



西北大学  
NORTHWEST UNIVERSITY

## 硕士学位论文

# 低分辨率环境下的微表情识别

作者姓名： 李桂锋

指导教师： 彭进业 教授 西北大学

学位类别： 工学硕士

学科专业： 电子与通信工程

培养单位： 信息科学与技术学院

2019 年 6 月



# **Micro-expression Recognition Under Low-resolution Case**

**A thesis submitted to  
Northwest University  
in partial fulfillment of the requirements  
for the degree of  
Master of Engineering  
in Electronics and Communication Engineering  
By  
Li Guifeng  
Supervisor: Peng Jinye Professor**

**June 2019**



## 西北大学学位论文知识产权声明书

本人完全了解西北大学关于收集、保存、使用学位论文的规定。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。本人授权西北大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时授权中国科学技术信息研究所等机构将本学位论文收录到《中国学位论文全文数据库》或其它相关数据库。

保密论文待解密后适用本声明。

学位论文作者签名：\_\_\_\_\_ 指导教师签名：\_\_\_\_\_

年      月      日                  年      月      日

## 西北大学学位论文独创性声明

本人声明：所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，本论文不包含其他人已经发表或撰写过的研究成果，也不包含为获得西北大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：\_\_\_\_\_

年      月      日



## 摘要

人脸表情在我们的社交互动中发挥着重要作用，因为它传达了丰富的信息。我们可以从一张人脸图像中阅读很多内容，但是如果没有特殊设备，我们也无法感知到这些信息。本文采用计算机视觉方法分析肉眼难以察觉的两种微妙的面部信息：微表情和心率。微表情是快速、不自主的面部表情，揭示了人们不打算表达的情感。人们很难感知微表情，因为它们太快和微妙，因此自动微表情分析是很有价值的工作，具有重大的应用前景。本文综述了微表情研究的进展，并分四部分工作进行描述。1) 我们介绍了第一个自发的微表情数据库—SMIC。缺乏数据阻碍了微表情的分析研究，因为很难收集自发的微表情。引入用于诱导和注释SMIC的协议以帮助未来的微表情收集。2) 引入了包括三个特征和视频放大过程的框架用于微表情识别，其优于两个微表情数据库上的其他最先进的方法。3) 描述了一种基于特征差异分析的微表情定位方法，该方法可以从自发的长视频中发现为微表情。4) 提出了一种自动微表情分析系统(MESR)，用于发现并识别微表情。心率是我们健康和情绪状态的重要指标。传统的心率测量需要皮肤接触，不能远程应用。我们提出了一种方法，可以对抗照明变化和头部运动，并从彩色面部视频远程测量心率。我们还应用该方法来解决面部反欺骗问题。我们展示了基于脉冲的特征比传统的基于纹理的特征更能够抵抗看不见的掩模欺骗。我们还表明，所提出的基于脉冲的特征可以与其他特征相结合，以构建用于检测多种类型的攻击的级联系统。最后，我们总结了工作的贡献，并基于当前工作的局限性提出了关于微表情和心率研究的未来计划。还计划将微表情和心率（可能还有来自面部的其他微妙信号）结合起来构建用于情感状态分析的多模式系统。

微表达是一种基本的非言语行为，它能忠实地表达人类隐藏的情感。它在国家安全、计算机辅助诊断等领域有着广泛的应用，促使我们对自动微表情识别进行研究。但从监控视频中获取的图像容易出现质量问题，导致实际应用困难。由于捕获的图像质量较低，现有的算法无法达到预期的效果。为了解决这个问题，我们进行了全面的研究

**关键词：**微表情识别，监控视频，低分辨率，超分辨率，Fast LBP-TOP



## ABSTRACT

The face plays an important role in our social interactions as it conveys rich sources of information. We can read a lot from one face image, but there is also information we cannot perceive without special devices. The thesis concerns using computer vision methodologies to analyse two kinds of subtle facial information that can hardly be perceived by naked eyes: the micro-expression (ME), and the heart rate (HR)

MEs are rapid, involuntary facial expressions which reveal emotions people do not intend to show. It is difficult for people to perceive MEs as they are too fast and subtle, thus automatic ME analysis is valuable work which may lead to important applications. In the thesis, the progresses of ME studies are reviewed, and four parts of work are described. 1) We introduce the first spontaneous ME database, the SMIC. The lacking of data is hindering ME analysis research, as it is difficult to collect spontaneous MEs. The protocol for inducing and annotating SMIC is introduced to help future ME collections. 2) A framework including three features and a video magnification process is introduced for ME recognition, which outperforms other state-of-the-art methods on two ME databases. 3) An ME spotting method based on feature difference analysis is described, which can spot MEs from spontaneous long videos. 4) An automatic ME analysis system (MESR) was proposed for firstly spotting and then recognising MEs

The HR is an important indicator of our health and emotional status. Traditional HR measurements require skin-contact which cannot be applied remotely. We propose a method which can counter for illumination changes and head motions and measure HR remotely from color facial videos. We also apply the method for solving the face anti-spoofing problem. We show that the pulse-based feature is more robust than traditional texture-based features against unseen mask spoofs. We also show that the proposed pulse-based feature can be combined with other features to build a cascade system for detecting multiple types of attacks.

At last, we summarize the contributions of the work, and propose future plans about ME and HR studies based on limitations of the current work. It is also planned to combine the ME and HR (maybe also other subtle signals from face) to build a multimodal system for affective

status analysis.

Micro-expression is an essential non-verbal behavior that can faithfully express the human's hidden emotions. It has a wide range of applications in the national security and computer aided diagnosis, which encourages us to conduct the research of automatic micro-expression recognition. However, the images captured from surveillance video easily suffer from the low-quality problem, which causes the difficulty in real applications. Due to the low quality of captured images, the existing algorithms are not able to perform as well as expected. For addressing this problem, we conduct a comprehensive study about the micro-expression recognition problem under low-resolution cases with face hallucination method. The experimental results show that the proposed framework obtains promising results on micro-expression recognition under low-resolution cases.

**Keywords:** Micro-expression recognition, Surveillance video, Low-resolution, Super-resolution, Fast LBP-TOP

## 插图索引

图 1	人脸宏表情样本示例 (a)厌恶、(b)快乐、(c)惊讶、(d)恐惧、(e)愤怒、(f)轻蔑、(g)沮丧、 (h)中性表情	12
图 2	一个消极的微表情片段示例 (SMIC数据集)	14
图 3	微表情数据采集示意图	15
图 4	人脸面部表情肌肉划分	17
图 5	LBP特征的光照鲁棒性	19
图 6	LBP-TOP特征提取过程	19
图 7	光流大小和方向的表现	20
图 8	三维卷积神经网络架构	23
图 9	ResNet的残差学习模块	24
图 10	低分辨率环境下微表情识别框架	28
图 11	来自SMIC-HS数据集的两帧	28
图 12	ASM算法标定的68个人脸关键点	30
图 13	LWM算法人脸对齐后的图像	31
图 14	输入为原始图像序列, 输出为TIM算法插值后图像序列	31
图 15	超分辨重建过程	33
图 16	低分辨率图像 (a) SMIC-HS/SMIC-subHS数据集低分辨率图像, (b) CASME II数据集低分辨率图像	40
图 17	不同分辨率图像重建结果比较	40
图 18	不同数据集不同分辨率的图像序列的识别准确度	41
图 19	不同分辨率图像序列的识别准确度混淆矩阵	42
图 20	直方图均衡化	47
图 21	伪3D残差网络三种设计模型	53



## 表格索引

表 1	微表情的研究方法	8
表 2	摆拍微表情数据集	13
表 3	自发微表情数据集	17
表 4	实验中使用的数据集	39
表 5	重建图像序列的平均PSNR(dB)指标	39
表 6	重建图像序列的平均SSIM指标	40
表 7	不同分辨率图像在不同数据集上的识别精度比较	43



## 符号对照表

符号	符号名称
$\Delta$	difference
$\nabla$	gradient operator
$\delta^\pm$	upwind-biased interpolation scheme



## 缩略语对照表

缩略语	英文全称	中文对照
ME	Micro Expression	微表情
TIM	Time In Model	时间插值模型
BART	Brief Affect Recognition Test	微表情识别标准测验
FACS	Facial Action Coding System	面部动作编码系统
AU	Action Unit	动作单元
METT	Micro Expression Training Tool	微表情训练工具
FD	Feature Difference	特征差异
PD	Peak Detection	峰值检测
LOSO	Leave One Subject Out	留一法



---

## 目 录

摘 要 .....	I
ABSTRACT .....	III
插图索引 .....	V
表格索引 .....	VII
符号对照表 .....	IX
缩略语对照表 .....	XI
<b>第一章 绪论 .....</b>	<b>1</b>
1.1 研究背景与意义 .....	1
1.1.1 微表情研究的意义 .....	1
1.1.2 计算机视觉对微表情研究的意义 .....	2
1.2 国内外研究现状 .....	4
1.2.1 人工微表情识别训练工具研究 .....	4
1.2.2 自动微表情识别研究 .....	4
1.3 本文的研究内容 .....	9
<b>第二章 相关工作 .....</b>	<b>11</b>
2.1 宏表情和微表情 .....	11
2.1.1 宏表情数据集 .....	11
2.1.2 早期微表情数据集 .....	12
2.1.3 自发微表情数据集 .....	13
2.1.4 宏表情与微表情比较 .....	17
2.2 微表情特征提取的一般方法 .....	18
2.2.1 基于纹理信息的LBP-TOP特征提取符 .....	18
2.2.2 基于光流法的特征提取 .....	19
2.2.3 基于Gabor的特征提取 .....	21
2.3 相关深度学习网络 .....	22
2.3.1 三维卷积神经网络 .....	22
2.3.2 残差神经网络 .....	24

---

<b>第三章 基于传统机器学习方法的低分辨率环境下微表情识别</b>	27
3.1 低分辨率微表情数据获取	28
3.2 数据预处理	28
3.2.1 主动形状模型	29
3.2.2 局部加权平均算法	30
3.2.3 时间插值模型	30
3.3 超分辨重建过程	33
3.3.1 基于块的方法	33
3.3.2 基于像素的方法	34
3.4 微表情的特征提取与分类	34
3.4.1 LBP-TOP特征提取	35
3.4.2 LSVM	35
3.5 实验设置及分析	36
3.6 总结	43
<b>第四章 基于深度学习方法的低分辨率环境下微表情识别</b>	45
4.1 数据集预处理	46
4.1.1 数据集混合	47
4.1.2 直方图均衡化	47
4.1.3 亮度调整	48
4.2 特征提取及识别	48
4.2.1 P3D 块	48
4.2.2 P3D ResNet	51
4.2.3 2	52
4.3 实验设置及分析	53
4.3.1 1	53
4.3.2 2	53
4.3.3 3	53
4.4 总结	53
<b>第五章 系统设计</b>	55
5.1 需求分析	55
5.2 功能设计	55
5.2.1 功能图	55
5.2.2 时序图	55
5.2.3 等	55
5.3 界面设计	55
5.4 小节	55

<b>第六章 总结与展望</b>	57
6.1 总结	58
6.2 存在的问题与展望	58
6.2.1 存在的问题	58
6.2.2 展望	58
<b>附录 A 附录</b>	59
A.1 论文无附录者无需附录部分	59
A.2 测试公式编号	59
A.3 测试生僻字	59
<b>参考文献</b>	61
<b>致谢</b>	65
<b>攻读硕士学位期间取得的科研成果</b>	67



# 第一章 绪论

微表情（Micro Expressions），心理学名词，心理应激微反应的一部分，是人类表达自身情感信息的重要非语言性行为。微表情从人类本能出发，在大多数情况下，不受思想的控制，无法掩饰，也不能伪装<sup>[1]</sup>。因为它无法伪装的特性起初被人们用来作为鉴谎的辅助工具，随着人们对其不断深入的研究发现它在临床诊断、司法系统等有着很高的应用价值。近几年来随着计算机技术的不断发展，人们利用计算机视觉对微表情识别研究有了突飞猛进的成果，但就目前而言，还没有团队在低分辨率环境下对微表情做任何研究。本文从实际应用的角度出发，分析实际场景中面临的各种低质量问题，分别使用传统机器学习方法和深度学习方法对微表情识别。

本章主要阐述微表情识别研究的意义和低分辨率环境下微表情识别的重要性，国内外对微表情识别相关的研究和发展趋势，最后概述了文章的内容和结构分配。

## 1.1 研究背景与意义

达尔文在1872年出版了《The Expression of Emotions in Man and Animals》，从此拉开了人类对面部表情的系统性研究。时至今日，人类对面部表情的研究已经非常丰富与成熟，但主要关注的是显而易见的宏观表情（Macro Expressions），虽然在1966年Haggard和Isaacs首次提出了微表情现象（Micro-momentary Facial Expressions），但当时并未引起人们的普遍重视。直到三年后（1969年），Ekman和Friesen在临床发现了微表情，这一发现奠定了微表情在临床辅助治疗上的重要地位<sup>[2]</sup>，也开启了微表情的研究热潮。

本节将从微表情研究的意义和计算机视觉对微表情研究的意义两方面分析。

### 1.1.1 微表情研究的意义

加利福尼亚大学洛杉矶分校的心理学教授Albert Mehrabian在上世纪六十年代发现了人际交流中的“55384”原则，他提出有效的沟通技巧应该包含三大要素：身体语言、声音和谈话内容<sup>[3]</sup>。其中谈话内容传递的信息量是总信息量的7%，声音（包括交谈时的语气、音调和音量）传递的信息量占总信息量的38%，剩下的55%来自身体语言（包括谈话期间身体姿势、肢体动作、面部表情、眼神和目光等），也就是说身体语言比谈话内容能传达更多有价值的信息。论文中阐明，出现这种情况的主要原因是谈话内容（口头语言）可以有意识地被控制，而身体语言这种非语言行为是无意识的举动，人类的主观意识很难控制动作语言行为。身体语言由三部分构成：表情语言、动作语言和空间语言。表情语言指的是通过面部肌肉运动和眼睛神态所传递出来的思想感情，动

作语言指人类通过身体各个部位的动作或姿态来传递感情，空间语言主要指由个体与个体之间所保持的间距所形成的一种信息表达方式。在这三种身体语言中最容易被观察到的就是表情语言，艾伯特教授的这项发现说明了表情语言的重要性。

神经学家Paul Donald MacLean于上世纪五十年代提出了“大脑三位一体”理论 (The Triune Brain)，他认为人类颅腔内的脑并非只有一个，而是三个，这三个脑作为人类不同进化阶段的产物，按照出现顺序依次覆盖在已有的脑层之上，如同考古遗址一样<sup>[4]</sup>。根据在进化史上出现的先后顺序，他将人脑分成“爬行动物脑”(Reptilian brain)、“古哺乳动物脑”(Paleomammalian Brain) 和“新哺乳动物脑”(Neomammalian Brain) 三大部分，它们分别对应人脑的脑干(Archipallium)、边缘系统(Limbic System) 和新皮质(Neocortex)，它们共同控制着人类的身体行为。新皮质被称作“爱说谎的大脑”，经常会因为当事人的某种需要而出现说谎的现象。语言等由新皮质大脑控制的行为是不可信的，欺骗的嫌疑很大，想要得知对方内心的真实感受，必须观察对方边缘系统所控制的表情或肢体动作。边缘系统是控制人类情感的中心，管理着人类的非语言行为表达，因此是分析身体语言的重点。让人不加思索的产生本能反应是它的一大特点，它反映出一个人最真实的一面，这很难被控制和掩饰。比如，当听到刺耳的噪音时你会不自主地捂住耳朵、手碰到高温或极寒物体时会马上缩回等。所以边缘系统的行为是诚实可信的行为，是人类的思想、感觉和意图的真实反应，也是人类生存、本能的反应，它属于微反应中除微语言以外的非语言行为反应，它包括了微动作(Micro Action)、微表情。

从上述例子可以看出，有关微表情的研究在心理学和神经学两大学科都有着充分的理论依据。Ekman等人在临幊上发现微表情是来自于观看一位有自杀倾向的精神病患者的视频，视频中患者在回答医生问题时表现的很开心，没有任何想要自杀的异常迹象，但在随后的二次会谈中患者向医生承认其状况并未好转，而且她曾隐藏了自杀的计划。Ekman和 Friesen在逐帧慢放视频时发现确实存在两帧和绝望有关的负面表情，这与患者的二次会谈内容相吻合，但只持续了1/12 s。之后的几十年里Ekman和他的同事继续研究微表情，在不断的实践中量化并定义了微表情，这也引起了越来越多学术界和商业界人士的兴趣，目前微表情已经被应用到了众多领域，比如国家安全、司法系统、政治选举、临床诊断、公共管理和教育领域等<sup>[5]</sup>。

### 1.1.2 计算机视觉对微表情研究的意义

我们人类是优秀的“人脸识别专家”，我们已经习惯甚至并没有意识到这一点。与其他类型的物种相比，我们人类为应对复杂的社交交互问题，大脑已经开发了特殊的识别脸部信息的功能模块，以便我们更好地从人脸中获取更丰富的信息，所谓“察言

观色”就是很好的佐证。当然，人脸也是丰富的视觉信息的来源之处，我们可以从人脸中读取很多信息。比如眼前之人如果是著名人士，我们可以立即认出他或她，如果是陌生人，我们可以对这个人的性别、年龄、种族等做出基本正确的猜测，同时如果该人脸存在表情，我们也可以大致感知他或她的情绪状态。然而尽管我们是人脸识别专家，但这并不意味着我们已经解析出了全部的人脸信息，因为仍然存在部分无法用肉眼读取的深层次信息。

与其他感官（例如听觉和嗅觉）相比，我们的视觉认知功能在我们的大脑中更加精巧地被构建。然而，我们获取视觉信息的能力仍然受到生理机制的限制。超出我们感知范围的视觉变化（在空间域中太微妙或在时域中太快）将被我们的眼睛忽略。比如我们很难从人脸上观察到某个微表情，因为微表情会短暂且快速地发生，所涉及的肌肉运动强度也非常微弱，甚至表情发出者和观察者都察觉不到，尤其在高风险条件下微表情出现的机率更高，被察觉的可能性也更低。研究人员经过严密的统计，发现微表情持续时间最长为1/2秒而最短只有1/25秒，所涉及的肌肉运动强度更是微乎其微，而正常的表情（宏观表情）一般持续时间在1/2秒到5秒之间，有一个起承转合的过程。[\[6-8\]](#)。

由于微表情识别务实的使用价值，其提出者Ekman从2005年开始对英国情报机构、美国中央情报局等各国机构进行微表情识别培训，而那时他已经71岁高龄了[\[9\]](#)。他教辩护律师、健康专家、扑克选手，甚至对配偶心怀猜疑的人识破谎言，并且制作了网络课程。但他坦言人类对于微表情的识别能力终归是有限的，不仅要花费大量的人力和物力培训微表情识别专家，而且准确度不高，同时还伴有影响正常生活的风险[\[10\]](#)。Ekman曾说自己的识谎能力影响到了日常生活，他从不试图去识破周围朋友、亲戚的微表情，“去揭露每个人的微表情，揭穿每个人的谎言，这只会让自己的生活痛苦万分”。

