现各种情绪状态。作为人机交互最重要的一环，面部表情的识别技术对于计算机和机器人能够理解人类的行为有着重要的意义。随着深度卷积神经网络技术的不断发展，许多方法也被引入到了表情识别领域。基于此，本文总结了深度卷积神经网络在基于视频的面部表情识别中应用的概况。

关键词：视频；深度学习；卷积神经网络；表情识别

引言

表情识别技术是人脸识别技术中的一个非常重要的技术，它涉及到解决计算机和机器人是否可以通过人类表情来判断人类需求的问题。作为人类，通过单张图片即可判断图片中人物的面部表情并且理解该人物在图片中那一刻甚至那一段时间的情绪。然而，想要在任何场景中都能够准确判断人物的表情是十分困难的，此时就需要判断动态的人脸，通过感知视频中人类连续变化的表情来理解人类需求。通常的，基于视频的表情识别步骤如下图1所示，首先从视频中提取出人脸区域，然后把人脸区域输入深度卷积神经网络，获得特征后再经过全连接层即可获得该段视频的表情类别。

图1

基于视频的人脸表情识别技术有着广泛的应用前景，近年来在人机交互、机器人、智能医疗、通信和自动驾驶等领域得到了广泛的关注，成为学术界和工业界的研究热点。随着深度卷积神经网络在表情识别领域的巨大成功，让很多研究者看到了将这些方法引入到视频表情识别领域的新思路。

1 基于深度学习的视频表情识别方法

面部表情识别的一个关键性挑战是从视频中捕获面部物理结构的动态变化，考虑到这些变化可以作为对面部表情的一种重要表现，而且深度RNN 在时间序列的上下文信息建模中有优势。基于此想法，zhang et al.提出了Part-based Hierarchical RecurrentNeural Network （PHRNN） 的方法，可以用来分析时间序列中的表情信息。一个面部表情视频的静止帧对于表情识别也有很强的识别力，但是表情识别的数据集普遍偏小，在训练中很容易造成过拟合。虽然浅层网络可以解决过拟合的问题，但是这对于使用深度学习方法去提取深度层面的特征是不利的，所以，zhang et al.在该方法中还提出了Multi-Signal Convolutional Neural Network （MSCNN）的方法，可以从静止帧提取空间特征。这样就可以捕获整体、外观和静止的信息。Temporal network （PHRNN）：首先基于面部动态结构，把人脸特征点分成4 部分，然后各自输入到模型中。局部特征沿着特征抽取方向级联在一起，通过面部形态变化和动态演变特性在上层形成全局的高维特征。Spatial network （MSCNN）：在训练阶段，MSCNN 取一对视频帧作为输入并且作为识别和监督的验证信号，这对增加不同表情的变化和减少相同表情的不同是有用的。

Kahou et al. 提出了EmoNets[1]做视频中人脸的表情识别，整个方法大致可以划分为下面四个步骤：（1）在额外的人脸数据集上训练网络。（2）为每一帧的脸通道提取7 个类别的概率。（3）在比赛数据集上，通过膨胀或者收缩视频的长度来把视频中所有帧的概率聚合到固定长度的视频描述。（4）先在比赛数据集的训练集的视频描述来训练SVM，然后做所有视频切片的分类。整个方法的传递途径如图-46所示。额外的数据集是FER2013和Toronto FaceDataset（TFD），而且还对这两个数据集做了相应的处理来让这两个数据集可以相互兼容（如特征点的匹配，降噪，过滤等操作）。而且在此方法中，Kahou et al. 仅仅使用额外的数据集来训练这个网络，在比赛数据集的训练集和验证集被用来做早停止和SVM 的子训练集。如果视频过长，就可以沿着时间一致性的方向独立的取10张视频帧组的概率向量，然后收缩面部通道去适合10帧视频描述。如果视频里能检测到人脸的帧数太少，就会重复展开一致性的帧直到总共获得10帧，然后训练数据集的视频描述就被用来训练SVM。

在视频情况下，一般必须通过聚合不同长度的帧序列获得的信息去产生一个分类结果。RNNs 提供了一个有吸引力的框架，通过使用一个连续的隐藏层表现值在一个序列中传播信息。基于此，Kahou et al.提出了使用两步法来把表情作为图片结构的时空演变来建模的方法。该方法首先使用CNN来训练分类包含表情的静态图片，然后基于CNN从各个帧中推断出来的高层表现训练一个RNN来预测整个视频的表情。

为了使更好的处理梯度的消失和爆炸问题，Kahou et al.使用了一个比LSTM更简单的网络IRNNs,该方法的特征融合网络，包含一个多层感知器（Multilayer Perceptron）而且每一个分支程序都含有独立的隐藏层。这些层的输出被连接起来并被作为另一的隐藏层的输入，然后接着一个输出数目与表情类别一样的softmax层。融合网络的第一层是包含了特殊形态的层，通过共享形态间隐藏单元的相似子集合，仍然可以在一些形态中保持有识别力的特征，此外，该方法还在决策层进行了融合。

随着多媒体信息数据的增加，从视频中挖掘有用的信息的需求将会呈指数级增长，这对于情感分析特别重要，因为服务和产品的评论逐步的从单向转变为多向，人们现在越来越喜欢通过视频来测评他们喜欢的产品，而且通过网络搜索来获取高质量的产品测评也是很有需要的。Poria et al.提出了通过结合音频、视频、文本中的感情特征的模型,并通过Multiple kernel learning （MKL）的特征选择方法，将特征组织并且分配到不同群组且每个群组有它自己的核心函数。该方法的输入为视频中的一个序列图片，为了捕获时间上的关联，Poria et al. 合并了在时间t和t+1上连续的每两张图片，使用不同的内核尺寸从转换的输入数据学习第一层的2D特征。相似的，第二层同样使用了不同的卷积核尺寸去学2D特征，上采样层用来转换不同内核尺寸的特征为一致的2D 特征，逻辑神经元层用来准备RNN 的输入数据。用一个神经元的相互连接层来使用延迟状态为长时间延迟建模。最终的输出层分类视频图片作“正的”或者“负的”。

