

**毕业设计（论文）任务书**

**设计（论文）题目 基于因果分析的鲁棒人脸动作单元识别**

**学院名称 计算机与信息学院**

**专 业 （班 级） 计算机科学与技术19-1班**

**姓 名 （学 号）**

**指 导 教 师**

**系（教研室）负责人**

|  |  |  |
| --- | --- | --- |
| 1. 毕业设计（论文）的主要内容及要求（任务及背景、工具环境、成果形式、着重培养的能力）   **任务及背景：**  **工具环境：**  **成果形式：**  **着重培养的能力：** | | |
| 1. 应收集的资料及主要参考文献 | | |
| 三、毕业设计（论文）进度计划 | | |
| 起 迄 日 期 | 工 作 内 容 | 备 注 |
| 1月中旬  3月中旬  4月中旬  5月中旬  5月下旬 | 完成开题报告；  熟悉开发工具，完成外文翻译；  进行系统设计，系统实现，完成中期检查；  完成系统设计，准备课题验收，撰写毕业设计论文；  完成论文撰写，准备答辩。 |  |

**开 题 报 告 （该表格由学生独立完成)**

|  |
| --- |
| 建议填写以下内容：1.简述课题的作用、意义，在国内外的研究现状和发展趋势，尚待研究的问题。2.重点介绍完成任务的可能思路和方案；3.需要的主要仪器和设备等；4.主要参考文献。 课题的作用和意义 许多研究人员通过对肌肉运动进行编码，来了解面部运动的方式和原因。面部动作编码系统为他们提供了一种可靠的面部运动和表情编码和分析技术。  面部动作编码系统（FACS）是由Paul Ekman和Wallace 于1978年首次开发，并由Ekman和Friesen，Hager于2002年进行修订的一种基于解剖学结构的系统，用于描述面部的肌肉运动。  FACS一共定义了44个面部运动单元(action unit, AU)，并具体定义了每个AU的作用区域、运动表观特征以及各种表情的AU构成。  近些年来，自动化的AU检测技术得到了广泛的关注和应用。通过计算机自动检测AU及其组合，有助于准确分析面部表情和理解个体情绪，并在驾驶员疲劳检测、病患疼痛估计、刑侦测谎、影视评估和广告评估等场景中具有良好的应用前景。  随着以深度学习为代表的人工智能技术的发展，以及人脸检测、对齐和识别技术的不断完善，AU检测技术也取得了明显的进步。作为一个典型的视觉模式识别问题，AU检测存在一些独有的难点，其中就包括面部AU的个体差异问题。由于不同人的面部表情和动作模式具有很大的个体差异性，且不同个体存在种族、年龄、性别和脸型等显著的属性差异，每个AU均存在巨大的类内变化。此外，面部动作还存在强度的差异，既可能非常明显,也可能极其细微,因此需要精细的特征才能准确区分不同人的面部AU。目前，面部动作单元识别亟需一种受试者无关的、鲁棒的深度学习模型。  同时，因果推断近年来逐渐应用于计算机视觉任务，因果推断使模型能考虑任务中存在的因果效应，并启发了不变风险最小化(Invariant Risk Minimization, IRM)这种学习范式来学习因果关系而不是相关关系。  本课题从因果分析和不变特征学习的角度出发，尝试学习一种鲁棒的面部动作单元识别模型，并根据FACS定义的AU组合计算公式得到相应的受试者情绪，最终通过前端将分析结果展示给用户。 课题在国内外的研究现状 FACS被提出以来，学术界克服AU数据采集、标注的困难，发布了多个AU数据集，促进了AU识别技术的发展。早期的数据集是在受控环境下采集的具有良好光照和简单背景的正面或近似正面的人脸图片，受试者被要求人为地显露出指定的面部表情。最近十年，研究人员更多地关注受试者被诱发而自发产生的表情，代表性数据集包括受控环境下采集的BP4D[1]和DISFA[2]等。  其中，BP4D数据库包含41名年龄分布在18岁到29岁的受试者，发布的视频记录了受试者在8个不同任务过程中的面部表情变化，8个任务分别是一对一访谈（诱发愉悦情绪），观看电影片段（诱发悲伤情绪），突然听到声音（诱发惊讶情绪），即兴创作一首滑稽歌曲（诱发尴尬情绪），感受威胁（诱发恐惧情绪），双手持续放入冰水，经历来自实验员的侮辱（诱发愤怒情绪）和闻到异味（诱发厌恶情绪）。该数据库同时记录了2维和3维的人脸表情视频， 其中2维视频总共含有超过160000 帧人脸图像，示例图像如图1所示，每一帧均标注了FACS编码。在该数据库上的测试方案一般是身份独立的三折交叉验证，即将全部41人分为3组，其中两组的数据作为训练集，剩余一组作为测试集，要求预测12种AU出现的概率。    图 1 BP4D数据集示例  DISFA数据库采集了27名受试者在观看电影片段时的自发表情。最终发布的数据库包含27个视频，每个视频包含4845帧人脸表情图像，示例图像如图2所示。该数据库同时标注了每帧图像的FACS编码和每个AU的强度等级（共5级）。在该数据库上的评测方法一般也是采用身份独立的三折交叉验证方法，要求预测8种AU出现的概率。    图 2 DISFA数据集示例  近年来，基于BP4D和DISFA两个数据集进行训练，国内外研究者从不同角度提出了大量的深度学习模型。例如：针对标签稀缺性，可以利用迁移学习将有用的知 识迁移到当前任务；针对特征难捕捉性，可以从准确捕捉AU的关联区域从而提取AU特征来切入；针对标签不均衡性，可以考虑利用AU间的关联对不均衡的AU进行平衡。具体而言：  **1.