同时当前对微表情研究的研究中，需要使用FACS（Facial Action Coding System）对包含被试微表情的视频进行逐帧的编码<sup>[? 1]</sup>。但是，不仅FACS编码的训练比较费时，编码者一般都需要接受100小时的训练才能达到初步熟练的程度；而且使用FACS进行编码也很费时，编码1分钟的视频平均需要2个小时<sup>[? 1]</sup>。为了更快地对基本表情编码，在FACS基础上，研究者发展出了一套附加的编码系统EMFACS（Emotion Facial Action Coding System），但人工对视频进行逐帧编码依然费时费力，这极大地限制了目前的微表情研究。因此，有效的微表情自动分析工具是开展微表情表达研究需要解决的一个重要问题。

计算机的发明是为了帮助人类更好地处理人类不想去处理的任务，而识别人脸微

表情这种会影响到日常生活的任务就是计算机存在的价值所在。而且当通过摄像机和计算机系统对待分析者分析时，不仅采集到的表情真实可靠（采集中采集对象并不知情，不存在任何干扰）而且通过计算机算法可以发现细微的人类无法察觉的表情变化，并且已经有大量的实验证明计算机的识别能力确实高于人类<sup>[11]</sup>。我们可以更好更快的训练计算机完成人类能够完成的任务，如人脸检测、人脸识别，同时我们还可以训练计算机执行我们无法完成的任务，如捕获肉眼难以察觉的细微信息<sup>[12]</sup>。

## 1.2 国内外研究现状

### 1.2.1 人工微表情识别训练工具研究

早期研究中，研究人员注重于测量或训练个体的微表情识别能力。Ekman 和 Friesen 在 1974 年制定了第一个微表情识别标准测验机制——BART (Brief Affect Recognition Test)，但当时的微表情识别标准测验有着很大的缺陷，它所呈现的微表情是孤立的呈现，这与现实生活中微表情的动态呈现方式完全不相符，这样的测验没有任何生态效度<sup>[? 1]</sup>。1978 年，Ekman 发布了面部动作编码系统 FACS，他们将人脸部的肌肉划分为 43 块，将它们随机组合获得了 1 万多种表情，但其中只有 3000 种具有情感意义，Ekman 等人又根据人脸解剖学特点，将这 43 块肌肉划分成相互独立又相互联系的运动单元 (Action Unit, AU)，分析这些运动单元的运动特征和其所控制的主要区域，将这些信息与相关的表情匹配就能得出面部表情的标准运动。为了克服 BART 的缺陷，Matsumoto 等人在 2000 年开发了更完善的微表情识别测量工具 (Japanese and Caucasian Brief Affect Recognition Test, JACBART)，该测验具有很好的可信度和严密的实验过程<sup>[? 1]</sup>。Ekman 等人在 2002 年根据日本人与高加索人短暂表情识别测验开发出了一个新的微表情识别训练工具 METT (Micro Expression Training Tool)，该训练工具有 7 种基本情绪的微表情，包括悲伤、恐惧、愤怒、厌恶、轻蔑、惊讶和高兴，METT 被应用在多种人群和领域，且对微表情受训者的识别能力有明显的提升<sup>[? 1]</sup>。

### 1.2.2 自动微表情识别研究

除上述通过训练提升人工识别能力外自动地微表情识别系统也在如火如荼的发展中，研究者们已经开发出很多相关的算法，甚至在某些数据集的准确度可达 90% 以上，但这些数据集有个明显的缺点，所有的微表情均为摆拍 (Posed)，这一缺点有其产生的必然性，但也严重违背了微表情的定义。为了解决这一问题，国内外的研究团队相继发表了自发微表情数据集，如中科院心理所分别在 2013 年和 2014 年发布了 CASME<sup>[? 1]</sup> 和 CAMSE II<sup>[? 1]</sup> 两个版本的数据集，后者比前者有着更高的时空分辨率和更多的数据量，但参与者全部为蒙古利亚人种（中国人），在数据的多样性上有一定的

不足；芬兰奥卢大学的CMVS团队在2013年发布了SMIC数据集<sup>[? 1]</sup>，包括8名高加索人种和8名蒙古利亚人种，同时数据集中包含了高速视频数据（High speed video, HS）、近红外视频数据（Near infrared videos, NIR）和普通彩色视频数据（Normal color video, VIS）；英国曼彻斯特城市大学在2017年发布了SAMM数据集<sup>[? 1]</sup>，是目前发表最新的数据集，包括了几乎全部的人种（蒙古利亚人种、高加索人种、尼格罗人种和大洋洲人种）和均衡的性别比，但其数据集只包含了高速灰度视频数据（见表3）。优秀的数据集提供了良好的实验基础，自动微表情识别系统的研究主要集中在微表情检测（Micro-expression Spotting）和微表情识别（Micro-expression Recognition）。

微表情的检测指在一个图像序列（视频帧）中检测微表情发生的起始时间点。有许多研究致力于类似的任务，如检测普通的人脸表情、眨眼和人脸AU<sup>[? ? ?]</sup>，以及各种有效的算法被提出，与微表情识别研究相比，微表情检测的研究较少。由于缺乏自发的微表达数据，大多数早期的微表达检测研究多采用摆拍的微表情数据。Shreve等首次先提出了一种基于应变的光流方法来识别视频中的宏表情(普通人脸表情，微表情的反义词)和微表情，在USF-HD数据集（摆拍的数据集）中进行了测试，同时他们还对从在线视频中收集的28个微表情进行了测试，但是这个数据集很小，没有发表<sup>[? ?]</sup>。在另一组研究中，Polikovsky等人提出了一种新的微表情检测方法，他们使用三维梯度直方图作为特征描述符，对中性面部微表情帧进行不同阶段(起始、顶峰和终止)分类<sup>[? ?]</sup>。在他们的研究中，将微表情检测任务作为分类任务，并训练模型根据所涉及动作的阶段将视频片段分为四类。关于Polikovsky的研究，有一点很好，那就是作者试图在一个精细的层面上构建微表情的时间范围。这可能适用于摆拍的微表情片段，因为它们具有相似的时间结构，但不适用于自发的微表情，因为在真实场景中，微表情在其时间范围上存在显著差异。分类任务需要对视频预分割，这在实际应用中也是一个问题。Wu等人提出利用Gabor滤波器构建微表情识别系统进行微表情检测<sup>[? 1]</sup>。该方法在METT训练数据上进行了测试，取得了良好的效果。但有一点需要说明的是，METT训练样本完全是合成的视频片段（在一系列相同的中性人脸图像中间插入一张有情感的人脸图像）。在这些视频片段中，表情的“开始”和“结束”非常的尖锐和突然，上下文界线也非常清晰，所以它们根本不能代表计算机真正的微表情检测问题。

虽然上述研究可能会为微表情检测提供潜在的贡献，但一个主要的缺点是，它们仅在摆拍的(或METT等合成的)微表达数据集上进行了测试。与自发微表情相比，提出的数据更容易完成微表情的检测任务。摆拍或合成的微表情数据通常具有相似的时间范围结构，即人为控制的起止时间点，这与自发微表情之间存在显著差异。考虑到视频的上下文，摆拍的微表情片段通常在视频中会禁止无关的动作，所以摆拍的微表

情片段通常具有更清晰的上下文。自发微表情视频的情况更加复杂，由于普通人脸表情(包含相同或相反的情绪价<sup>[? 1]</sup>)、眨眼和其他头部动作也可能发生，并且还可能与自然情绪相互重叠。这些在自发微表达数据中遇到的挑战，在以往的研究中利用摆拍的微表情数据都无法解决，因此利用自发微表情数据进行微表情检测还需要更多的工作。论文[26]和其他几项研究使用了一种更简单的方法，即微表达“探测”来解决这个问题<sup>[27? ?]</sup>。在这些研究中，微表情检测被视为二分类问题，一组带标签的微表达片段与另一组非微表达片段进行分类。这些研究都是在自发微表情数据集中测试的，这是一个很大的优点。但是对于分类任务，训练和测试视频都需要进行适当的分割，这在实际应用中可能会遇到困难。二分类方法与直接从长视频中提取自发微表情的实际应用目标仍有较大差异。Li等人提出了一种将特征差异（Feature Difference, FD）比较和峰值检测（Peak Detection, PD）相结合的微表情识别框架，这种方法是第一个用于真实微表情数据集检测微表情且行之有效的方法<sup>[26]</sup>。在最近的另一篇文章中作者提出了一种基于概率框架的随机游走模型（Random walk model）来检测微表情，通过几何变形建模从视频片段中识别自发微表情片段，该方法被证明对SMIC和CASMEII都是有效的<sup>[?]</sup>。

微表情识别的任务类似于普通人脸表情识别，差异处在于微表情片段(包含微弱面部运动的起始到终止的帧序列)的标签，具体是指根据表达的情感内容训练一个分类器将其分类为两个或两个以上的类别(例如快乐、悲伤等)。在众多已发表的文献中微表情识别的研究比微表情检测的研究更为突出，在摆拍和自发的数据集上提出并测试了很多算法。早期的工作都是从摆拍的微表情数据开始的。Polikovsky和Kameda等人采用三维梯度描述符对AU标记的微表情进行识别，将提出的方法在自己采集的摆拍微表情数据上进行了测试<sup>[?]</sup>。Wu等人将Gentleboost和支持向量机（Support Vector Machine, SVM）分类器结合，在METT训练工具中识别合成的微表达样本。随着自发微表达数据库的出现，利用自发微表达数据库进行微表达识别的研究也越来越多。2011年，论文[26]提出了第一个微表情识别方法，并在第一版的SMIC数据集上取得了很好的效果（第一版的SMIC数据集包含77个自发的微表情样本）。该方法采用时间插值模型(temporal interpolation model, TIM)对微表情计算帧数，将其与和多核学习（Multiple Kernel Learning, MKL）结合捕获图像序列的主要变化，使用三正交平面的局部二值模型（Local Binary Patterns on Three Orthogonal Planes, LBP-TOP）特征作为描述符提取动态纹理作为微表情识别的特征，再用随机森林（Random forest, RF）作为分类器分类。这是首次尝试构建一个可行的方法来识别这些细微的面部行为，并在77个微表情样本上取得了很好的效果(二分类准确度为71.4%)。同样的方法在新版本的SMIC数据集上进

行了测试，在包含164个微表情的SMIC-HS数据集上得到了48.78%(三分类)的识别结果。此后，论文[26]的研究结果被许多其他研究人员引用为微表情识别研究的基准。Ruiz-Hernandez和Pietikäinen等人利用二阶高斯流的再参数化来生成更鲁棒的直方图，在第一版SMIC数据库上得到了比论文[26]更好的识别结果<sup>[27]</sup>。Huang等人通过在时空域中的积分投影获得被试者（目标）的形状属性，将形状属性与时空域上的纹理信息结合组成新的特征，提出了时空局部量化模式(SpatioTemporal Completed Local Quantization Patterns, STCLQP)作为微表情识别的特征，在SMIC上实现了64.02%的精度<sup>[?]</sup>。

几个其他的微表情识别研究使用了CASMEII微表情数据库。Wang等人从张量独立颜色空间(非普通的RGB颜色空间，Tensor independent color space, TICS)中提取LBP-TOP进行微表情识别，并在CASMEII数据集进行测试<sup>[31]</sup>。Wang等人在其另一篇论文中，将局部时空方向特征与鲁棒主成分分析（Principal Component Analysis, PCA）的稀疏部分相结合一起用于微表情识别，在CASMEII上实现了65.4%的准确率<sup>[28]</sup>。Wang等人提出利用6个交叉点的局部二值模式（LBP-Six Intersection Points, LBP-SIP）进行微表情识别，并在CASMEII和SMIC上进行了测试，该方法是减少了LBP-TOP中的冗余信息<sup>[29]</sup>。Huang等人为了提高微表情的辨别力，提出一种新的基于Laplacian的特征选择方法，在已发表的数据集中得到了很好的识别效果<sup>[35]</sup>。Wang等人提出了一种紧实的LBP-TOP描述符（Super-compact LBP-Three Mean Orthogonal Planes, MOP），MOP所描述的紧实鲁棒形式不仅保留了基本模式而且减少了影响编码特征判别的冗余<sup>[30]</sup>。Hong等人为提高LBP-TOP在时空信息上的计算效率，引入了张量的概念，这加速从三维空间到二维空间的实现过程<sup>[51]</sup>。

除LBP及其变体外，部分研究人员致力于对光流特性的研究。如Liu等人提出利用主方向平均光流特征（Main direction Mean opticflow, MDMO）进行微表情识别，同时还考虑了局部统计运动和空间位置信息，在SMIC和CASMEII数据库上都取得了良好的性能，但是他们只使用了SMIC的第一个版本，而不是SMIC的完整版<sup>[36]</sup>。Liong等人从光学应变量值得到一个时间段内人脸的细微相对位移量，并对局部特征赋予不同的权重，形成新的特征<sup>[33]</sup>。Xu等人利用光流估计对微表情图像序列选择的粒度进行像素级对齐，得到主光流方向，将其作为精细的面部动态特征描述符<sup>[38]</sup>。

近年来，对微表情识别的研究十分活跃。到目前为止，大多数提出的方法都考虑使用基于纹理的特性来完成任务。时空纹理特征是描述面部运动的合适选择，但单独使用它们可能不足以进行微表情识别，识别性能仍有很大的提升空间。Li等人将低强度的微表情视频经过欧拉视频放大，在三个正交平面上利用不同的特征提取符提取特征对微表情进行识别。随后Li等人基于LBP-TOP的思想在三个正交平面上扩展了梯度

方向直方图（HOG）和图像梯度方向直方图（HIGO）提出了HOG-TOP和HIGO-TOP。Song等人通过从面部和身体微弱运动中学习的稀疏编码来识别情绪，他们的微表情定义更加广泛，将身体部位（脸部除外）的姿势包括在内<sup>[2]</sup>。Ngo等人提出了一种用于微表情图像序列预处理的选择性转移机（Selective Transfer Machine, STM），用于解决数据库中不平衡和不同面部形态的问题<sup>[32]</sup>。Lu等人发现微表情的图像序列在时空域中基于Delaunay三角归一化，提出了基于Delaunay的时间编码模型（Delaunay-based temporal coding model, DTOM）<sup>[37]</sup>。Oh等人通过Riesz小波变换获得多尺度单原信号，提取其幅值、相位、方向特征组成新的特征描述符进行微表情识别<sup>[34]</sup>。He等人提出了一种多任务的中层特征学习方法进行特征提取，该方法能够获得更具识别能力和泛化能力的中层特征<sup>[39]</sup>。由于深度学习模型需要大量的数据进行训练，而现有的微表情数据还远远不够，但最近，Patel等人提出了一种利用深度学习模型解决微表情识别问题的方法<sup>[?]</sup>。作者建议选择性使用基于普通人脸表情数据库训练的卷积神经网络(Convolutional Neural Networks, CNN)模型的深度特征，但效果并不理想。Li等人发现微表情的峰值帧能够表达更丰富的情感，他们提出了一种在峰值帧应用深度神经网络的方法检识别表情，在CASME II上取得了不错的成果<sup>[?]</sup>。表 1 对目前提出的大多数方法做了列举。

表 1 微表情的研究方法

摆拍微表情数据集（Posed）		自发微表情数据集（Spontaneous）	
微表情检测	微表情识别	微表情检测	微表情识别
3D梯度	Gentleboost和SVM	特征差异与峰值检测	TIM+MKL+LBP-TOP+RF 二阶高斯射流的再参数化 从张量独立色彩空间中提取LBP-TOP 局部时空方向特征+鲁棒PCA 低强度的微表情视频经过欧拉视频放大 基于普通表情的迁移学习
			LBP-SIP、STM、MOP、STCLQP、DTOM Riesz小波变换、光流估计 引入张量概念的Fast LBP-TOP MDMO+局部运动+空间位置 HOG/HIGO-TOP
Gabor滤波器	人脸与身体微弱运动结合	基于概率的随机游走模型	
应变的光流方法			

### 1.3 本文的研究内容

本文简单介绍了微表情识别的研究现状和基本方法，以及其不容小觑的应用价值。然而，目前的微表情研究都是基于高质量数据集的基础上展开的，例如SMIC数据集的人脸分辨率为 $190 \times 230$ 像素，最新发布的SAMM数据集的人脸分辨率达到 $400 \times 400$ 像素之高，但在实际应用中由于图像采集设备的机能限制，难免会遇到低像素的视频数据，这严重影响了几乎所有的微表情识别算法的性能，为了解决这一问题，本文提出了专门针对低分辨率环境下的微表情识别方法：按照常规方法对视频做预处理，然后再使用超分辨重建技术从低维图像中近似的重建出高维图像，对重建出的高维图像进行微表情识别，最后比较重建后的微表情识别效率与未重建前的低维图像的识别效率，同时分别从传统机器学习和深度学习两个角度介绍了系统框架和识别结果。

本文共分为六章，内容具体安排如下：

第一章：绪论。阐述了微表情研究的背景及意义，概括了国内外最新的微表情研究现状，最后对文章的整体结构做出安排；

第二章：相关工作。简单介绍了微表情的概念和微表情数据集，总结了微表情识别在传统方法和深度学习方法的进展，阐述了低分辨率微表情的识别的意义和最新进展；

第三章：基于传统方法的低分辨率环境下微表情识别的研究。主要介绍了人脸对齐与分割，其中包括对单一目标检测中主动形状模型（Active Shape Model，ASM）准确度的算法改进，图像序列的超分辨重建，LBP-TOP和SVM；

第四章：基于深度学习的低分辨率环境下微表情识别的研究。主要介绍了伪3D残差网络（Pseudo-3D Residual Networks，P3D ResNet），数据增强以及实验分析；

第五章：低分辨率环境下微表情识别可视系统。通过对需求分析设计出系统的时序图和功能图，最后设计出功能完善的可视化系统；

第六章：总结与展望。总结全文的工作，分析不足和展望未来前景。



## 第二章 相关工作

本章的主要内容分为三个部分，第一部分主要介绍微表情研究中使用的数据集以及微表情和常见的宏表情之间的差异，第二部分介绍近几年微表情研究中所使用的特征提取的几种方法，第三部分介绍本文将在第四章中使用的相关深度网络的基础知识。

### 2.1 宏表情和微表情

普通的人脸表情可以是自然产生的，也可以是根据需要人为的摆拍出来。摆拍的人脸表情是有意表现出某种情绪，而自然产生的人脸表情是在表达自己的真实情感。所以在人脸宏表情（普通人脸表情）的研究中，既有自然产生的表情数据，也有摆拍的表情数据，微表情研究也存在类似的问题。在微表情研究领域用“自发”这个词来强调微表情的产生(或被诱导)是自然的，因为参与者的内心存在实际的情感。

所以根据表情的产生方式将微表情分为“自发”产生的自然微表情和“摆拍”的微表情。目前摆拍的数据集主要有USF-HD数据集和Polikovsky数据集，自然状态下诱发产生的自发数据集，主要包括SMIC、SMIC2、CASME、CASME II和SAMM等数据集，本节将简单介绍几种数据集的产生方式和优缺点，同时列举典型的宏表情数据集做对比，明确微表情的概念。。