为了更好的解决视频表情分类的问题，Vielzeuf et al.提出了[2]方法。该方法中，每一个形态都有自己的特征向量和分数，这些特征和分数可以被不同的方法融合使用。特征向量的维度被选择用来平衡各个形态的贡献，同时也在参数的数量和性能之间做权衡。在该方法中，每个视频的音频使用OpenSmile toolkit来提取一个长度为1582 的描述符向量。用两种不同的模型分析人脸，使用VGG-16 模型提取一个特征集，使用C3D模型提取另一个特征集，这两个特征集的大小都是4096-d 的。通过序列帧获得描述符经过时序上的融合，可以产生每个形态的得分和紧凑的描述符。然后三个形态（语音，VGG-face 和C3D-face）使用预测得分和紧凑的表现力结合在一起。然后使用一个包含形态丢失的全连接获得的隐藏展现再作为输入数据输入第二个可以输出得分的正常的连接层。这样做比简单的输入连接的形态特征到一个全连接神经网络要好，还可以让网络学习到联合的展现力。

为了结合从高层信息获得的分数和从低层信息获得的特征，Vielzeuf et al. 提出来一种新的方法，对不同的形态的特征分开使用一个全连接的分类神经网络，可以获得一个尺寸为7的向量输出。然后这个向量连接另外两个形态的得分产生一个尺寸为21的向量产生一个尺寸为21的向量。这个尺寸为21 的特征作为一个全连接分类神经网络的输入，然后可以输出一个尺寸为7 的预测向量，这样做可以结合其他形态的预测值来做预测。最终，这三个新的预测向量被连接到一起作为最后一个全连接分类器的输入。

2 数据集

基于视频的人脸表情识别的方向研究及所提方法有效性的验证都需要视频表情数据，所以建立视频表情数据库对研究基于视频的表情识别具有重大意义。目前，使用较多的视频表情数据库主要有：MMI Facial Expression 视频表情数据库、美国CMU的Extended Cohn-Kanade视频表情数据库、Oulu-CASIA视频表情数据库，CAER视频表情数据库，DFEW视频表情数据库以及其他大学实验室或科研机构建立的数据库。

2.1 MMI Facial Expression Database

该数据集不但含有表情标签，还含有单个FACS动作单元（AU）的表情，包含超过2900个视频。它对视频中的AUs进行了充分注释（事件编码），并在帧级上进行部分编码，指示每一帧的AU是处于中性、起始、顶点还是偏移相位，其中一小部分是根据视听笑声作注解，该数据库可供科学界免费使用。

2.2 Extended Cohn-Kanade Dataset

CK+表情数据库发布于2010年，包含从123个志愿者采集到的593个表情序列，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。其中327个表情序列添加了表情标签，包含8种基本的表情：愤怒、蔑视、高兴、悲伤、惊讶、厌恶、恐惧、中立。

2.3 Oulu-CASIA （Oulu-CASIA NIR&VIS facial expression database）

该数据库共有2472个视频序列，包含六种典型的表情:愤怒、厌恶、恐惧、快乐、悲伤和惊喜。视频序列来自80个受试者，年龄分布在23岁到58岁之间，其中73.8%的受试者为男性，受试者被要求坐在观察室的一张离摄像机的距离约为60厘米椅子上，然后根据图片序列中显示的表情示例做出面部表情。成像硬件的工作速率为25帧/秒，图像分辨率为320X240像素。

2.4 Context-Aware Emotion Recognition database（CAER）

CAER数据库包含13201个视频剪辑，其中有1.1M帧是可用的。视频的长短是不一样的，最短的视频大约有30帧，较长的视频大约有120帧，平均序列长度为90帧。此外，该数据库制作者们提取大约70K的静态图像创建一个静态图像子集,称为CAER-S。数据集是随机分割的训练（70%）、验证（10%）和测试（20%）集。

2.5 Dynamic Facial Expression in the Wild （DFEW）

野外动态面部表情（DFEW）是一个大规模的面部表

情数据库，该数据库种类繁多，数量庞大，注释丰富，由16372个从电影中剪辑出来的视频片段组成。DFEW数据库中的剪辑片段都是极具挑战性的，包含了实际使用场景中很多经常出现的干扰信息，如极端照明，遮挡，和姿态的各种变化。该数据库的制造者们聘请了12位专家注解者，每个剪辑都被这些专家独立标注10次，标注质量很高。

3 结束语

本文重点总结了表情识别中基于视频的方法，介绍了基于一般深度神经网络在视频表情识别中的应用，也介绍了基于多特征的视频表情识别方法，此外针对高低层信息特征，介绍了相关方法在视频表情识别中的应用，也介绍了相关的视频表情识别数据集。

参考文献：

[1] Kahou, Samira Ebrahimi, et al. "Emonets: Multimodal deep learning approaches for emotion recognition in video." Journal on Multimodal User Interfaces 10.2（2016）: 99-111.

[2] Vielzeuf, Valentin, et al. "Temporal multimodal fusion for video emotion classification in the wild." Proceedings of the 19th ACM International Conference on Multimodal Interaction. 2017.