基于迁移学习的方法**  最常见的迁移学习方法是在当前数据集上微调其他图像数据集上预训练的模型，由于不同类型的图像时常具有相似的颜色分布和背景环境等属性，预训练模型所携带的知识也有利于当前模型的训练。Zhou等人[6]基于一个在 ImageNet[7]上预训练的 VGG16[8]网络，实现AU强度估计和头部姿态估计。Ji等人[9]在整体表情识别和人脸识别这2个与AU相关联任务的数据集上分别预训练ResNet-34网络[10]，接着在AU数据集上分别微调2个网络，并将2个网络预测的AU出现概率取平均作为最终的预测值。预训练的数据集与当前数据集之间存在域(Domain)差异，且微调过程可能会丢失一些有用信息，因而限制了微调预训练模型的有效性。  **2.基于区域学习的方法**  AU为人脸局部肌肉动作，因而提取其特征需要准确定位关联区域，每个AU的关联区域包括其所在部位以及存在一定关联的其他部位。可以使用人脸特征点（landmark）辅助进行区域学习也可以让模型自适应地学习局部区域特征。  使用特征点辅助学习的方法中，具有代表性的是Jaiswal等人[11]利用特征点为每个AU预定义方形的感兴趣区域（Region of Interest，ROI）以及对应的二进制掩码(Mask)，其中掩码上特征点形成的多边形区域内点的值为1而其他点的值为0，然后基于CNN从裁剪的ROI和掩码提取每个AU的特征。Ali等人[12]先利用一个卷积层提取低层特征，然后根据特征点位置在这一特征图(Feature Map)上裁剪与AU的ROI 对应的方块，并分别利用一个CNN从每个方块进一步提取特征。  自适应区域学习则利用AU标签来有监督地训练深度神经网络，网络在特征学习过程中会隐式地自适应捕捉AU的关联区域。一些工作在网络中加入注意力学习模块， 显式地捕捉AU关联区域.。Shao等人[13]不依赖特征点的先验约束，直接通过AU识别的监督信号自适应地学习通道级注意力和空间注意力,同时利用全连接条件随机场(Conditional Random Field，CRF)捕捉像素级关系来优化空间注意力，从而选择和提取每个AU的关联特征。Ertugrul等人[14,15]分别采用一个CNN从裁剪的人脸块提取特征，接着利用注意力机制对各个块所提取的特征进行加权，实现AU识别。虽然上述工作能够较好地捕捉AU特征，但仍包含了一些不相关的信息，影响 AU识别的精度。  **3.基于关联学习的方法**  人脸表情涉及多个局部位置的肌肉动作，因而像素位置间的关系可以被利用起来 。表情中会时常出现多个AU，但不会所有 AU 都出现，因而除部分 AU 相互独立（不相关）外，多数AU之间并不独立，可能同时出现（正相关），也可能相互排斥（负相关）。而且，在视频中AU是动态变化的，挖掘时域关联可以促进AU识别。  在像素级进行关联学习，Niu等人[16]首先利用CNN提取人脸特征，这一特征的空间上每一点沿通道的特征向量被作为一个局部特征，接下来利用长短期记忆(Long Short Term Memory，LSTM)网络学习局部特征间的关系，由于不同AU涉及不同位置的肌肉动作，该方法对每个AU分别采用一个LSTM来学习不同局部特征的贡献。鉴 于密集的人脸特征点可以描述人脸几何结构，Fan等人[17]利用图卷积网络(Graph Convolutional Network， GCN)从特征点空间位置形成的几何图结构中学习一个隐向量，该隐向量包含人脸形状模式以及特征点间的相互依赖关系，在特征学习过程中被用来增强表征能力。在这些工作中像素与AU的对应并不明确，使得像素级关系对AU识别的促进作用较有限。  在AU级进行关联学习，为了抑制标签不均衡导致的预测偏置，许多工作通过调整采样率和权重来进行平衡。Li等人[18]在深度神经网络的训练过程中对训练集中出现频率较低的AU采用更大的随机采样率，使得每个小批量（MiniBatch）中不同AU 出现的频率较均衡。另一些工作[19，5，20]在计算AU识别损失时，给每一AU所赋的权重与该AU出现的频率成反比，从而加强了出现频率较低的AU。此外，为了平衡每个AU的出现频率和不出现频率，Li等人[21]对交叉熵损失中出现频率的熵项乘以训练集中该AU的不出现频率，而对不出现频率的熵项乘以该AU的出现频率，这样，若某一AU的不出现频率大于出现频率，其对应于出现的损失项被加强。Song等人[22]提出自适应加权损失函数，通过自适应地学习认知不确定性(Epistemic Uncertainty)来计算小批量中每个样本的权重，不确定性越高的样本被赋以越大的权 重，从而抵消数据不均衡。  当前采用时域关联学习的方法一般先提取视频中每帧人脸图像的空间特征，然后利用LSTM等时间序列模型对时域上帧间关联进行建模。Chu等人[23]采用CNN提取各帧空间特征，并用LSTM对帧间的时域信息进行建模，最后在CNN和LSTM的末端将时空特征进行融合。Bishay等人[24]设计一个三层级的框架：在第一层级利用 CNN学习人脸外观特征，并利用多层感知机从人脸特征点学习几何特征；在第二层级利用 RNN 从连续帧学习时域上的关联；在第三层级将各网络的预测结果进行融合。Zhang等人[25]利用注意力机制实现特征融合和标签融合，其中前者用于捕捉人脸局部块间的空间关系，而后者用于捕捉时域动态关系。  作为一种客观的人脸表情分析手段，面部动作单元检测相关技术一方面取得了相当的进展，另一方面仍然不能完全满足实际应用需求。可以说，相关方法仍然存在较大的探索空间。 完成课题的可行方案  1. 