#### 2.1.1 宏表情数据集

随着图像识别和表情研究的深入，宏表情数据集的建立越来越完备。目前常用的人脸宏表情数据集有日本女性表情库（The Japanese female facial expression, JAFFE）<sup>[2]</sup>、Yale表情数据集<sup>[3]</sup>和CK+人脸表情数据集<sup>[4]</sup>等。本文提出的算法中没有用到宏表情，此处以最常用的CK+数据集举例只为说明宏表情和微表情之间的差异。

CK+数据集是在Cohn-Kanade Dataset的基础上扩展来的，发布于2010年。2000年，Cohn-Kanade（CK）数据集发布，目的是促进人脸面部表情的自动检测研究。从那时起，CK数据集成为算法开发和评估中使用最广泛的测试平台之一。在此期间，产生了三个明显的局限性：1)虽然AU编码经过了很好的验证，但是情感标签是错误的，因为标定的标签是被要求的，而不是实际产生的；2)对新算法的评估缺乏一个通用的性能指标；3)不包含常见数据集的标准协议。因此，CK数据集被用于AU和情感检测时缺少与基准算法的比较，并且使用原始数据集的随机子集使得对元数据的分析变得困难。为了解决这些问题，发布了扩展版的Cohn-Kanade数据集CK+。CK+数据集中的表

情分为三类，包括了摆拍和自发的表情以及其他类型的元数据（Metadata）。对于摆拍的表情，序列的数量比第一版的增加了22%，参与者的数量增加了27%。与初始版本一样，每个序列的目标表情都是完整的FACS编码，情感标签也经过了修改和验证。此外，元数据中还添加了经过验证的情感标签。数据集中还包含使用主动外观模型（Active Appearance Models, AAMs）和线性支持向量机分类器给出的基线结果，使用留一法交叉验证对摆拍的数据进行AU和情绪检测。

CK+数据集包括123个参与者，593个图像序列，每个图像序列的最后一张帧有AU标签，而在这593个图像序列中，有327个序列有情感标签。CK+数据集是人脸表情识别中比较流行的一个数据集，该数据集将情感分为8类，包括中性，厌恶，愤怒，蔑视，恐惧，高兴，悲伤，惊讶。图1给出了该数据集中的样本图片。综合来说，CK+是目前较为理想的表情数据集。

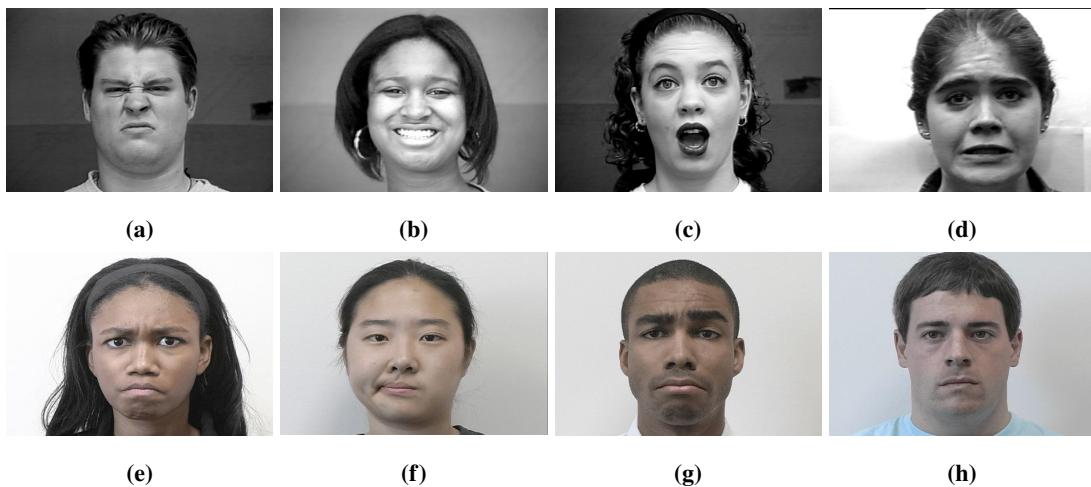


图1 人脸宏表情样本示例

(a)厌恶、(b)快乐、(c)惊讶、(d)恐惧、(e)愤怒、(f)轻蔑、(g)沮丧、(h)中性表情

### 2.1.2 早期微表情数据集

由于微表情在镜头下很难产生，所以微表情数据的缺乏是微表情研究的第一障碍。虽然微表情已经被心理学家研究了很长一段时间，但网络上广泛传播的微表情样本仅仅只有片段，并没有发现任何心理学研究小组分享的大数据集。作者认为第一个原因是心理学研究更关注微表情本身的性质，比如什么时候出现或者看起来是什么样子，所以他们不需要像使用计算机研究那样需要大量的微表情数据。其次，在某些情况即使涉及到大量的微表情数据，但由于保密性限制，数据无法公开共享，例如患者的病历或司法讯问记录等。

以2009年为时间点，在此以前属于微表情的早期研究阶段，一些研究人员在他们的研究中使用了摆拍的微表情数据，这些数据避免了自发微表情数据获取时的困

难，作为早期自动微表情识别研究的尝试，使用摆拍的数据是一次历史性的突破。例如Shreve等人收集了一个名为USF-HD的数据集，其中包含100个摆拍的微表情片段，视频长度平均在1分钟左右，最长2分钟，最短20秒。研究人员要求参与者模仿屏幕上显示的微表情样本作为数据的来源，同时在他们的文章中明确写到可以通过要求参与者尽可能快地模仿来收集一个摆拍的微表情数据集。Polikovsky等人也收集了一个摆拍的微表情数据集，要求受试者在低强度下表演七种基本情绪，并尽快回到中性表情，数据由一台每秒200帧的高速摄像机记录。表2列出了摆拍的微表情数据集的详细属性。

**表2 摆拍微表情数据集**

	USF-HD	Polikovsky
微表情片段	100	N/A
参与者	N/A	10
分辨率	720 × 1280	480 × 640
FPS	29.7	200
FACS	NO	YES
表情类	N/A	7
人种	N/A	3

值得注意的是摆拍的微表情数据不可以代替或与自发的微表情数据一起使用，因为这两种数据是不同性质的。例如在研究微表情的起始点时，由于两者是在不同的机制下产生的，所以摆拍的微表情在时空特性上与自发的微表情存在很大的差异，其次在研究视频的上下文时，由于模仿的表情与通过视频编辑生成的摆拍微表情片段通常在起止点很突兀，而且期间禁止其他无关的动作发生，这也与自发的微表情有很大的不同。另一方面，自然环境下自发的微表情可能伴随着复杂的场景，比如头部运动和眨眼等动作。基于这些事实，利用摆拍的微表情数据进行研究并不能真正解决实际中微表情分析的自动化问题，所以努力收集自发的微表情数据是后续工作的正确路径。需要说明的是在本文接下来的内容中，所有的工作都是关于自发微表情的，如果没有特别说明，“微表情”一词表示自发微表情。

### 2.1.3 自发微表情数据集

#### A. SMIC数据集介绍

2011年，论文[26]首次提出了一种诱导和收集自发微表情的方法，将获得的数据集命名为“自发微表情语料库”，简称SMIC，它是第一个使用自然诱发状态的微表情数据集，对后续的数据集建立具有很好的指导性意义。第一版的SMIC数据集只包含了6位参与者的数据，论文[?]对其进行了扩充，包含了16名参与者的164个自发微表情片段，

由三种相机记录的3个数据集组成：100fps的高速相机记录的HS数据集、25fps的普通彩色相机记录的VIS数据集和25fps的近红外摄像机记录的NIR数据集，且所有数据集具有相同的图像分辨率 $640 \times 480$ 。增加VIS和NIR摄影机有三个考虑：（1）提升数据集的多样性；（2）研究高速相机在微表情分析方面是否优于普通速度相机；（3）研究时间插值方法是否可以应用于普通高速相机，以解决相机的短时插值问题。图2给出一个消极的微表情序列，通过该样例可以看出，一个微表情是由一组图片序列构成的，本样例的变化主要表现为嘴角的细微下沉。



图2 一个消极的微表情片段示例（SMIC数据集）

### a) 视频采集

研究表明通过图像、视频和音乐等一定的刺激能够诱发真实的表情<sup>[2]</sup>。自发微表情是由人的内心感受触发的非自愿行为，所以可以使用上述方法引发情感反应。但必须找到一种能确保诱发的表情足够短的方法（满足微表情的标准）。一些心理学著作研究了微表情发生的条件，Ekman等人认为当人们试图隐藏自己的真实情感时，尤其是当被抓住后果会很严重时，微表情就会出现，这被称为高风险条件，例如嫌疑人正在接受警察或测谎专家的审问，这是一种很自然的高风险场景。但是对于采集数据而言，营造真正的审问现场显的不太可能。所以为了诱导参与者自发的微表情，需要找到一种模拟高风险环境的方法。设计的情景必须满足以下两个要求：（1）激发参与者情绪的刺激必须有效，使激发的情绪反应强烈到无法完全隐藏；（2）应该制造高压力，这样参与者才会有动力去尽力隐藏自己的真实感受。

作者使用了心理学研究中诱导抑制情绪产生的方法。采集前向参与者详细说明研究内容和过程，显示器上显示如下提示性语句：“（1）将向您展示几段诱发情绪的短片，请尽量保持头部稳定并仔细观看。（2）每段视频片段后，您将有一个短暂的休息。请根据您对刚才看到的视频的真实感受填写调查表（报告中的情感反馈是标注过程中的重要参考。）。（3）当您在看视频的时候，我会待在另一个房间，通过摄像头观察您的面部和身体动作，并尝试猜测您所看的视频片段(片段是随机播放的)。您的任务是装出若无其事的样子，而不是表露您的真实感情。如果您不能隐藏您的感觉，您将不得不填写一份超过500个冗长而乏味的问题问卷。”

在每段影片结束后，参与者将在问卷中回答以下问题：“（1）您在看视频时感受到

了什么样的情绪(快乐、悲伤、厌恶、恐惧、惊讶、愤怒或困惑)? (2) 看视频的时候您是否感觉到愉悦? (从1到7愉悦程度逐渐上升)”。选择最有效的刺激作为情感诱导因子是后续数据采集的关键因素之一。通过查阅文献比较不同类型的情绪诱导材料,如图像、音乐、视频和互动。最终决定使用短视频作为微表情采集诱导剂的原因有三个: (1) 视频包含音频和视觉信息, 因此比图像和音乐的影响更强大; (2) 视频能够持续一段时间, 更能激发强烈情绪, 更容易产生微表情; (3) 从获取稳定额叶面部视频的实际角度来看, 观看视频的参与者比多人参与的互动场景更容易控制。

20名来自奥卢大学的学生和研究人员自愿参与数据集的制作, 参与者的年龄从22岁到34岁不等, 其中7位是女性, 13位是男性, 9人是白种人, 11人是亚洲人。研究人员在电脑显示器上向参与者展示了16段精心挑选的能引发强烈情绪的视频片段。当参与者观看视频片段时, 三个固定在电脑显示器上的摄像头记录参与者的面部反应, 操作人员在前方通过另外一台电脑监控参与者的面部反应, 图3给出了环境设置示意图。



图3 微表情数据采集示意图

### b) 视频的标注

录制的视频需要进行分割和标注, 目的是得到适合研究人员进行训练与测试的微表情样本和相应的标签。采集的三种视频中高速视频由于具有最佳的时间分辨率所以非常适合标注, 另外两个摄像头拍摄的视频需要同步后再进行标注。首先, 从原始的长视频中分割出微表情的起始和终止帧, 微表情序列的开始表示的是与之前中立(或接近中立)的人脸表情相比的可见运动的第一帧, 而微表情序列的终止指在与下一帧相比可以发现任何运动时结束的最后一帧。关于微表情精确的长度限制目前还存在争议, SMIC数据集参考论文[6]和[7]等人的建议, 设置了1/2秒较宽松的分割节点。注意, 并不是所有的微表情都结束于一个完全的中性表情, 有些表情可能会上升, 然后下降到接近中性的状态, 并保持这种状态很长时间, 这也被认为是微表情的结束。

随后，对所有的视频片段都用情感标签进行标注。情感标签的证据有两种：视频片段的内容和参与者提交的报告。虽然在视频播放前已经预知将可能产生某种确定情绪，但研究发现，参与者在某些视频刺激下可能会产生不同的情绪表现(甚至相反的情绪)。在少数情况下，当参与者提交的问卷报告与视频内容相反对时(例如，一些参与者反馈在观看恐怖视频片段时感到快乐或有趣)，SMIC数据集使用参与者的提交的报告作为微表情标签的标准。起初，SMIC数据集根据视频内容分配了五种情感标签，包括快乐，悲伤，恐惧，厌恶，惊讶。后来，SMIC数据集将五类标签合并为三类：积极（Positive）、惊喜（Surprise）和负面（Negative）。将第一版本的快乐类别改为新的积极类别，而消极类别则是由第一版的悲伤、恐惧和厌恶这三个类别组合而成。将三种消极情绪融合在一起的原因是：首先，参与者在该段视频的报告中选择了三种情绪中的一种以上；其次，三种标签的样本量都太小，合并后会更好的平衡。根据具体情况，惊喜类别可以分为正面惊喜和负面惊喜。同时为了验证数据标签的有效性，标注由两位标注者分别执行。然后，两位标注者相互交换检查各自的标注，只有当两位标注者的标注结果一致时的标签有效。

### B. CASMEC数据集介绍

在SMIC发表后不久，另一组研究人员收集了新的自发微表情数据集。由中国科学院Yan等人采用与SMIC相似的情绪诱导方式收集了中国科学院微表情数据集(CASME)，包含19位中国籍参与者的195个微表情片段。CASME由两台相机记录，一台是明基M31相机，帧率为60fps，分辨率为 $1280 \times 720$ (CASME-A)，另一台是灰点GRAS-03K2C相机，帧率为60fps，分辨率为 $640 \times 480$  (CASME-B)。CASME数据集中的微表情首先使用AU标记，然后被分为八类情绪，包括娱乐、悲伤、厌恶、惊讶、蔑视、恐惧、压抑和紧张。之后又发布了第二版数据集CASME II，CASME II提供了更多具有更高时空分辨率的微表情样本。新数据集的平均人脸尺寸为 $280 \times 340$ ，每秒200帧，是从26名中国参与者中获得的247个微表情样本。CASMEII样本有五个类的AU标签和情感标签，即幸福、厌恶、惊讶、压抑和其他。

CASME II数据集相比之前的数据集引入了AU标签，它是在充分考虑了主客观因素（除了FACS编码的基本判断）外，还参考了参与者自己的主观回忆来辅助标注的样本标签。除此之外，该数据集还对微表情的起始帧、峰值帧和终止帧都做了详细的标注。

### C. 其他数据集介绍

最近又有一个新的自发微表情数据集SAMM发布，它也使用了类似于SMIC和CASME的情绪诱导方法，从13个不同民族的32名参与者中获得159个微表情。SAMM数据具有更

高的帧分辨率 $2040 \times 340$ , 帧速率为200fps。数据提供了AU标签和七种表情标签, 包括生气、开心、蔑视、恐惧、惊讶、厌恶和其他。

表 3 列出了当前所有提到的微表情数据集的详细参数。

表 3 自发微表情数据集

	SMIC-HS	SMIC-subHS	SMIC-NIR	SMIC-VIS	CASME II	SAMM
微表情片段	164	71	71	71	247	159
参与者	16	8	8	8	26	32
分辨率	640 × 480	640 × 480	640 × 480	640 × 480	640 × 480	2040 × 1088
人脸分辨率	190 × 230	190 × 230	190 × 230	190 × 230	280 × 340	400 × 400
FPS	100	100	100	100	200	200
性别比 (F/M)	6/10	2/6	2/6	2/6	15/11	16/16
FACS	NO	NO	NO	NO	YES	YES
表情类	3	3	3	3	5	7
平均年龄 (SD)	26.7 (N/A)	26.7 (N/A)	26.7 (N/A)	26.7 (N/A)	22.03 (SD=1.6)	33.24 (SD=11.32)
人种	2	2	2	2	1	4

#### 2.1.4 宏表情与微表情比较

通过前两小节的介绍可以看出微表情和宏表情最大的区别是在面部肌肉的运动强度和持续时间。如图 4 所示，人脸根据不同的部位被划分为不同的肌肉组如皱眉肌、降眉间肌、鼻肌等，图 1 中上述肌肉运动明显比图 2 更加剧烈也更加夸张；另一方面，微表情的持续时间一般在  $1/25$  秒到  $1/2$  秒之间，而宏表情则不然。

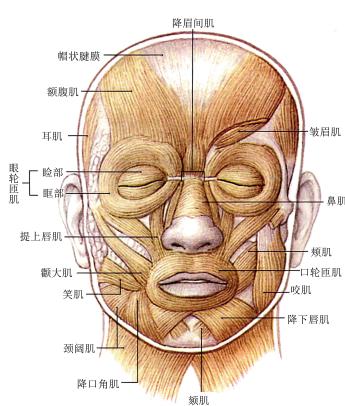


图 4 人脸面部表情肌肉划分

目前存在的宏表情和微表情数据集种类繁多且各具特色，但因为各类研究目的的差异，各类数据集间的迁移效果并不显著，存在着很大的局限性和个体差异性，存在的问题集中表现为以下几点：

首先，各个数据集建立标准不统一。对于常用的宏表情数据集，例如JAFFE和CK+两个宏表情数据集，他们在情感分类的标准上存在差异，JAFFE数据集将表情分为开心、愤怒、沮丧、讨厌、吃惊、恐惧和中性七类，而CK+数据集则在此基础上加入了轻蔑。同样与微表情数据集的分类方式也不一样，例如SMIC微表情数据集将表情分类为积极、消极和惊讶三类，CASME系列数据集将表情分类高兴、惊讶、悲伤、厌恶、恐惧、抑制和其他七种类型。另一方面，对于微表情数据集，不同的数据集之间诱发方式可能存在差异，例如有的数据集使用自然诱发（SMIC、CASME），有的使用非自然状态下的模拟方式（USF-HD、Polikovsky）。此外差异性还表现在SMIC数据集没有对AU单元进行标注、CASME数据集的光照条件为。所以不同的数据库建立标准导致了难以使用统一的方法对数据集进行性能的评价。

其次，样本数量有限。目前各个数据集的样本数量较少，尤其是参与者的数量有限，这使得某些微表情现象不具有普遍意义，同时也难以提出较强的可说服性理论。

## 2.2 微表情特征提取的一般方法

### 2.2.1 基于纹理信息的LBP-TOP特征提取符

局部二值模式（Local Binary Pattern，LBP）是一种用来描述图像局部纹理特征的运算符，它具有旋转不变性和灰度不变性等显著的优点，在人脸识别方面被广泛应用<sup>[?]</sup>。简单来说，就是对图像中的某一像素点的灰度值与其邻域的像素点的灰度值做比较，如果邻域像素值比该点大，则赋为1，反之，则赋为0。从图像的左上角开始，形成一个LBP编码，然后将该编码转换为一个十进制数。通过这种转换，可以将一个像素点与邻域的差值关系用一个数表示。

因为LBP特征记录的是像素点与邻域像素的差值关系，所以光照变化只会引起像素值的同增或同减，不会改变LBP值的大小，特别是在局部区域，所以LBP可以很好的保存图像中像素值的差值关系（如图5所示）。进一步的将LBP做直方图统计，这个直方图就可以用来作为纹理分析的特征算子。

但是LBP只能处理单张的二维图像，对于视频或者图像序列则无能为力，2007年芬兰奥卢大学的Zhao等人提出了一种处理动态纹理的LBP特征提取符LBP-TOP，但是现在已经被广泛用于基于视频的人脸表情识别<sup>[49]</sup>。

LBP-TOP是LBP从二维空间到三维空间的拓展，是在二维图像的X、Y方向上增加了一个沿着时间方向的T轴，形成了互相正交的XY、XT和YT三个平面。以一个图像序列为单位组成一个立方体，对其划分合适的块参数（Blocksize），将一个小块作为一个小单元从三个不同的平面提取LBP特征生成该平面的特征直方图，将三个直方图级联为

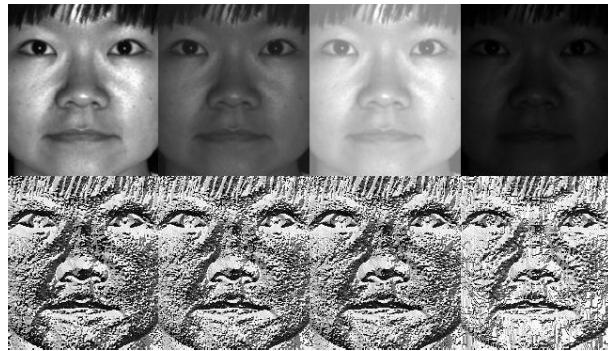


图 5 LBP 特征的光照鲁棒性

一个基于块的直方图，以同样的方式遍历整个图像序列立方体，生成最终的LBP-TOP特征直方图，如图 6 所示，LBP-TOP<sup>1</sup>指一个小单元（图中的红色立方体）的LBP-TOP特征直方图，LBP-TOP<sup>2</sup>指整个立方体的LBP-TOP特征直方图，关于Blocksize的划分将在章节三中讨论。

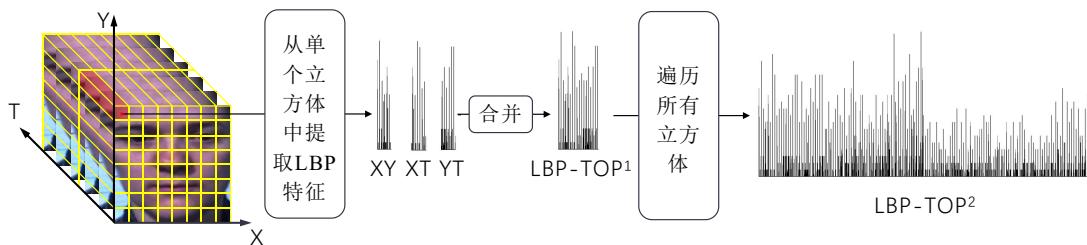


图 6 LBP-TOP 特征提取过程

## 2.2.2 基于光流法的特征提取

二十世纪五十年代心理学家Gibson在其著作[?]中首次提出了环境光（Ambient optic）、环境光阵（Ambient optic array）、光流（Optic flow）和光流阵（Optic flow array）等基本概念。随后，一系列关于昆虫视觉机理方面的实验结果表明，大多数昆虫都可以通过光流测量自身的运动，进而通过积分获得自身飞行的距离。1976年，Poggio和Reichardt等人在研究昆虫视觉时提出了关于光流的粗略计算形式[? ]，1981年Horn和Lucas等人将二维速度场与灰度相联系，引入了光流约束方程，为光流计算做出了奠基性的工作[? ?]。

光流是空间运动物体在观察成像平面上的像素运动的瞬时速度，是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系，从而计算出相邻帧之间物体的运动信息的一种方法。一般而言，光流是由于场景中前景目标本身的移动、相机的运动，或者两者的共同运动所产生的。Barron等人对多种光流计算技术进行了总结，按照理论基础与数学方法的区别把它们分成四种：（1）基于区域或者基于特征的匹配方法；（2）基于频域的方法；（3）基于梯度的

方法；（4）基于能量的方法<sup>[2]</sup>。光流的研究是利用图像序列中的像素强度数据的时域变化和相关性来确定各自像素位置的“运动”，即研究图像灰度在时间上的变化与景象中物体结构及其运动的关系，将二维图像平面特定坐标点上的灰度瞬时变化率定义为光流矢量。光流场（optical flow field）是运动场在二维图像平面上的投影（运动场，物体在三维真实世界中的运动），是一个二维矢量场，包含的信息是图像中像素点的灰度值发生变化趋势的瞬时速度矢量信息，那通俗的讲就是通过一个图像序列，把每张图像中每个像素的运动速度和运动方向找出来就是光流场。研究光流场的目的就是为了从图像序列中近似计算出不能直接得到的运动场。图 7 提供了光流场中光流的大小和方向的表现。

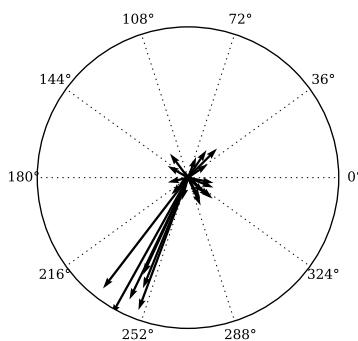


图 7 光流大小和方向的表现

光流法检测运动目标，其基本思想是赋予图像中的每一个像素点一个速度矢量，从而形成了该图像的运动场。图像上的点和三维物体上的点在某一特定的运动时刻是一一对应的，根据各像素点的速度矢量特征对图像进行动态的分析。若图像中不存在运动目标，那么光流矢量在整个图像区域则是连续变化的，而当物体和图像背景中存在相对运动时，运动物体所形成的速度矢量则必然不同于邻域背景的速度矢量，从而将运动物体的位置检测出来。但光流法的计算过于复杂，而且在实际情况中由于各种复杂的环境因素常常会导致计算结果出现很大的误差，故对它的使用必须是在建立一定前提假设的基础之上的：（1）相邻帧之间保持亮度恒定不变；（2）相邻帧的取帧时间保持连续或物体运动幅度小；（3）具有空间一致性，同一子图的像素点具有相同的运动。

光流不能由运动图像的局部信息来唯一的确定，例如，亮度等值线上的点或者亮度比较均匀的区域都无法唯一的确定其点的运动对应性，但是运动是可以进行观察得到。当人的眼睛观察运动物体时，物体的景象在人眼的视网膜上形成一系列连续变化的图像，这一系列连续变化的信息不断“流过”视网膜（即图像平面），由于它包含了目标运动的信息，因此可被观察者用来确定目标的运动情况。由此说明运动场和光流

不一定是唯一对应的，即光流不一定是由物体运动产生的，反之如果物体发生了运动也不一定就能产生光流。但是一般情况下，表观运动和物体真实运动之间的差异是可以忽略的，可以用光流场代替运动场来分析图像中的运动目标及其相关的运动参数。

### 2.2.3 基于Gabor的特征提取

Gabor小波与人类视觉系统中简单细胞的视觉刺激响应非常相似，它在提取目标的局部空间和频率域信息方面具有良好的特性<sup>[? ?]</sup>。虽然Gabor小波本身并不能构成正交基，但在特定参数下可构成紧框架。Gabor小波对于图像的边缘敏感，能够提供良好的方向选择和尺度选择特性，而且对于光照变化不敏感，能够提供对光照变化良好的适应性。上述特点使Gabor小波被广泛应用于视觉信息理解。二维Gabor小波变换是在时频域进行信号分析处理的重要工具，其变换系数有着良好的视觉特性和生物学背景，因此被广泛应用于图像处理、模式识别等领域。与传统的傅立叶变换相比，Gabor小波变换具有良好的时频局部化特性。即非常容易地调整Gabor滤波器的方向、基频带宽及中心频率从而能够最好的兼顾信号在时空域和频域中的分辨能力；Gabor小波变换具有多分辨率特性即变焦能力。即采用多通道滤波技术，将一组具有不同时频域特性的Gabor小波应用于图像变换，每个通道都能够得到输入图像的某种局部特性，这样可以根据需要在不同粗细粒度上分析图像。此外，在特征提取方面，Gabor小波变换与其它方法相比：一方面其处理的数据量较少，能满足系统的实时性要求；另一方面，小波变换对光照变化不敏感，且能容忍一定程度的图像旋转和变形，当采用基于欧氏距离进行识别时，特征模式与待测特征不需要严格的对应，故能提高系统的鲁棒性。

无论从生物学的角度还是技术的角度，Gabor特征都有很大的优越性。研究表明，在基本视觉皮层里的简单细胞的感受野局限在很小的空域范围内，并且高度结构化。Gabor变换所采用的核与哺乳动物视觉皮层简单细胞2D感受野剖面非常相似，具有优良的空间局部性和方向选择性，能够抓住图像局部区域内多个方向的空间频率（尺度）和局部性结构特征。这样，Gabor分解可以看作一个对方向和尺度敏感的有方向性的显微镜。同时，二维Gabor函数也类似于增强边缘以及峰、谷、脊轮廓等底层图像特征，这相当于增强了被认为是面部关键部件的眼睛、鼻子、嘴巴等信息，同时也增强了诸如黑痣、酒窝、伤疤等局部特征，从而使得在保留总体人脸信息的同时增强局部特性成为可能。它的小波特性说明了Gabor滤波结果是描述图像局部灰度分布的有力工具，因此，可以使用Gabor滤波来抽取图像的纹理信息。由于Gabor特征具有良好的空间局部性和方向选择性，而且对光照、姿态具有一定的鲁棒性，因此在人脸识别中获得了成功的应用。然而，大部分基于Gabor特征的人脸识别算法中，只应用了Gabor幅值信息，而没有应用相位信息，主要原因是Gabor相位信息随着空间位置呈周期性变化，而幅值的

变化相对平滑而稳定，幅值反映了图像的能量谱，Gabor幅值特征通常称为Gabor能量特征(Gabor energy features)。Gabor小波可像放大镜一样放大灰度的变化，人脸的一些关键功能区域(眼睛、鼻子、嘴、眉毛等)的局部特征被强化，从而有利于区分不同的人脸图像。Gabor小波核函数具有与哺育动物大脑皮层简单细胞的二维反射区相同的特性，即具有较强的空间位置和方向选择性，并且能够捕捉对应于空间和频率的局部结构信息；Gabor滤波器对于图像的亮度和对比度变化以及人脸姿态变化具有较强的健壮性，并且它表达的是对人脸识别最为有用局部特征。Gabor小波是对高级脊椎动物视觉皮层中的神经元的良好逼近，是时域和频域精确度的一种折中。

通过上面的分析，我们知道了，一个Gabor核能获取到图像某个频率邻域的响应情况，这个响应结果可以看做是图像的一个特征。那么，我们如果用多个不同频率的Gabor核去获取图像在不同频率邻域的响应情况，最后就能形成图像在各个频率段的特征，这个特征就可以描述图像的频率信息了，由于纹理特征通常和频率相关，因此Gabor核经常用来作为纹理特征。

## 2.3 相关深度学习网络

深度学习方法作为机器学习尤其是神经网络方法的一种，在图像处理、视频分析和语音识别等领域产生了许多成功的案例。端到端的神经网络模型能够通过学习大量高维数据进行分类和预测，这减少了人工特征工程的繁杂操作。卷积神经网络作为应用最广泛的深度学习方法之一，目前在很多图像相关领域都处于无可替代的地位。卷积神经网络最早出现在论文[?]的研究中，在过去的几年里，卷积神经网络经过了大量的分层、分块等设计的修改，诞生了许多成功的衍生网络如AlexNet<sup>[?]</sup>、VGG-Net<sup>[?]</sup>和GoogLeNet<sup>[?]</sup>等。最近，有很多将深度学习应用于微表情识别的研究，然而，由于是在小数据集上使用的深度网络，所获得的识别准确度仅在随机水平左右。

### 2.3.1 三维卷积神经网络

二维卷积神经网络被认为是解决图像识别问题的一种强大的模型，但将二维卷积神经网络应用到视频中却并非易事，在视频中应用二维卷积神经网络一个简单的方法就是对每一帧运用卷积来识别，但是这种方法并没有考虑到连续帧间的时间维度信息。一些研究表明，三维卷积能够从空间和时间两个维度提取特征，从而获取多个相邻帧中编码的运动信息。三维卷积神经网络从输入帧中生成多个通道的信息，并结合所有通道的信息得到最终的特征表示。该方法应用到现实环境中的人类行为识别中，在不依赖于人工提取的特性的情况下取得了优异的性能<sup>[?]</sup>。

三维卷积是通过将一个三维核与叠加在一起的多个连续帧形成的立方体进行卷积来实现的。在这个结构中，卷积层中每一个特征映射都会与前一层中多个相邻的连续帧相连，从而获取运动信息。需要注意的是：三维卷积核只能从立方体中提取一类特征，因为在整个立方体中卷积核的权值都是一样的，也就是共享权值。所以我们可以采用多种卷积核，以提取多种特征。对于卷积神经网络有一个通用的设计规则，在后面的层（离输出层近的）特征映射的个数应该增加，这样就可以从低级的特征映射组合产生更多类型的特征。

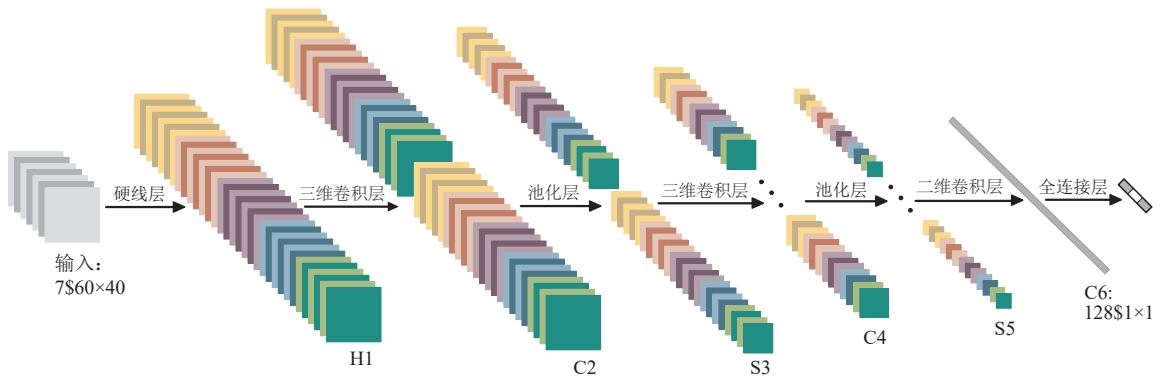


图 8 三维卷积神经网络架构

如图 8 所示，该架构包含一个硬线层、三个卷积层、两个池化层和一个全连接层。每个三维卷积核卷积的立方体是连续的七帧，每帧块大小为  $60 \times 40$ ；第一层应用了一个固定的硬线核去对原始的帧进行处理，产生多个通道的信息，然后对多个通道分别处理。最后再将所有通道的信息组合起来得到最终的特征描述。

H1 为硬线层 (Hardwired)，每帧提取五个通道的信息，分别是灰度、 $x$  和  $y$  方向的梯度， $x$  和  $y$  方向的光流。其中，前面三个都可以每帧都计算。然后水平和垂直方向的光流场需要两个连续帧才确定。

C2 为卷积层 (Convolution)，以硬线层的输出作为该层的输入，用一个三维卷积核在五个通道的每一个通道分别进行卷积。同时为了增加特征映射的个数（实际上就是提取不同的特征），在每一个位置都采用两个不同的卷积核。

S3 为池化层 (Sub-sampling)，得到相同数目但是空间分辨率降低的特征映射。

C4 为卷积层，对五个通道分别采用三维卷积核，同样，为了增加特征映射个数，在每个位置都采用三种不同的卷积核。

S5 为池化层，对每个特征映射采用核进行降采样操作，在这个阶段，每个通道的特征映射都很小。

C6为卷积层，此时对每个特征映射采用二维卷积核进行卷积操作，然后输出为减小到11大小的特征映射。

经过多层的卷积和下采样后，每连续7帧的输入图像都被转化为一个128维的特征向量，这个特征向量捕捉了输入帧的运动信息。输出层的节点数与行为的类型数目一致，而且每个节点与C6中这128个节点是全连接的。最后，采用一个线性分类器来对这128维的特征向量进行分类，实现行为的识别。

### 2.3.2 残差神经网络

增加网络的宽度和深度可以很好的提高网络的性能，深的网络一般都比浅的的网络效果好，但对于原来的网络，如果简单地增加深度，会导致梯度弥散或梯度爆炸。对于该问题的解决方法是正则化初始化和中间的正则化层（Batch Normalization），这样的话可以训练几十层的网络。虽然通过上述方法能够训练，但是又会出现退化问题。在实验中的表现就是虽然网络层数增加，但在训练集上的准确率却饱和甚至下降了（这个不能简单的解释为过拟合现象，因为过拟合时应该在训练集上表现的更好）。通过浅层网络等同映射构造深层模型，结果深层模型并没有比浅层网络有等同或更低的错误率，推断退化问题可能是因为深层的网络并不是那么好训练，也就是求解器很难去利用多层网络拟合同等函数。

针对这个问题He等在2015年提出了一种全新的网络，叫深度残差神经网络（Deep residual neural network），它允许网络尽可能的加深，其中引入了全新的结构，如图9所示<sup>[?]</sup>。

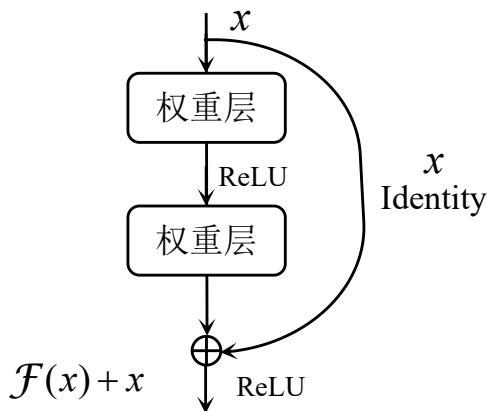


图9 ResNet的残差学习模块

理论上，对于“随着网络加深，准确率下降”的问题，ResNet提出了两种Mapping：一种是Identity mapping，指图9中的弧线，另一种是Residual mapping，指除了弧线以外的部分，所以最后的输出是 $y = \mathcal{F}(x) + x$ 。Identity mapping指自身，也就是公式中的 $x$ ，而Residual mapping指的是“差”，也就是 $y - x$ ，所以残差指的就是 $\mathcal{F}(x)$ 部分。如

果网络已经到达最优，继续加深网络，Residual mapping将被转化为0，只剩下Identity mapping，这样理论上网络一直处于最优状态了，网络的性能也就不会随着深度增加而降低了。

由上可以看出残差网络的主要思想是在网络中增加了直连通道，即Highway network的思想，它允许原始输入信息直接传到后面的层中<sup>[?]</sup>，这样的话这一层的神经网络可以不用学习整个的输出，而是学习上一个网络输出的残差。