训练鲁棒的人脸动作单元识别模型   本课题关键在于基于因果分析，训练一种去除受试者混淆因子的人脸动作单元识别模型[4]。 数据预处理 对于DISFA数据集，使用OpenCV读取视频序列每一帧；对于BP4D数据集，从中筛选出所有2D图像，构建2D图像数据库；  使用Dlib提供的人脸检测和人脸关键点(face landmark)模型，检测出68个人脸关键点构建关键点列表。从检测出的关键点列表中提取眼睛和嘴巴的坐标信息，以此为基准对整张图片进行仿射变换，实现人脸对齐；  最后将所有图片裁剪至256\*256大小，完成图像数据的预处理；按照AU标签的激活强度0~5，认为大于等于2的强度下AU被激活，按照此规则制作标签数据集，完成标签数据的处理。 因果图分析 AU检测存在面部AU的个体差异问题。以“快乐”这一面部情绪表示作为案例分析，“快乐”的面部表情由AU6（抬起脸颊）和AU12（抬起嘴角）组成，因此AU6和AU12的共现是所有受试者共有的一种普遍的AU语义关系。  然而，除了普遍的AU语义关系之外，不同受试者表达情感的习惯有所差异，这意味着还存在特定于受试者的特殊AU语义关系。如图3所示，受试者1在微笑时倾向于挑起她的内眉毛（AU1），受试者2倾向于挑起他的整个眉毛（AU1和AU2）并降低下巴（AU26）。如果使用前三个受试者作为训练样本来训练模型，就可能使模型学到特定于受试者的语义关系，而非简单的“快乐”=AU6+AU12。当出现新的受试者4时，在训练模型时的特定于受试者的语义关系就可能导致预测结果向AU1,AU2或AU26偏倚。    图 3 AU检测的受试者差异问题  该方案通过后门调整的因果干预方法来去除“受试者”这一混淆因子，进而使得模型学习到通用的AU语义关系，使得模型更加鲁棒。  想要使用因果干预方法，首先要根据具体问题构建因果图，该任务的因果图如图4所示：    图 4 AU检测因果图  其中，X代表面部的情绪表达，S代表受试者习惯，R代表潜在的AU语义关系，Y代表模型估计的AU出现概率。对于因果图的解释如下：  S → X：受试者表达情绪的习惯导致在他们表达情绪时记录的面部图像中出现特定于受试者的面部表情。  S → R ← X:R表示潜在的AU语义关系，它由普遍的AU关系和受试者特定的AU关系组成。X → R反应了面部图像的潜在AU语义关系；S → R主要反映在预训练的AU识别模型中，模型在训练过程中包含了特定受试者的AU语义关系。  X → Y ← R: X→Y表示AU模型从输入的面部图像中提取面部外观特征来估计Y；R → Y表示模型学习的潜在AU语义关系通过利用训练数据的先验来影响估计的Y。  在因果推断的理论中，在面部情绪表达X和模型预测AU的出现概率Y之间，S是一个混淆因子，由于S的存在，使得Y的预测不可靠。为了消除混杂因素带来的不利影响，并获得仅根据X中的内容 (即输入面部图像的面部外观特征) 估计Y的模型，通过将do运算符应用于变量X来干预X。do-operator擦除所有进入X的箭头，并通过这种方式防止任何关于X的信息在非因果方向上流动。这样，从S到X的因果联系就被切断了，我们得到了一个近似P(Y | do(X)) 而不是P(Y | X) 的AU识别模型。  因果干预最基础的方法是后门调整，如果Z满足(X,Y)的后门准则，那么X对Y的因果关系可以写作：    显然，本图中我们需要对S进行调整，公式写作：   模型结构 如图5所示，CISNet以一张面部图像X作为输入，使用 Resnet34作为骨干网络提取面部图像特征，然后将提取出来的特征送入因果干预模块（即CIS模块），实现受试者特征的去混淆，之后将因果干预模块输出的特征与原始特征拼接送入分类器进行AU检测，在设定阈值为0.5的情况下，将分类器输出的AU出现概率估计值作为最终的二值AU预测结果。    图 5 模型架构(CISNet)  该模型中的因果干预模块是一个即插即用的插件，根据公式：    该公式中P代表模型预测的概率，代表每个受试者都要依次输入模型进行前向运算，这将花费大量时间。在这里使用归一化加权几何平均(Normalized Weighted Geometric Mean, NWGM)技巧，可以将目标公式近似为：    对于这个条件是X和R的条件概率，我们可以使用一个线性模型来近似它：    其中 是Resnet34提取的原始特征， 是因果干预模块输出的特征。  根据 ，将特征 设计为：  具体模块内设计如图6所示：    图 6 因果干预模块  因果干预模块由混淆因子注意力、混淆因子字典、混淆因子先验概率组成。其中，混淆因子先验概率由每个受试者面部图片除以总图片数得到；混淆字典计算受试者的所有面部图像特征的平均值，在每个epoch训练结束时对字典进行更新；混淆因子注意力则通过缩放点积注意力(scaled dot attention)计算当前输入的图片特征和字典特征的相似度权重，计算公式为：    最后，为了减轻AU识别中的数据不平衡问题，使用自适应的交叉熵损失函数进行训练：    其中，表示训练集中第i个AU的出现频率， 代表二值的AU标签， 代表模型预测的概率。  在pytorch框架上，使用动量为0.9和权重衰减为0.0005的随机梯度下降(SGD)优化器，学习率设置为0.001，批量大小设置为4，训练周期设置为15对模型进行训练。   1. 基于FACS计算情绪并展示  计算情绪 如图7所示，根据模型检测出的面部动作单元，从七种表情中进行匹配，计算出相似度最近的情绪。  