传统的卷积网络或者全连接网络在信息传递的时候或多或少会存在信息丢失、损耗等问题，同时还会导致梯度弥散或梯度爆炸，导致很深的网络无法训练。残差网络在一定程度上解决了这个问题，通过直接将输入信息绕道传到输出，保护信息的完整性，整个网络只需要学习输入、输出差别的那一部分，简化学习目标和难度。残差网络的结构可以极快的加速神经网络的训练，模型的准确率也有比较大的提升。残差网络在ImageNet比赛Classification任务上获得第一名，因为它“简单与实用”并存，之后很多方法都建立在ResNet50或者ResNet101的基础上完成，检测、分割、识别等领域都纷纷使用残差网络，Alpha zero也使用了残差网络，所以可见残差网络确实很好用。



### 第三章 基于传统机器学习方法的低分辨率环境下微表情识别

微表情是一种基本的非言语行为，可以忠实地表达人类隐藏的情感。它在国家安全、计算机辅助诊断等领域有着广泛的应用，鼓励我们开展微表情自动识别的研究。然而，从监控视频中获取的图像容易出现质量低下的问题，给实际应用带来困难。由于所捕获图像的质量较低，现有的算法无法达到预期的效果。针对这一问题，我们对面部幻觉法在低分辨率情况下的微表情识别问题进行了全面的研究。实验结果表明，该框架在低分辨率情况下的微表情识别中取得了良好的效果。

目前的微表情识别算法虽然取得了较为合理的效果，但其性能在很大程度上取决于人脸视频剪辑的质量。一旦用于识别的人脸视频片段质量较差(如分辨率较低)，上述算法就无法正常工作。原因主要有两个方面:(1)低分辨率图像丢失了大量的细节信息，导致从低分辨率图像序列中提取可用特征的困难<sup>[41]</sup>。(2)低分辨率图像与高分辨率图像(如不同分辨率、不同清晰度)不一致，导致我们在测试阶段无法直接使用低分辨率图像作为输入。在现实世界中，从监控录像中获取的面部图像通常只占整个画面的一小部分。例如，用于微表情识别的SMIC数据集的面部分辨率为 $190 \times 230$ 。然而，监控视频捕捉到的面部图像序列的分辨率往往在 $50 \times 50$ (或更低)以下。这意味着以前的微表情识别方法不能直接用于处理低分辨率的情况。因此，低分辨率情况下的微表情识别研究具有重要意义和挑战性。

为了解决上述问题，我们利用最近的面部幻觉方法进行了低分辨率微表情识别的研究。我们首先对低质量的人脸图像序列产生幻觉，以恢复丢失的动态特征。然后，我们采用传统的微表情识别方法来探索低质量的微表情识别。我们评估了不同分辨率下微表情识别准确率的表现，研究了分辨率与识别准确率之间的关系。一般来说，本文的目标是全面研究分辨率对微表情识别的影响，同时开发一个框架来处理低质量条件下的微表情识别任务。

低分辨率微表情识别过程包括图像序列预处理、超分辨率重建、特征提取和分类，如图 10 所示。详细描述如下：

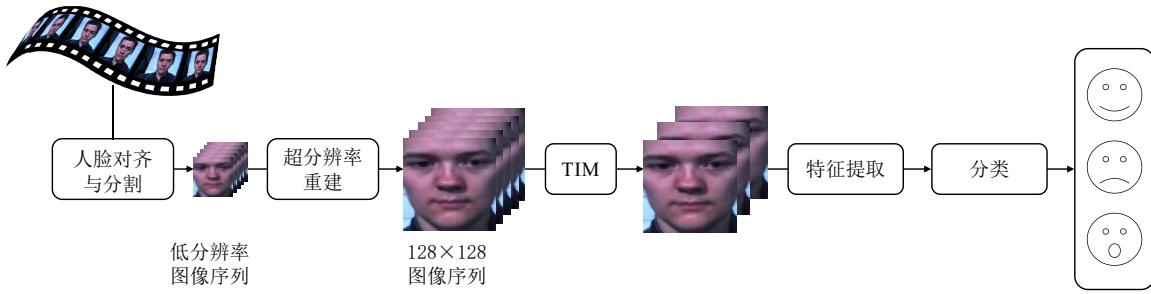


图 10 低分辨率环境下微表情识别框架

### 3.1 低分辨率微表情数据获取

微表情识别的数据集有很多，如SMIC和CASME II。然而，这些数据集都是专业相机在特定环境下获取的高清图像序列。图 11 显示了SMIC-HS数据集的视频片段中的两帧。我们可以在红色框中发现面部表情的细微变化。特别是白色椭圆区域的运动和白色箭头的位置更为明显。如果图像分辨率太低，这些细节很难被注意到。



图 11 来自SMIC-HS数据集的两帧

由于现有的自发微表情数据集中不存在低分辨率的图像序列，我们采用图像恶化处理的方法来获得模拟的低分辨率的微表情图像序列。本文将低分辨率图像分为三大类：小尺寸、低质量和小尺寸放大器、质量较差<sup>[42]</sup>。我们考虑第三种类型的图像(小尺寸、质量差)，即更接近真实应用的情况，如模拟图像。

在图像重建任务中，低分辨率图像序列是通过对高分辨率图像序列进行模糊、下采样和噪声处理得到的<sup>[45]</sup>：

$$\mathbf{L} = \mathbf{DBH} + \mathbf{n} \quad (3.1)$$

其中  $\mathbf{D}$  和  $\mathbf{B}$  分别是下采样和模糊处理， $\mathbf{H}$  是高分辨率图像， $\mathbf{n}$  是加性噪声， $\mathbf{L}$  是低分辨率图像。

### 3.2 数据预处理

在我们提出的框架中，预处理主要包括三个步骤：人脸对齐、人脸分割和TIM。原

始视频中有自然的姿势变化和无意识的运动。同时，收集不同性别、年龄、种族的参与者的微表情视频片段。因此，为了避免上述非表达因子的干扰，进行人脸对齐和人脸分割是必不可少的。

Since MEs are very subtle, other differences (e.g., face size and face shape) between clips need to be minimized in order to reduce intra-class variations and highlight the inter-class differences generated by ME movements. For this purpose we align all faces to a model face in the following way. 由于MEs是非常细微的，为了减少类内变化和突出ME运动产生的类间差异，需要最小化剪辑之间的其他差异(如面大小和面形状)。为此，我们以以下方式将所有面与模型面对齐。

First, we select a neutral face image  $I_{mod}$  as the model face. Sixty eight facial landmarks of the model face  $\psi(I_{mod})$  are detected using the Active Shape Model (Cootes et al. 1995). For the  $i$ th ME clip, the 68 landmarks are detected on the first frame  $I_1$ . Then we use the Local Weighted Mean (LWM) (Goshtasby 1988) to compute a transform matrix between the landmarks of  $I_1$  and  $I_{mod}$ . The transform matrix  $T_{RAN}$  is:  $T_{RANi}$ , where  $\psi(I_i, 1)$  is the coordinates of 68 landmarks of the first frame of the ME clip  $v_i$ . All rest frames of this ME clip were normalized using matrix  $T_{RANi}$ . The normalized image  $I_0$  was computed as a 2D transformation of the original image:  $I_1$ , where  $I_1$  is the  $j$ th frame of the normalized ME clip  $v_1$ . At last, the face areas were cropped out from normalized images of each ME clip using a rectangular defined according to the eye locations in the first frame  $I_1$ . 首先，我们选择一个中性的人脸图像 $I_{mod}$ 作为模型人脸。六十八面部地标模型的脸 $\psi(I_{mod})$ 发现使用主动形状模型(傻瓜et al . 1995年)。对于第*i*个ME片段，在第一帧I<sub>1</sub>上检测到68个地标，然后使用局部加权平均(LWM) (Goshtasby 1988)计算I<sub>1</sub>和I<sub>mod</sub>地标之间的变换矩阵。变换矩阵T跑:T王妃,  $\psi(2,1)$ 的坐标68地标的第1帧我剪辑vi。该ME剪辑的所有静止帧均使用矩阵T RANi进行归一化。将归一化图像i0计算为原始图像I1的二维变换，其中I1为归一化ME clip v1的第j帧。最后，根据第一帧I1中眼睛位置定义的矩形，从每个ME剪辑的归一化图像中裁剪出人脸区域。

### 3.2.1 主动形状模型

我们从一个特定的片段中选取一个正面和中性表情的帧作为标准模板，手工定位两只眼睛的位置。然后，利用主动形状模型(Active Shape Model, ASM)检测68个面部地标<sup>[47]</sup>。利用局部加权平均(LWM)建立正则帧68个面部地标与其他帧68个面部地标之间的关系，将微表情图像与正则帧对齐，减少非表情因素的干扰<sup>[48]</sup>。

从采集的视频中选取其中没有表情且正面人脸的一帧作为模型脸 $I_{mod}$ ，应用ASM算

法标定68个人脸关键点 $\psi(\mathbf{I}_{mod})$ ，若错误标定其他物品或其他人脸（非主要人脸）时，计算任意对应位置关键点的相对距离 $L(\mathbf{I}_{mod})$ ：

$$L^{(i)}(\mathbf{I}_{mod}) = \|\psi_{\mathbf{I}_{mod}}(p \pm 68) - \psi_{\mathbf{I}_{mod}}(q \pm 68)\|_2^2 \quad (3.2)$$

其中*i*指标定的关键点组数；*p*、*q*指一个关键点组中任意两点；*n*为关键点的个数，数量为68的倍数。选择相对距离 $L^{(i)}(\mathbf{I}_{mod})$ 最大的一组关键点确定视频帧中主要人脸。如图12所示，1框内为主要人脸，2框为错误识别。

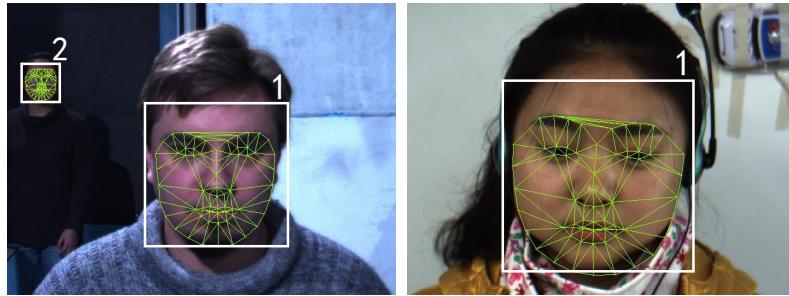


图 12 ASM算法标定的68个人脸关键点

### 3.2.2 局部加权平均算法

对每段个视频的第一帧 $\mathbf{I}_{j,1}$ 应用ASM算法标定68个人脸关键点 $\psi(\mathbf{I}_{j,1})$ ，使用LWM函数建立模型脸关键点集 $\psi(\mathbf{I}_{mod})$ 与视频第一帧关键点集 $\psi(\mathbf{I}_{j,1})$ 之间的对应关系：

$$\mathbf{T}_j = LWM(\psi(\mathbf{I}_{mod}), \psi(\mathbf{I}_{j,1})), \quad j = 1, \dots, l \quad (3.3)$$

其中*j*是视频段号，*l*是视频片段总数。对视频片段的所有帧应用该关系，使视频的每一帧具有与模型脸 $\psi(\mathbf{I}_{mod})$ 统一的姿态：

$$\mathbf{I}'_{j,k} = \mathbf{T}_j \times \mathbf{I}_{j,k}, \quad k = 1, \dots, n_j \quad (3.4)$$

其中 $\mathbf{I}_{j,k}$ 为第*j*个视频片段的第*k*帧， $\mathbf{I}'_{j,k}$ 为统一姿态后的第*j*个视频片段的第*k*帧，*n<sub>j</sub>*为*j*视频片段的帧数。如图13所示，两图均按照模型脸统一姿态，左图剔除了错误识别，右图改变了头部的倾斜角度。

### 3.2.3 时间插值模型

微表情视频有不同的长度，从4帧到50帧(如果用100帧/秒的相机拍摄)。为了解决视频片段长度不同的问题，Li等人利用TIM算法将序列的所有帧映射到曲线上，对新合成的人脸图像进行固定间隔采样，最终得到相同的预定义序列长度。实验结果表明，该算法提高了识别精度。图16显示了TIM的映射过程<sup>[46]</sup>。



图 13 LWM 算法人脸对齐后的图像

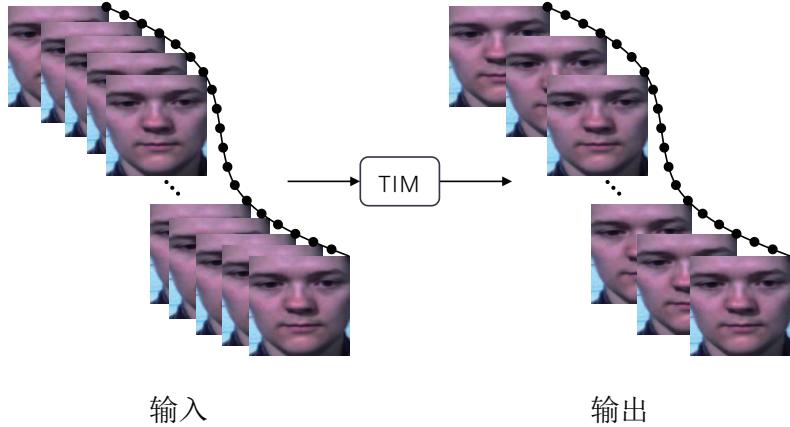


图 14 输入为原始图像序列，输出为 TIM 算法插值后图像序列

对由重建出的高分辨率图像组成的图像序列应用TIM算法统一帧数。具体操作为：将重建的图像序列映射到一条非线性曲线上 $\mathcal{F}^n : [1/n, 1] \rightarrow \mathbb{R}^{n-1}$ ，

$$\mathcal{F}^n(t) = \begin{bmatrix} f_1^n(t) \\ f_2^n(t) \\ \vdots \\ f_{n-1}^n(t) \end{bmatrix} \quad (3.5)$$

其中 $f_k^n(t) = \sin(\pi kt + \pi(n-k)/n)$ ,  $t \in [1/n, 1]$ ， $n$ 为视频段中帧数（图像序列个数），根据实验需求等间距采样，获得统一帧数。

Another special challenge of ME recognition is the short duration. For example, the shortest clip in SMIC lasts for 3/25 seconds, which has only three frames (at 25 fps). Such short sequences strictly limited the application of many spatial-temporal feature descriptors, e.g., for the LBP-TOP feature feasible radius along the time dimension can only be  $r = 1$ . Besides, there are also considerable big length variations between ME clips. This also poses a challenge for some features that are sensitive to the frame number. 另一个对自我认知的特殊挑战是持续时间短。例如，SMIC中最短的剪辑持续3/25秒，只有3帧(25 fps)。这样的短序列严格限制了许多时空特征描述符的应用，例如LBP-TOP特征沿时间维的可行半径只能为 $r = 1$ 。

除此之外，ME夹之间也有相当大的长度变化。这也对一些对帧号敏感的特性提出了挑战。

In Zhou et al. (2011), a temporal interpolation model (TIM) was proposed, which in the original paper was for the purpose of lip-reading. We employ the TIM method in our ME recognition framework to counter for the problem related with ME durations and frame number variances. Zhou等人提出了一种时间插值模型(TIM)，该模型在原论文中是用于唇读的。我们在ME识别框架中使用TIM方法来解决与ME持续时间和帧数方差相关的问题。

The TIM method relies on a path graph to characterize the structure of a sequence of frames. A sequence-specific mapping is learned to connect frames in the sequence and a curve embedded in the path graph so that the sequence can be projected onto the latter. The curve, which is a continuous and deterministic function of a single variable  $t$  in the range of [0,1], governs the temporal relations between the frames. Unseen frames occurring in the continuous process of an ME are also characterized by the curve. Therefore a sequence of frames after interpolation can be generated by controlling the variable  $t$  at different time points accordingly. TIM方法依赖于一个路径图来描述帧序列的结构。学习序列特定映射，将序列中的帧与路径图中嵌入的曲线连接起来，从而将序列投影到路径图中。曲线是[0,1]区间内单个变量t的连续确定性函数，控制着帧间的时间关系。在ME的连续过程中出现的不可见的帧也可以用曲线来表征。因此，通过控制变量t在不同时间点的变化，可以生成插值后的帧序列。

With the TIM method we are able to change the frame sequences into any arbitrary length, for either down-sampling or up-sampling. In the current framework, the TIM method was used to interpolate all ME clips (of one dataset) into one fixed length, e.g., of 10, 20, or 40 frames. By unifying the clips length, we can solve both the problem of short duration, and the problem of varied sequence lengths. The purpose of the current step is for 1) allowing more options when selecting feature parameters, and 2) achieving more stable performance with spatial-temporal feature descriptors. The problem of how to select the most suitable length for TIM interpolation is explored and discussed in the Section 2.4.2. 使用TIM方法，我们可以将帧序列改变为任意长度，无论是向下采样还是向上采样。在目前的框架中，TIM方法被用来将所有的ME剪辑(一个数据集)插入到一个固定的长度，例如，10帧、20帧或40帧。通过统一剪辑长度，既可以解决短持续时间的问题，也可以解决序列长度变化的问题。当前步骤的目的是1)在选择特征参数时允许更多的选项，2)利用时空特征描述符实现更稳定的性能。如何选择最合适的TIM插值长度的问题在2.4.2节中进行了探讨和讨论。

### 3.3 超分辨重建过程

低分辨率图像和高分辨率图像在质量和分辨率上都是不同的。高分辨率图像序列的微表情识别方法不能直接应用于低分辨率图像序列。在2.1节中，我们介绍了从高分辨率图像生成低分辨率图像的过程。

为了重建高分辨率图像，论文[45]提出了一种新的人脸幻觉算法。将基于块的正则化项与基于像素的正则化项相结合，对目标函数进行约束。重构后的高分辨率图像 $\mathbf{H}$ 可以通过最小化以下目标函数得到：

$$f(\mathbf{H}) = \|\mathbf{L} - \mathbf{DBH}\|_2^2 + \alpha \mathbf{F}_{patch} + \eta \mathbf{F}_{pixel} + \lambda \mathbf{F}_{penalty} \quad (3.6)$$

其中右侧第一项为重建误差，后三项分别是基于块的正则项、基于像素的正则项以及惩罚项。图 17 展现了重建工作的具体流程。具体将在下文详细介绍。

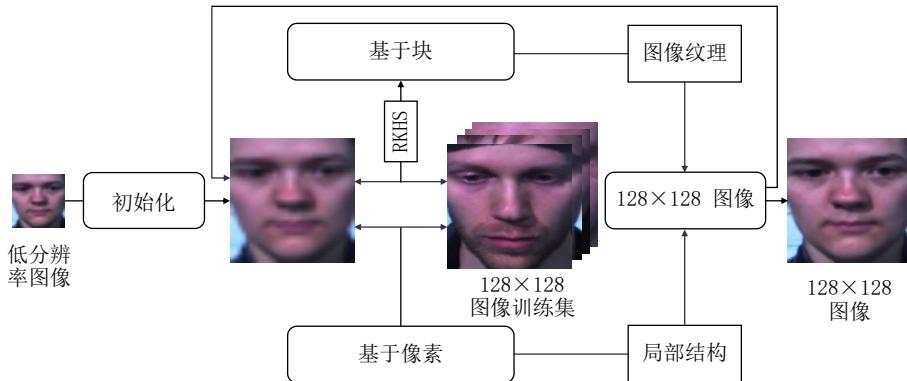


图 15 超分辨重建过程

#### 3.3.1 基于的块方法

将分割后的图像 $\mathbf{L}$ （低分辨率图像）应用三线性插值法调整为与训练样本图像 $\mathbf{H}$ （高分辨率图像，来自公开的高分辨率图像集）大小相等的尺寸，将此图命为 $\mathbf{H}^{(0)}$ （超分辨重建过程的初始图像），对图像分块处理（如分块数为 $8 \times 8$ ），如图7所示，最小化代价函数：

$$\begin{aligned} J_\tau(\omega_\tau, \mathbf{H}^{(k)}) &= \left\{ \|\phi(\mathbf{R}_\tau \mathbf{H}^k) - \phi(\mathbf{H}_\tau) \omega_\tau\|_2^2 + \lambda \left\| \mathbf{d}_\tau \bigcircledast \omega_\tau \right\|_2^2 \right\} \\ \text{s.t. } \mathbf{1}^T \omega_\tau &= 1, 2, \dots, M \end{aligned} \quad (3.7)$$

估算结合系数 $\omega_\tau$ ，将其初始结合系数命名为 $\omega_\tau^0$ ，此时 $\mathbf{H}^{(k)}$ 已知，即 $\mathbf{H}^{(0)}$ ，其中 $\tau$ 为图像中图像块的位置， $\mathbf{R}_\tau$ 为提取所有图像集位置 $\tau$ 的图像块的矩阵， $\mathbf{H}_\tau = [h_\tau^1, \dots, h_\tau^N]$ 为高分辨率图像位置 $\tau$ 处图像块集（ $N$ 为训练样本的数量，即高分辨率图像集的数量）， $\phi(\cdot)$

为从原始空间到无限维再生核希尔伯特空间 (Reproducing Kernel Hilbert Space, RKHS) 的映射,  $\mathbf{d}_p$  描述了目标高分辨率块 (重建出的块) 与相应的训练样本块在核映射空间的核相似性,  $\otimes$  表示哈达玛积,  $\lambda$  为惩罚参数,  $\mathbf{1}$  为全为1的列向量,  $M$  为图像的块数。

将获得的  $\omega_\tau$  ( $\omega_\tau^0$ ) 代入公式 (5), 并最小化:

$$\begin{aligned} f(\mathbf{H}) = & \|\mathbf{L}^k - \mathbf{D}\mathbf{B}\mathbf{H}^k\|_2^2 + \eta \sum_{\tau} \|\mathbf{x}_\tau \mathbf{h}^k - \boldsymbol{\beta}_\tau \mathbf{H}_\tau\|_2^2 + \\ & \alpha \sum_{\tau} (\|\boldsymbol{\phi}(\mathbf{R}_\tau \mathbf{H}^k) - \boldsymbol{\phi}(\mathbf{H}_\tau) \omega_\tau\|_2^2 + \lambda \|\mathbf{d}_\tau \otimes \omega_\tau\|_2^2) + \sigma MSE \end{aligned} \quad (3.8)$$

$$s.t. \quad \mathbf{1}^T \omega_\tau = 1, 2, \dots, M$$

获得重建的高分辨率图像  $\mathbf{H}^{(k)}$  (初始结果命名为  $\mathbf{H}^{(1)}$ ), 其中  $\mathbf{D}$  为下采样矩阵,  $\mathbf{B}$  为模糊处理矩阵,  $\boldsymbol{\beta}_\tau$  为规范全局优化的像素间关系矩阵,  $MSE$  为均方误差,  $\eta$  为基于像素正则项的权重,  $\alpha$  为基于块正则项的权重,  $\sigma$  为均方误差的权重。

### 3.3.2 基于像素的方法

## 3.4 微表情的特征提取与分类

如图 10 所示, 微表情识别主要分为两部分: 特征提取和分类。在以往的微表达分析方法中研究人员展示了 LBP-TOP 及其变体作为特征描述符的优势。与传统的基于单个图像的 LBP 特征不同, LBP-TOP 可以捕捉到空间和时间域的动态变化, 这对于微表情识别是必不可少的。我们首先将整个人脸图像序列划分为几个长方体, 如  $5 \times 5 \times 1$ ,  $8 \times 8 \times 2$  等, 其中前两个参数决定了空间域中的块数, 最后一个参数是时间方向上的段数。每个长方体都可以看作一个新的单位。LBP 特征提取自新单元中三个不同的正交平面 (XY、XT、YT)。我们遍历所有长方体, 得到图像序列的 LBP-TOP 特征, 然后将每个长方体的 LBP-TOP 特征串联起来。如图 5 所示。

在分类部分, 我们使用线性支持向量机 (LSVM) 作为分类器<sup>[50]</sup>。为了进行公平的比较, 我们在实验中采用了一主体退出协议。根据数据集发布方提供的微表情表签, 我们将来自 SMIC 的样本分为三类 (positive, negative, surprise), 来自 CASME II 的样本分为五类 (happiness, surprise, repression, disgust, and other)。

As mentioned in the literature review section, spatial-temporal descriptors are the major stream in most of ME analysis studies. Three kinds of spatial-temporal features are considered here in the proposed ME recognition framework. Details of each feature are briefly described below, and the comparison of their performance will be discussed with experimental results in Section 2.4.2. 如文献综述部分所述, 时空描述符是 ME 分析研究的主流。本文提出的 ME 识别框架考虑了三种时空特征。下面将简要描述每个特性的细节, 并将在 2.4.2 节

中与实验结果进行比较。

### 3.4.1 LBP-TOP特征提取

The first feature is the local binary pattern on three orthogonal planes (LBP-TOP), proposed by Zhao & Pietikäinen (2007). LBP-TOP is an extension of the original LBP for dynamic texture analysis in spatial-temporal domain. According to our literature review, LBP-TOP and its variants are the most frequently used features in current ME recognition studies. 第一个特征是三个正交平面上的局部二元模式(LBP-TOP)，由Zhao & Pietikainen(2007)提出。LBP-TOP是原LBP的扩展，用于时空域的动态纹理分析。根据我们的文献综述，LBP-TOP及其变体是目前ME识别研究中最常用的特征。

A video sequence can be thought as a cuboid of pixels on X, Y and T dimension. Traditional LBP code can be extracted from either XY, XT or YT plane, as shown in Figure 5(a). To summarize spatial-temporal attributes of the 3D cuboid, the three LBP histograms from each plane are concatenated into a big histogram as the final LBP-TOP feature vector, as illustrated in Figure 5(b). 视频序列可以看作是X、Y和T维上像素的长方体。传统的LBP代码可以从XY、XT或YT平面中提取，如图5(a)所示。为了总结三维长方体的时空属性，将每个平面的三个LBP直方图拼接成一个大直方图作为最终的LBP-top特征向量，如图5(b)所示。

### 3.4.2 LSVM

Although the selection of classifier is also important, it is not considered as the main target for the current research. To keep it well-controlled and put more focus on previous steps of the framework, in all the following ME recognition experiments, we use a linear SVM (Chang & Lin 2011) as the classifier and use the leave-one-subject-out protocol for validation. For the tests on SMIC, ME samples are classified into three categories; for the tests on CASMEII, ME samples are classified into five categories. 虽然分类器的选择也很重要，但它并不是目前研究的主要目标。为了保持良好的控制，并将更多的注意力放在框架的前几个步骤上，在接下来的ME识别实验中，我们使用了线性SVM (Chang & Lin 2011)作为分类器，使用leave-one-subject-out协议进行验证。对于中芯国际的测试，ME样本分为三类；对于CASMEII的测试，ME样本分为五类。

支持向量机(Support Vector Machine, SVM)是曾经打败神经网络的分类方法，从90年代后期开始在很多领域均有举足轻重的应用，近年来，由于深度学习的兴起，SVM的风光开始衰退，但是其仍然不失为一种经典的分类方法。SVM最初由Vladimir N. Vapnik 和 Alexey Ya. Chervonenkis于1963年提出，之后经过一系列改进，现今普遍使用的

版本由Corinna Cortes 和 Vapnik于1993年提出，并在1995年发表。深度学习兴起之前，SVM被认为是机器学习近几十年来最成功、表现最好的方法。

本文讨论线性可分的支持向量机，详细推导其最大间隔和对偶问题的原理。简单起见，以二分类为例，如下图，设训练集为 $D=(x_1,y_1), \dots, (x_n,y_n)$ ，蓝色圆点为一类，红色方块为另一类，分类的目标是寻找一个超平面，将两类数据分开。在二维平面中，分类超平面就是一条直线，从图中可以看出，能将训练样本分开的超平面有很多可能(图中绿色虚线)，超平面除了要将训练集中的数据分开，还要有较好的泛化性能，需要把测试集中的数据也划分开。从直观上看，绿色实线是比较好的一个划分，因为该直线距离两类数据点均较远，对于数据局部扰动的容忍性较好，能够以较大的置信度将数据进行分类。

### 3.5 实验设置及分析

The proposed framework was tested on two databases of SMIC and CASMEII. In order to explore the effect of each individual step of the framework, four sub-experiments were carried out each for a different purpose. The sub-experiments and their results are described in below. 该框架在中芯国际和CASMEII两个数据库上进行了测试。为了探究框架中每个单独步骤的效果，我们针对不同的目的分别进行了四个子实验。子实验及其结果如下所述。

#### Effect of TIM Interpolation TIM插值的效果

In the first sub-experiment we would like to evaluate how the interpolation process will affect the performance of the framework. We also hope to find a suitable sequence length (or length range) for the interpolation process that would be efficient for the ME recognition task. 在第一个子实验中，我们想要评估插值过程将如何影响框架的性能。我们也希望找到一个合适的序列长度(或长度范围)的插值过程，将是有效的ME识别任务。

To avoid the impact from other factors and focus on the TIM process, we skip the motion magnification step and use only LBP-TOP (with fixed parameters of 8 8 1 blocks,  $r = 2$ ,  $p = 8$ ) as the feature descriptor. We choose eight interpolation lengths (10, 20, ..., 80) for the TIM step, and evaluate the framework on SMIC-HS, SMIC-VIS and SMIC-NIR datasets. The average sequence length of the original ME clips is 33.7 frames for SMIC-HS and 9.66 frames for SMIC-VIS and SMIC-NIR. 原始ME片段的平均序列长度为SMIC-HS为33.7帧，SMIC-VIS为9.66帧，SMIC-NIR为9.66帧。为了避免其他因素的影响，将重点放在TIM过程上，我们跳过了运动放大的步骤，只使用LBP-TOP(参数固定为881个block,  $r = 2$ ,  $p = 8$ )作为特征描述符。我们选择8个插值长度(10,20, ...)对于TIM步骤，对SMIC-HS、SMIC-

VIS和SMIC-NIR数据集的框架进行评估。

Test results are shown in Figure 6. The results can be summarized in two aspects. First, interpolation to 10 frames (TIM10 for short) leads to significantly better performance than without the TIM process. Compared to the original sequences, TIM10 barely changed the average sequence lengths for SMIC-VIS and SMIC-NIR, and it was a down-sampling process for SMIC-HS. Thus we think the improved performance was caused by the unifying of the sequences length. Secondly, if we compare the result of TIM10 to those of longer TIM sequences, it shows that longer interpolated sequences do not lead to better performance. One possible explanation for this result might be that, the time-dimension changes are diluted if the ME clips are interpolated into much longer sequences. According to the two findings, it appears that TIM of 10 frames is the best option for the current framework. In all the following experiments, TIM10 is applied in the framework as default if not otherwise specified. 测试结果如图6所示。研究结果可以归纳为两个方面。首先，插值到10帧(简称TIM10)比不使用TIM过程的performance要好得多。与原始序列相比，TIM10几乎没有改变SMIC-VIS和SMIC-NIR的平均序列长度，是SMIC-HS的下采样过程。因此，我们认为序列长度的统一导致了性能的提高。其次，如果我们将TIM10的结果与较长的TIM序列的结果进行比较，就会发现较长的插值序列并不会带来更好的性能。对这个结果的一种可能的解释是，如果ME剪辑被插值成更长的序列，时间维度的变化就会被稀释。根据这两个发现，对于目前的框架来说，TIM of 10 frames似乎是最好的选择。在接下来的所有实验中，如果没有另外指定，TIM10将作为默认值应用于框架中。

#### Comparison of features 特性的比较

The purpose of the second sub-experiment is to compare the performance of three features. Five combinations of histograms on three orthogonal planes of each feature are evaluated separately. 第二个子实验的目的是比较三个特征的性能。在每个特征的三个正交平面上分别计算五种直方图组合。

After face alignment, TIM10 were applied to interpolate all sequences into 10 frames. The motion magnification step was temporally skipped for later discussion. Three kinds of features were extracted from evenly divided blocks of sequences with varied parameters. For the LBP feature, we vary the radius  $r$ , neighbour points  $p$  and the number of divided blocks; for the HOG and HIGO features, we fixed the number of bins as  $b = 8$  and vary the number of divided blocks. Tests were carried out on three datasets of SMIC and CASMEII, and the results are listed in Table 5. Note that results of the five combinations of three orthogonal planes of each

feature are listed separately. For each combination of feature (one cell in the table), only the best result (with corresponding parameters) achieved among all parameter combinations is listed. 人脸对齐后，应用TIM10对所有序列进行10帧插值。在后面的讨论中暂时跳过了运动放大步骤。从不同参数序列的均匀分割块中提取三种特征。对于LBP特征，我们改变半径r，邻点p和划分块数；对于HOG和HIGO特性，我们将桶的数量固定为 $b = 8$ ，并改变划分块的数量。对SMIC和CASMEII的三个数据集进行了测试，结果如表5所示。注意，每个特征的三个正交平面的五个组合的结果被单独列出。对于每个特征组合(表中的一个单元格)，只列出所有参数组合中获得的最佳结果(具有相应的参数)。

Two phenomena can be found from the result table. First, the TOP combination (three orthogonal planes) doesn't always lead to the best performance, especially for the HIGO feature. In many cases better results can be achieved using only XOT, YOT or XYOT plan features, and it is true for all four datasets. On the other hand the XY plane feature always get the lowest performance than other plane combinations. The results indicate that dynamic changes along T dimension carry the most important information for ME recognition, while the XY plane features carries more about facial appearance information which maybe redundant for the ME recognition task. Similar findings were also reported in Davison et al. (2014). Secondly, comparing the three kinds of features, gradient-based features HOG and HIGO outperform LBP on three out of the four test datasets (except SMIC-NIR). HIGO seems to perform slightly better than HOG, and the highest performance obtained on SMIC is 76.06% using HIGO-XOT. One possible explanation is that the HIGO feature is not affected by local gradient magnitude, which might vary due to the diversity of muscle movement speeds among ME clips. Results on the NIR data shows another trend. Skin textures recorded by an NIR camera are different from that of visible color videos. For the SMIC-NIR dataset, the LBP feature performed better than the other two features, which is consistent with previous results in Zhao et al. (2011). 从结果表中可以发现两个现象。首先，最上面的组合(三个正交平面)并不总是带来最好的性能，特别是对于HIGO特性。在许多情况下，仅使用XOT、YOT或XYOT计划特性就可以获得更好的结果，对于所有四个数据集都是如此。另一方面，XY平面特征总是比其他平面组合得到最低的性能。结果表明，T维上的动态变化是ME识别中最重要的信息，而XY平面特征所携带的面部特征信息更多，这可能是ME识别任务中多余的信息。Davison et al.(2014)也报道了类似发现。其次，比较这三种特征，基于梯度的特征HOG和HIGO在四个测试数据集中的三个(SMIC-NIR除外)上优于LBP。HIGO的性能似乎略优于HOG，在SMIC上使用HIGO-xot获得的最高性能为76.06%。一种可能的解释是HIGO特征不受局部梯度大小的影响，

局部梯度大小可能由于ME片段中肌肉运动速度的多样性而变化。近红外数据显示了另一种趋势。近红外相机记录的皮肤纹理不同于可视彩色视频。对于SMIC-NIR数据集，LBP特征优于其他两个特征，这与Zhao等(2011)之前的研究结果一致。

我们现在在三个不同的自发微表达数据集上展示实验和结果，即SMIC-HS, SMIC-subHS和CASME II。实验参数的设置和结果分析将在下面的小节中讨论。

表4 实验中使用的数据集

	SMIC-HS	SMIC-subHS	CASME II
微表情数	164	71	247
参与者	16	8	26
分类	3	3	5

SMIC-HS和SMIC-subHS是SMIC的两个子集。SMIC-HS数据集包含来自16名参与者的164个自发微表情片段，分为三类:阳性(51个片段)、阴性(70个片段)和惊奇(43个片段)。SMIC-subHS数据集是SMIC-HS的子集，只包含最后8个参与者。前8名受试者的微表达片段数量差异较大，其中3名受试者贡献了整个组近一半的微表达样本，这可能会影晌“一受试者退出”的表现，而后8名受试者的s (SMIC-subHS)片段数量分布较为均匀。SMIC-subHS数据集中，正负、惊喜片段分别为28、23、20个。同时，CASME II数据集包含26名参与者，分别属于5个不同的类别:惊讶(25个片段)、幸福(32个片段)、其他(99个片段)、厌恶(64个片段)和压抑(27个片段)。表1显示了实验中使用的数据集的摘要。面部高分辨率图像的分辨率设置为 $128 \times 128$ 的实验。 $128 \times 128$ 分辨率的图像下采样乘以2,4,8次获得低分辨率图像(如图6)。这意味着我们评估三个不同层次的低分辨率的面部图像序列(如 $16 \times 16$ ,  $32 \times 32$ ,  $64 \times 64$ )的微识别任务。

在本节中，利用论文[45]提出的方法将低分辨率图像重建为高分辨率图像，该方法在2.3节中进行了简要介绍。表2-3列出了不同分辨率下重建图像序列的平均峰值信噪比(PSNR)和结构相似度(SSIM)指数。在这里，我们分别使用S64、S32和S16来命名分辨率为 $64 \times 64$ ,  $32 \times 32$ 和 $16 \times 16$ 的重建图像序列。

表5 重建图像序列的平均PSNR(dB)指标

PSNR (dB)	$16 \times 16$	$32 \times 32$	$64 \times 64$
SMIC-HS	31.25	37.67	44.30
SMIC-subHS	31.67	38.26	43.22
CASME II	31.80	36.49	37.83