7599328e9db0465a8ac32a483deb95ee  图 7 AU计算公式 前端展示 在后端得到动作单元的检测结果与对应情绪后，将预测结果传至前端，面部表情分析系统的形式如图8所示。  facereader-screenshot-action-units-timeline-happy-circumplex-model-asian-woman  图 8 面部表情分析系统 主要的软件和硬件环境 开发平台：Windows10, Ubuntu18  开发工具：Pytorch, Django, Vue  开发语言：python  硬件平台：[Nvidia RTX 3090 TI, 11th Gen Intel(R) Core(TM) i5-1135G7](https://www.amazon.com/Nvidia-RTX-3090-TI-Founders/dp/B09X4JVZB5)   主要参考文献 [1]Zhang, Xing, et al. "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database." Image and Vision Computing 32.10 (2014): 692-706.  [2] Mavadati, S. Mohammad, et al. "Disfa: A spontaneous facial action intensity database." IEEE Transactions on Affective Computing 4.2 (2013): 151-160.  [3] Chen, Yingjie, et al. "Causal intervention for subject-deconfounded facial action unit recognition." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 1. 2022.  [4] Luo, Cheng, et al. "Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition." arXiv preprint arXiv:2205.01782 (2022).  [5] Shao, Zhiwen, et al. "Jaa-net: joint facial action unit detection and face alignment via adaptive attention." International Journal of Computer Vision 129 (2021): 321-340.  [6] ZHOU Y, PI J, SHI B E. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning[C]/Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 872-877.  [7] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.  [8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations. San Diego: OpenReview, 2015: 1-14.  [9] JI S, WANG K, PENG X, et al. Multiple transfer learning and multi-label balanced training strategies for facial AU detection in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Work⁃ shops. Seattle: IEEE, 2020: 1657-1661.  [10] 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. Las Vegas: IEEE, 2016: 770-778.  [11] JAISWAL S, VALSTAR M. Deep learning the dynamic appearance and shape of facial action units[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Placid: IEEE, 2016: 1-8  [12] ALI A M, ALKABBANY I, FARAG A, et al. Facial action units detection under pose variations using deep regions learning[C]//Proceedings of the International Conference on Affective Computing and Intelligent Interaction. San Antonio: IEEE, 2017: 395-400.  [13] SHAO Z, LIU Z, CAI J, et al. Facial action unit detection using attention and relation learning[EB/ OL]. (2019-10-23)[2022-05-09]. https://doi. org/10.1109/ TAFFC.2019.2948635.  [14] ERTUGRUL I O, JENI L A, COHN J F. Pattnet: Patch-attentive deep network for action unit detection[C]//Proceedings of the British Machine Vision Conference. Car⁃ diff: BMVA Press, 2019: 114.1-114.13.  [15] ERTUGRUL I O, YANG L, JENI L A, et al. Dpattnet: Dynamic patch-attentive deep network for action unit detection[J]. Frontiers in Computer Science, 2019, 1(11): 1-13.  [16] NIU X, HAN H, YANG S, et al. Local relationship learning with person-specific shape regularization for facial action unit detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 11917-11926.  [17] FAN Y, LIN Z. G2rl: Geometry-guided representation learning for facial action unit intensity estimation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. Virtual Conference: IJCAI, 2020: 731-737.  [18] LI W, ABTAHI F, ZHU Z, et al. Eac-net: Deep nets with enhancing and cropping for facial action unit detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2583-2596.  [19] SHAO Z, LIU Z, CAI J, et al. Facial action unit detection using attention and relation learning[EB/ OL]. (2019-10-23)[2022-05-09]. https://doi. org/10.1109/ TAFFC.2019.2948635.  [20] CORNEANU C A, MADADI M, ESCALERA S. Deep structure inference network for facial action unit recognition[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 309-324.  [21] LI G, ZHU X, ZENG Y, et al. Semantic relationships guided representation learning for facial action unit recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8594-8601.  [22] SONG T, CHEN L, ZHENG W, et al. Uncertain graph neural networks for facial action unit detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Conference: AAAI, 2021: 5993-6001.  [23] CHU W S, DE LA TORRE F, COHN J F. Learning spatial and temporal cues for multi-label facial action unit detection[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 25-32  [24] BISHAY M, PATRAS I. Fusing multilabel deep net⁃ works for facial action unit detection[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 681-688  [25] ZHANG Y, JIANG H, WU B, et al. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data[C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 733-742. |
| **指导教师评语：**（建议填写内容：对学生提出的方案给出评语，明确是否同意开题，提出学生完成上述任务的建议、注意事项等）      **指导教师签名：**  **20 年 月 日** |