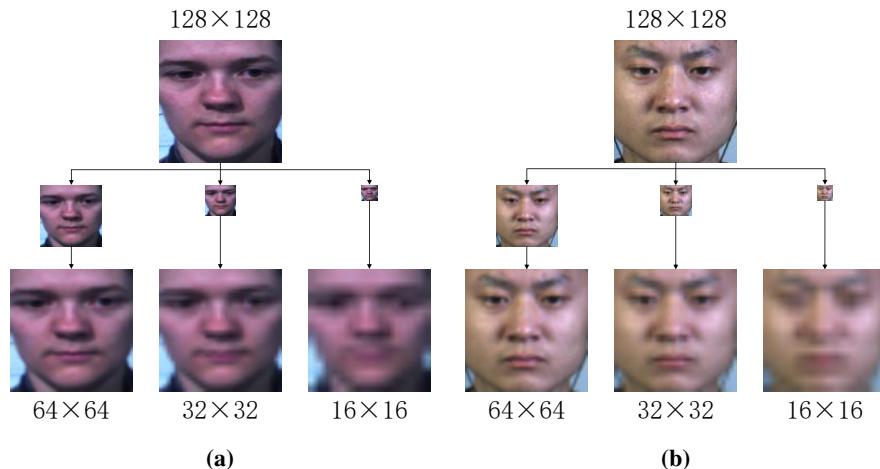


图 16 低分辨率图像

(a) SMIC-HS/SMIC-subHS数据集低分辨率图像, (b) CASME II数据集低分辨率图像

表 6 重建图像序列的平均SSIM指标

SSIM	$16 \times 16$	$32 \times 32$	$64 \times 64$
SMIC-HS	0.9397	0.9775	0.9883
SMIC-subHS	0.8970	0.9346	0.9424
CASME II	0.9439	0.9761	0.9882

如表2和表3所示，重建的人脸图像序列的定量指标(PSNR/SSIM)与输入人脸图像序列的分辨率成正比。例如，在SMIC-HS数据集中，S16的PSNR指数为31.25dB，比S32低6.42dB，比S64低13.05dB。对于SSIM指数，S16达到0.9397，比S32低0.0378，比S64低0.0486。此外，图7给出了重建后的图像序列的视觉表现，也表明了与上述观点相同的结论。

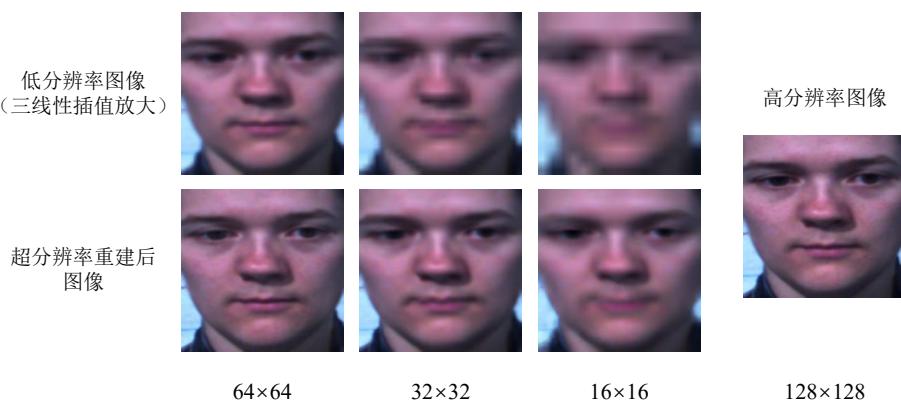


图 17 不同分辨率图像重建结果比较

为了对视频片段的时长进行归一化，使用TIM算法将视频片段的帧数插值为10帧，如2.2节所述。我们应用快速LBP-TOP将视频剪辑成不同的长方体和提取每个长方体的LBP-TOP特征构成一个完整的功能，使用统一的映射，半径设置为 $r = 2$ ，和相邻点 $p$ 的

数量设置为 $p = 8$ 。我们使用leave-one-subject-out协议进行实验，即，将一个受试者的所有样本作为测试集，其他受试者的所有样本作为训练集。我们采用LSVM作为分类器，其中惩罚系数 $c = 1$ 。

在本节中，我们提出了低分辨率图像序列的微表情识别性能的基线。为了适应不同分辨率的测试样本，我们将训练集从 $128 \times 128$ 下采样到对应的分辨率(即)，以便进行分类程序。注意，下采样操作导致微表达式缺乏判别特征。接下来的实验也表明，在非常低的分辨率下，识别的准确率会急剧下降。

图8显示了不同分辨率图像序列在不同数据集上的识别精度。在这里，我们分别使用L64、L32和L16来命名低分辨率图像序列。从图8可以看出，当输入图像序列的分辨率从 $64 \times 64$ 降低到 $32 \times 32$ 时，SMIC-SubHS数据集(蓝色折线)的识别准确率显著降低。同时，我们可以看到低分辨率图像序列(如L16)的准确率相对较低。这一现象表明，低分辨率的图像序列很难获得满意的结果。主要原因是因为低分辨率导致微表情描述缺乏高频信息和纹理细节。

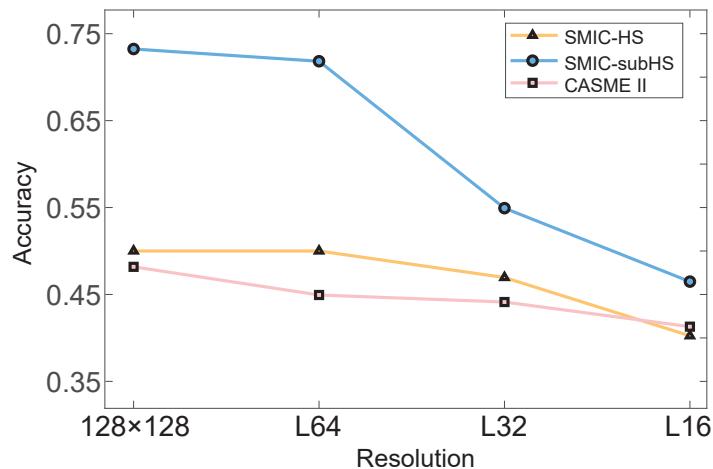


图 18 不同数据集不同分辨率的图像序列的识别准确度

图9为低分辨率图像序列分类结果的混淆矩阵，并以 $128 \times 128$ 分辨率下的性能为参考。我们可以发现，在SMICSubHS数据集(图9第二列)中，当图像序列的分辨率为 $128 \times 128$ 时，混淆矩阵更加集中在对角线上，说明微表情识别方法的识别效果较好。然而，当图像序列的分辨率降低时，混淆矩阵逐渐变差。我们还可以发现，在SMICHs(图9第一列)和CASME II(图9第三列)中，误分类的比例大于SMICsubHS，图8所示的识别准确率也相对较低。对于SMIC-HS来说，上述问题的主要原因可能是后8个受试者(SMICsubHS数据集)的分布比前8个受试者更加均衡。对于CASME II来说，主要是由于分布不平衡和类别过多。例如，在CASME II数据集中，来自类other的视频片段数量占40.08%。

			R	D	O	H	S	
			22.22	11.11	44.44	22.22	0.00	
			1.56	34.38	57.81	6.25	0.00	
			1.01	18.18	71.72	9.09	0.00	
			3.13	18.75	18.75	53.13	6.25	
			0.00	28.00	36.00	24.00	12.00	
N	34.29	41.43	24.29					
P	23.53	60.78	15.69					
S	27.91	9.30	62.79					
N	73.91	21.74	4.35					
P	21.43	78.57	0.00					
S	15.00	20.00	65.00					
N	N	P	S					
N	34.29	31.43	34.29					
P	17.65	70.59	11.76					
S	27.91	20.93	51.16					
N	52.17	30.43	17.39					
P	10.71	85.71	3.57					
S	10.00	15.00	75.00					
N	N	P	S					
N	27.14	28.57	44.29					
P	15.69	60.78	23.53					
S	27.91	9.30	62.79					
N	39.13	47.83	13.04					
P	25.00	71.43	3.57					
S	30.00	20.00	50.00					
N	N	P	S					
N	38.57	20.00	41.43					
P	41.18	23.53	35.29					
S	18.60	18.60	62.79					
N	26.09	47.83	26.09					
P	21.43	71.43	7.14					
S	35.00	30.00	35.00					
N	N	P	S					
R	7.41	40.74	48.15	3.70	0.00			
D	4.69	26.56	59.38	6.25	3.13			
O	0.00	14.14	78.79	6.06	1.01			
H	0.00	28.13	59.38	12.50	0.00			
S	0.00	28.00	56.00	12.00	4.00			
R	R	D	O	H	S			

图 19 不同分辨率图像序列的识别准确度混淆矩阵

水平方向，从左往右依次是SMIC-HS数据集、SMIC-subHS数据集和CASME II数据集。垂直方向，从上到下依次是 $128 \times 128$ 分辨率混淆矩阵、 $64 \times 64$ 分辨率混淆矩阵、 $32 \times 32$ 分辨率混淆矩阵和 $16 \times 16$ 。其中P代表积极（positive）、N代表消极（negative）、S代表惊喜（surprise）、R代表蔑视（repression）、D代表沮丧（disgust）、O代表其他（others）、H代表开心（happiness）。

在本小节中，我们将在三个数据集上对我们提出的框架进行实验。首先将测试集重构为 $128 \times 128$ 的分辨率，然后在高分辨率空间中进行分类。通过实证分析，选择了最优参数。表4给出了识别精度以及相应的块大小参数设置。从表4可以看出，实验结果有了显著的改善。例如在SMIC-subHS数据集中， $64 \times 64$ 分辨率下的图像序列识别准确率从71.83%提高到74.65%，提高了2.82%。S32的识别准确率为74.65%，比L32高19.72%。S16的识别准确率为73.24%，比L16高26.76%。这表明该方法对低分辨率图像序列的微表情识别精度有较好的提高。我们还注意到，与直接将输入作为原始的128128个图像序列相比，该框架使用S64得到了更好的结果。这可能是因为原始SMIC-HS/subHS数据集中的样本在记录过程中存在明显的噪声，使得人脸序列实际上包含了冗余和噪声信息。

超分辨率重建图像序列识别精度的混淆矩阵如图 19 所示。我们展示了 $128 \times 128$ 、S64、S32 和 S16 图像序列根据不同数据集的识别精度。从图 19 可以看出，我们提出的框架的混淆矩阵比图 9 更集中在对角线上。特别是在 $16 \times 16$  SMIC-HS 数据集(左下)，积极的识别精度有显著提高。此外，我们可以从SMIC-subHS 数据集(第二列)的结果中看出，将阴性误分类为阳性的比例显著降低，而将阳性正确分类的比例也得到了大幅提高。不幸的是，尽管CASME II 数据集(第三列)的结果有所改善，但每个类别的分类仍然很差，通常错误地划分为其他类别。也许是因为其他微表情包含了所有其他类型的微表情，不包括惊讶、快乐、厌恶和压抑，所以它的分类是混合的。综上所述，从图 19 和图 9 的对比中我们可以看出，该框架对于低分辨率的微表情识别具有很好的性能提升。

表 7 不同分辨率图像在不同数据集上的识别精度比较

Accuracy (%)	High-resolution	Super-resolution Reconstruction			Low-resolution		
	$128 \times 128$	$64 \times 64$	$32 \times 32$	$16 \times 16$	$64 \times 64$	$32 \times 32$	$16 \times 16$
SMIC-HS	50.00 ( $8 \times 8 \times 2$ )	52.44 ( $5 \times 5 \times 2$ )	51.83 ( $5 \times 5 \times 2$ )	51.83 ( $5 \times 5 \times 2$ )	50.00 ( $6 \times 6 \times 5$ )	46.95 ( $6 \times 6 \times 5$ )	40.24 ( $3 \times 3 \times 6$ )
	73.24 ( $5 \times 5 \times 2$ )	74.65 ( $5 \times 5 \times 2$ )	74.65 ( $5 \times 5 \times 2$ )	73.24 ( $8 \times 8 \times 3$ )	71.83 ( $6 \times 6 \times 2$ )	54.93 ( $6 \times 6 \times 2$ )	46.48 ( $4 \times 4 \times 1$ )
SMIC-subHS	48.18 ( $7 \times 7 \times 3$ )	48.18 ( $7 \times 7 \times 5$ )	44.53 ( $7 \times 7 \times 3$ )	42.92 ( $7 \times 7 \times 5$ )	44.94 ( $7 \times 7 \times 1$ )	44.13 ( $4 \times 4 \times 2$ )	41.30 ( $2 \times 2 \times 5$ )
CASME II							

High-resolution是指分辨率为 $128 \times 128$ 的图像序列，Super-resolution Reconstruction是指将低分辨率图像序列通过超分辨率重建方法为 $128 \times 128$ ，Low-resolution是指重建前的低分辨率图像序列。 $X \times Y \times T$ 指水平、垂直和时间方向块的数量。

### 3.6 总结

This Chapter focuses on the studies of ME analysis. We first reviewed the state-of-the-art progress about ME studies, including psychological studies about the concept and phenomenon of the ME, the challenges of collecting ME data, earlier studies using posed MEs, and also more recent studies of automatic ME spotting and recognition using spontaneous ME datasets. Then four parts of works about spontaneous ME analysis done by us were introduced, including 1) the collection of the first spontaneous ME database SMIC; 2) a framework for ME recognition;

3) an ME spotting method using feature difference analysis; and 4) an automatic ME analysis system (MESR) for firstly spotting and then the recognition of MEs. 本章着重于ME分析的研究。我们首先回顾了ME研究的最新进展，包括关于ME的概念和现象的心理学研究、收集ME数据的挑战、早期使用所提出的MEs的研究，以及最近使用自发ME数据集进行ME自动定位和识别的研究。然后介绍了我们所做的关于自发ME分析的四部分工作，包括1)收集了第一个自发ME数据库SMIC;2)自我认知的框架;3)基于特征差异分析的ME定位方法;4)自动ME分析系统(MESR)，用于先定位后识别MEs。

The topic of ME analysis concerns facial movements at a very fine level. It is at an early stage by now, but it is attracting more attentions and developing rapidly. New databases are emerging, and new methods have been proposed with better performance both for ME recognition and for ME spotting. In future computers might be able to sense people's hidden feelings better than us with the ability to accurately spot and recognize MEs. We were the first few researchers that devoted to this topic. The main contributions of our exploratory works are 1) to break the ground and attract more researchers to the topic of ME; 2) to provide data as a benchmark for future studies; and 3) to propose basic framework and possible solutions for countering ME-specific challenges (e.g., the short duration and the intensity of movements) that may inspire future work. I plan to continue research on ME analysis in future, and detailed plans is described in Section 4. ME分析的主题是非常精细的面部运动。虽然目前还处于起步阶段，但已引起越来越多的关注，发展迅速。新的数据库正在出现，新的方法已经被提出，它们在ME识别和ME定位方面都有更好的性能。在未来，计算机可能比我们更好地感知人们隐藏的情感，并能准确地发现和识别MEs。我们是第一批致力于这一课题的研究人员。我们探索性工作的主要贡献是:1)开拓创新，吸引更多的研究者关注我的课题;2)提供数据作为未来研究的基准;3)提出基本框架和可能的解决方案，以应对可能激励未来工作的ME-specific挑战(如短期和运动强度)。我计划在未来继续对ME分析进行研究，具体计划见第4节。

本文对低分辨率微表情识别问题进行了全面的研究。我们使用模糊和下采样模型来生成和模拟低分辨率的微表情人脸图像序列。我们在每一帧上使用面部幻觉的方法重建高质量的面部图像序列，增强局部细节，将低质量的图像序列放大到高分辨率的图像序列。然后利用快速LBP-TOP提取动态特征，利用SVM分类器对微表情进行识别。实验结果表明，在低分辨率的微表情识别问题上，该框架在可公开获取的微表情数据集(SMIC-HS、SMIC-subHS、CASME II)上表现良好。未来，我们将重点研究低分辨率情况下微表情识别的深度特征。

## 第四章 基于深度学习方法的低分辨率环境下微表情识别

我们考虑在不受控制的环境中对动作的全自动识别。大多数现有的工作依赖领域知识从输入构建复杂的手工特性。此外，通常假定环境是受控的。卷积神经网络(tional neural network, CNNs)是一种深度模型，它可以直接作用于原始输入，从而实现特征构造过程的自动化。然而，这些模型目前仅限于处理2D输入。在本文中，我们开发了一种新的三维CNN动作识别模型。该模型通过三维卷积从空间和时间两个维度提取特征，从而获取多个相邻帧中编码的运动信息。所建立的模型从输入帧中生成多个通道的信息，并结合所有通道的信息得到最终的特征表示。我们将开发的模型应用到现实环境中的人类行为识别中，在不依赖于手工制作的特性的情况下取得了优异的性能。

Micro-expression is one of important clues for detecting lies. Its most outstanding characteristics include short duration and low intensity of movement. Therefore, video clips of high spatial-temporal resolution are much more desired than still images to provide sufficient details. On the other hand, owing to the difficulties to collect and encode micro-expression data, it is small sample size. In this paper, we use only 560 micro-expression video clips to evaluate the proposed network model: Transferring Long-term Convolutional Neural Network (TLCNN). TLCNN uses Deep CNN to extract features from each frame of micro-expression video clips, then feeds them to Long Short Term Memory (LSTM) which learn the temporal sequence information of micro-expression. Due to the small sample size of micro-expression data, TLCNN uses two steps of transfer learning: (1) transferring from expression data and (2) transferring from single frame of micro-expression video clips, which can be regarded as “big data”. Evaluation on 560 micro-expression video clips collected from three spontaneous databases is performed. The results show that the proposed TLCNN is better than some state-of-the-art algorithms. 微表情是检测谎言的重要线索之一。它最突出的特点是运动时间短，强度低。因此，高时空分辨率的视频片段比静止图像更需要提供足够的细节。另一方面，由于微表达数据的采集和编码困难，样本量较小。在本文中，我们仅使用560个微表情视频片段来评价所提出的网络模型:transfer long tional Neural network (TLCNN)。TLCNN利用深度CNN从微表情视频片段的每一帧中提取特征，然后将其输入到长短时记忆(Long Short-Term Memory, LSTM)中，LSTM学习微表情的时序信息。由于微表情数据的样本容量较小，TLCNN采用了两个迁移学习的步骤:(1)从表情数据进行迁移，(2)从单帧微表情视频片段进行迁移，可视为“大数据”。对采集自三个自发性数据库的560个微表

情视频片段进行评价。结果表明，所提出的TLCNN算法优于现有的一些算法。

Recently, owing to the rapid development of computer hardware, especially Graphical Processor Unit (GPU), deep learning is applied on many areas such as face recognition [35] and verification [36], and shows outstanding performances. These deep learning methods use multiple processing layers to discover patterns and structures in very large data sets. Each layer learns a concept from the data that subsequent layers build on; the higher the level, the more abstract the concepts that are learned. Deep learning does not depend on prior data processing and automatically extracts features [37]. These advantages and good performances of deep learning are ascribed to big data. However, the number of micro-expression video clips is usually small. Deep learning on data with small sample size may not achieve good performances. To address this problem, we use transfer learning to pretrain a deep convolutional neural network and we propose the Transferring Long-term Convolutional Neural Network (TLCNN) model for micro-expression recognition. In TLCNN, there are two steps of transfer learning: (1) transferring from expression data and (2) transferring from single frame of micro-expression video clips, which can be regarded as big data . To fully use the temporal information in micro-expression videos, TLCNN also uses Long Short Term Memory (LSTM) to extract temporal features of micro-expression videos from mid-level image representation for each frame images.

近年来，随着计算机硬件特别是图形处理器单元(GPU)的飞速发展，深度学习在人脸识别[35]、验证[36]等领域得到了广泛的应用，表现出了优异的性能。这些深度学习方法使用多个处理层来发现非常大的数据集中的模式和结构。每一层从后续层构建的数据中学习一个概念；层次越高，所学的概念就越抽象。深度学习不依赖于先验数据处理，自动提取特征[37]。这些优势和深度学习的良好表现都归功于大数据。然而，微表情视频剪辑的数量通常很少。在样本容量较小的数据上进行深度学习可能不会取得良好的效果。针对这一问题，我们利用转移学习对深度卷积神经网络进行预处理，提出了用于微表情识别的转移长期卷积神经网络(TLCNN)模型。在TLCNN中，迁移学习有两个步骤：(1)从表达数据进行迁移，(2)从单个微表达视频片段帧进行迁移，这可以看作是大数据。为了充分利用微表情视频中的时间信息，TLCNN还利用长短时记忆(Long Short Term Memory, LSTM)从每帧图像的中层图像表示中提取微表情视频的时间特征。

## 4.1 数据集预处理

**数据增强(Data Augmentation):** 是指对图片进行随机的旋转、翻转、裁剪、随机设置图片的亮度和对比度以及对数据进行标准化(数据的均值为0，方差为1)。通过这些操

作，我们可以获得更多的图片样本，原来的一张图片可以变为多张图片，扩大了样本容量，对于提高模型的准确率和提升模型的泛化能力非常有帮助，在进行数据增强的同时也会需要消耗大量的系统资源。

#### 4.1.1 数据集混合

#### 4.1.2 直方图均衡化

直方图均衡化(Histogram Equalization)又称直方图平坦化,实质上是对图像进行非线性拉伸,重新分配图像象元值,使一定灰度范围内象元值的数量大致相等。这样,原来直方图中间的峰顶部分对比度得到增强,而两侧的谷底部分对比度降低,输出图像的直方图是一个较平的分段直方图:如果输出数据分段值较小的话,会产生粗略分类的视觉效果。

直方图是表示数字图像中每一灰度出现频率的统计关系。直方图能给出图像灰度范围、每个灰度的频度和灰度的分布、整幅图像的平均明暗和对比度等概貌性描述。灰度直方图是灰度级的函数,反映的是图像中具有该灰度级像素的个数,其横坐标是灰度级 $r$ ,纵坐标是该灰度级出现的频率(即像素的个数) $p_r(r)$ ,整个坐标系描述的是图像灰度级的分布情况,由此可以看出图像的灰度分布特性,即若大部分像素集中在低灰度区域,图像呈现暗的特性;若像素集中在高灰度区域,图像呈现亮的特性。

图 20 所示就是直方图均衡化,即将随机分布的图像直方图修改成均匀分布的直方图。基本思想是对原始图像的像素灰度做某种映射变换,使变换后图像灰度的概率密度呈均匀分布。这就意味着图像灰度的动态范围得到了增加,提高了图像的对比度。

图, 来自直方图均衡化

图 20 直方图均衡化

通过这种技术可以清晰地在直方图上看到图像亮度的分布情况,并可按照需要对图像亮度调整。另外,这种方法是可逆的,如果已知均衡化函数,就可以恢复原始直方图。

设变量 $r$ 代表图像中像素灰度级。对灰度级进行归一化处理,则 $0 \leq r \leq 1$ ,其中 $r = 0$ 表示黑,  $r = 1$ 表示白。对于一幅给定的图像来说,每个像素值在 $[0, 1]$ 的灰度级是随机的。用概率密度函数 $p_r(r)$ 来表示图像灰度级的分布。

为了有利于数字图像处理,引入离散形式。在离散形式下,用 $r^k$ 代表离散灰度级,用 $p_r(r^k)$ 代表 $p_r(r)$ ,并且下式成立:  $p_r(r^k) = nk/n$ , 其中,  $0 \leq r^k \leq 1$ ,  $k = 0, 1, 2, \dots, n-1$ 。式中 $n^k$ 为图像中出现 $r^k$ 这种灰度的像素数,  $n$ 是图像中的像素总数, 而 $nk/n$ 就是概率论中的频数。图像进行直方图均衡化的函数表达式为:

$$S_i = T(r_i) = \sum_{i=0}^{k-1} \frac{n_i}{n} \quad (4.1)$$

式中,  $k$ 为灰度级数。相应的反变换为:

$$r^i = T^{-1}(S_i) \quad (4.2)$$

#### 4.1.3 亮度调整

### 4.2 特征提取及识别

#### 4.2.1 P3D 块

Inspired by the recent successes of Residual Networks (ResNet) [7] in numerous challenging image recognition tasks, we develop a new family of building modules named Pseudo-3D (P3D) blocks to replace 2D Residual Units in ResNet, pursuing spatio-temporal encoding in ResNet-like architectures for videos. Next, we will recall the basic design of Residual Units in ResNet, followed by presenting how to devise our P3D blocks. The bottleneck building architecture on each P3D block is finally elaborated. 受最近残差网络(ResNet)[7]在众多具有挑战性的图像识别任务中取得的成功启发, 我们开发了一个名为伪3d (P3D)块的构建模块家族, 以取代ResNet中的2D残差单元, 在视频的ResNet-like架构中实现时空编码。接下来, 我们将回顾ResNet中剩余单元的基本设计, 然后介绍如何设计P3D块。最后阐述了每个P3D块上的瓶颈构建体系结构。

Residual Units 残差单位 ResNet consists of many stacked Residual Units and each Residual Unit could be generally given by ResNet由许多桩状残余单元组成, 每个残余单元通常由eq (1)

$$\mathbf{x}_{t+1} = \mathbf{h}(\mathbf{x}_t) + \mathbf{F}(\mathbf{x}_t) \quad (4.3)$$

where  $\mathbf{x}_t$  and  $\mathbf{x}_{t+1}$  denote the input and output of the  $t$ -th Residual Unit,  $\mathbf{h}(\mathbf{x}_t) = \mathbf{x}_t$  is an identity mapping and  $\mathbf{F}$  is a non-linear residual function. Hence, Eq.(1) can be rewritten as

其中 $\mathbf{x}_t$ 和 $\mathbf{x}_{t+1}$ 表示第 $t$ 个残差单元的输入和输出,  $\mathbf{h}(\mathbf{x}_t) = \mathbf{x}_t$ 为恒等映射,  $\mathbf{F}$ 为非线性残差函数。因此, 式(1)可以改写为eq(2) where  $\mathbf{F} \cdot \mathbf{x}_t$  represents the result of performing residual function  $\mathbf{F}$  over  $\mathbf{x}_t$ . The main idea of ResNet is to learn the additive residual function  $\mathbf{F}$

with reference to the unit inputs  $\mathbf{x}_t$  which is realized through a shortcut connection, instead of directly learning unreferenced non-linear functions.

$$(\mathbf{I} + \mathbf{F}) \cdot \mathbf{x}_t = \mathbf{x}_t + \mathbf{F} \cdot \mathbf{x}_t := \mathbf{x}_t + \mathbf{F}(\mathbf{x}_t) = \mathbf{x}_{t+1} \quad (4.4)$$

其中  $\mathbf{F} \cdot \mathbf{x}_t$  表示在  $\mathbf{x}_t$  上执行残差函数  $\mathbf{F}$  的结果。ResNet 的主要思想是通过一个快捷的连接，而不是直接学习未引用的非线性函数，来学习参考单元输入  $\mathbf{x}_t$  的可加性残差函数  $\mathbf{F}$ 。

### P3D Blocks design P3D块设计

To develop each 2D Residual Unit in ResNet into 3D architectures for encoding spatiotemporal video information, we modify the basic Residual Unit in ResNet following the principle of Pseudo 3D as introduced in Section 3.1 and devise several Pseudo-3D Blocks. The modification is not straightforward for involvement of two design issues. The first issue is about whether the modules of 2D filters on spatial dimension (S) and 1D filters on temporal domain (T) should directly or indirectly influence each other. Direct influence within the two types of filters means that the output of spatial 2D filters is connected as the input to the temporal 1D filters (i.e., in a cascaded manner). Indirect influence between the two filters decouples the connection such that each kind of filters is on a different path of the network (i.e., in a parallel fashion). The second issue is whether the two kinds of filters should both directly influence the final output. As such, direct influence in this context denotes that the output of each type of filters should be directly connected to the final output. 为了将ResNet中的每个2D残差单元发展成用于编码时空视频信息的3D架构，我们按照3.1节中介绍的伪3D原理对ResNet中的基本残差单元进行了修改，并设计了几个伪3D块。由于涉及两个设计问题，所以修改并不简单。第一个问题是空间维度上的二维滤波器模块和时间域上的一维滤波器模块是否应该直接或间接地相互影响。两种滤波器之间的直接影响是将空间二维滤波器的输出连接为时间一维滤波器的输入(即，以级联的方式)。两个过滤器之间的间接影响使连接解耦，使每一种过滤器在网络的不同路径上(即)。第二个问题是这两种过滤器是否都应该直接影响最终的输出。因此，在此上下文中，直接影响表示每种过滤器的输出都应该直接连接到最终输出。

Based on the two design issues, we derive three different P3D blocks as depicted in Figure 2, respectively, named as P3D-A to P3D-C. Detailed comparisons about their architectures are provided as following: 基于这两个设计问题，我们推导出如图2所示的三个不同的P3D块，分别命名为P3D- a到P3D- c。关于它们的架构的详细比较如下：

(1) P3D-A: The first design considers stacked architecture by making temporal 1D filters (T) follow spatial 2D filters (S) in a cascaded manner. Hence, the two kinds of filters can directly

influence each other in the same path and only the temporal 1D filters are directly connected to the final output, which could be generally given by eq (3) (1) P3D-A:第一种设计考虑了层叠结构，使时间一维滤波器(T)以级联方式跟随空间二维滤波器(S)。因此，两种滤波器在同一路径上可以直接相互影响，只有时间1D滤波器直接连接到最终输出，一般可以给出eq(3)

(2) P3D-B: The second design is similar to the first one except that indirect influence between two filters are adopted and both filters are at different pathways in a parallel fashion. Although there is no direct influence between S and T, both of them are directly accumulated into the final output, which could be expressed as eq (4) (2) P3D-B:第二种设计与第一种设计相似，只是采用了两个滤波器之间的间接影响，两个滤波器以并行的方式在不同的路径上。虽然S和T之间没有直接的影响，但是它们都直接累加到最终输出中，可以表示为eq (4)

(3) P3D-C: The last design is a compromise between P3D-A and P3D-B, by simultaneously building the direct influences among S, T and the final output. Specifically, to enable the direct connection between S and final output based on the cascaded P3D-A architecture, we establish a shortcut connection from S to the final output, making the output  $xt+1$  as eq (5) (3) P3D-C:最后的设计是在P3D-A和P3D-B之间进行折衷，同时建立S、T和最终输出之间的直接影响。具体来说，为了基于级联P3D-A架构实现S与最终输出的直接连接，我们建立了S到最终输出的快捷连接，使输出 $xt+1$ 为eq(5)

### Bottleneck architectures 瓶颈的架构

When specifying the architecture of 2D Residual Unit, the basic 2D block is modified with a bottleneck design for reducing the computation complexity. In particular, as shown in Figure 3(a), instead of a single spatial 2D filters (3 3 convolutions), the Residual Unit adopts a stack of 3 layers including 1 1, 3 3, and 1 1 convolutions, where the first and last 1 1 convolutional layers are applied for reducing and restoring dimensions of input sample, respectively. Such bottleneck design makes the middle 3 3 convolutions as a bottleneck with smaller input and output dimensions. Thus, we follow this elegant recipe and utilize the bottleneck design to implement our proposed P3D blocks. Similar in spirit, for each P3D block which purely consists of one spatial 2D filters (1 3 3 convolutions) and one temporal 1D filters (3 1 1 convolutions), we additionally place two 1 1 1 convolutions at both ends of the path, which are responsible for reducing and then increasing the dimensions. Accordingly, the dimensions of the input and output of both the spatial 2D and temporal 1D filters are reduced with this

bottleneck design. The detailed bottleneck building architectures on all the three P3D blocks are illustrated in Figure 3(b) to 3(d). 在确定二维残差单元的体系结构时，对二维基本块进行瓶颈设计，以降低计算复杂度。特别是，如图3所示(一个)，而不是一个单一的空间2 d过滤器(3 3旋转)，剩余单位采用一堆3层包括1 1 3 3 1 1曲线玲珑，第一个和最后一个1 1卷积层在哪里申请减少和恢复方面的输入样本，分别。这样的瓶颈设计使得中间的3 3个卷积成为输入和输出维数较小的瓶颈。因此，我们遵循这个优雅的配方，并利用瓶颈设计来实现我们提出的P3D块。相似的精神，为每个P3D块纯粹由一个空间二维滤波器卷积(1 3 3)和一个时间1 d过滤器(3 1 1旋转)，我们另外两个1 1 1两端的卷积路径，负责降低，然后增加尺寸。因此，该瓶颈设计减少了空间二维和时间一维滤波器的输入和输出的维数。图3(b) - 3(d)展示了这三个P3D块上详细的瓶颈构建体系结构。

#### 4.2.2 P3D ResNet

In order to verify the merit of the three P3D blocks, we first develop three P3D ResNet variants, i.e., P3D-A ResNet, P3D-B ResNet and P3D-C ResNet by replacing all the Residual Units in a 50-layer ResNet (ResNet-50) [7] with one certain kind of P3D block, respectively. The comparisons of performance and time efficiency between the basic ResNet-50 and the three P3D ResNet variants are presented. Then, a complete version of P3D ResNet is proposed by mixing all the three P3D blocks from the viewpoint of structural diversity. 为了验证这三个P3D模块的优点，我们首先开发了三个P3D ResNet变体，即，P3D- a ResNet、P3D- b ResNet和P3D- c ResNet分别用一种P3D块替换50层ResNet (ResNet-50)[7]中的所有残余单元。比较了基本ResNet-50和三种P3D ResNet变体的性能和时间效率。然后，从结构多样性的角度，将三个P3D块混合在一起，提出了一个完整的P3D ResNet版本。

Mixing different P3D Blocks. Further inspired from the recent success of pursuing structural diversity in the design of very deep networks [38], we devise a complete version of P3D ResNet by mixing different P3D blocks in the architecture to enhance structural diversity, as depicted in Figure 4. Particularly, we replace Residual Units with a chain of our P3D blocks in the order P3D-A->P3D-B->P3D-C. Table 1 also details the performance and speed of the complete P3D ResNet. By additionally pursuing structural diversity, P3D ResNet makes the absolute improvement over P3D-A ResNet, P3D-B ResNet and P3D-C ResNet by 0.5%, 1.4% and 1.2% in accuracy respectively, indicating that enhancing structural diversity with going deep could improve the power of neural networks. 混合不同的P3D块。从最近在深度网络[38]的设计中追求结构多样性的成功中得到进一步的启发，我们设计了一个完整的P3D ResNet版本，通过在架构中混合不同的P3D块来增强结构多样性，如图4所示。特别地，我们

用P3D- a ->P3D- b ->P3D- c顺序的P3D区块链替换剩余的单元。表1还详细说明了完整P3D ResNet的性能和速度。通过进一步追求结构多样性， P3D ResNet相对于P3D- a ResNet、 P3D- b ResNet和P3D- c ResNet的准确率分别提高了0.5%、 1.4%和1.2%， 说明随着深度的增加， 结构多样性的增强可以提高神经网络的能力。

在视频分类或理解领域， 容易从图像领域的2D卷积联想到用3D卷积来做， 虽然用3D卷积进行特征提取可以同时考虑到spatial和temporal维度的特征， 但是计算成本和模型存储都太大， 因此这篇文章针对视频领域中采用的3D卷积进行改造， 提出Pseudo-3D Residual Net (P3D ResNet)， 思想有点像当年的Inception v3中用1\*3和3\*1的卷积叠加代替原来的3\*3卷积， 这篇文章是用1\*3\*3卷积和3\*1\*1卷积代替3\*3\*3卷积（前者用来获取spatial维度的特征， 实际上和2D的卷积没什么差别； 后者用来获取temporal维度的特征， 因为倒数第三维是帧的数量）， 毕竟这样做可以大大减少计算量， 而如果采用3D卷积来做的话， 速度和存储正是瓶颈， 这也使得像C3D算法的网络深度只有11层， 参看Figure1。该文章的网络结构可以直接在3D的ResNet网络上修改得到。顺便提一下， 除了采用3D卷积来提取temporal特征外， 还可以采用LSTM来提取， 这也是当前视频研究的一个方向。

Figure1是几个模型在层数、 模型大小和在Sports-1M数据集上的视频分类效果对比， 其中的P3D ResNet是在ResNet 152基础上修改得到的， 深度之所以不是152， 是因为改造后的每个residual结构不是原来ResNet系列的3个卷积层， 而是3或4个卷积层， 详细可以看Figure3， 所以最后网络深度是199层。官方github代码中的网络就是199层的。ResNet 152是直接在Sports-1M数据集上fine tune得到的。可以看出199层的P3D ResNet虽然在模型大小上比ResNet-152（此处ResNet-152是在sports-1M数据集上fine tune得到的）大一些， 但是准确率提升比较明显， 与C3D（此处C3D是直接在sports-1M数据集上从头开始训练得到的）的对比在效果和模型大小上都有较大改进， 除此之外， 速度的提升也是亮点， 后面有详细的速度对比。

#### 4.2.3 2

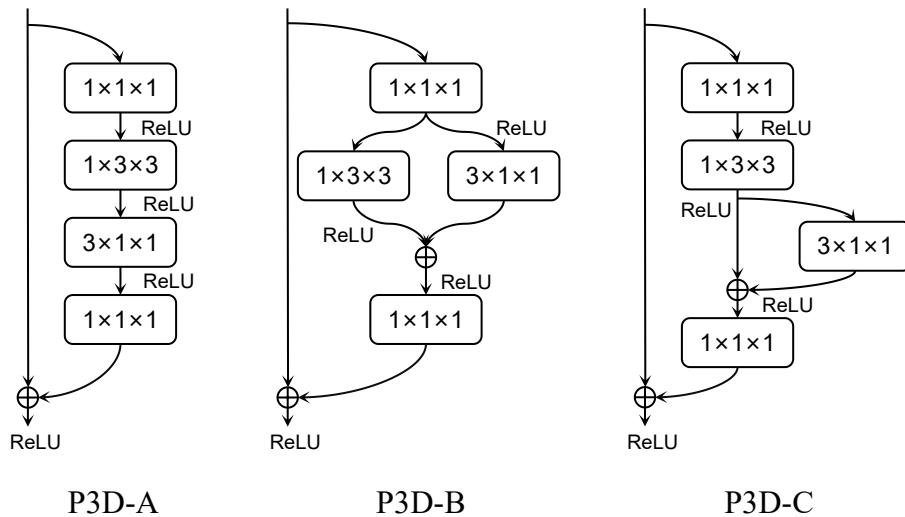


图 21 伪3D残差网络三种设计模型

### 4.3 实验设置及分析

4.3.1 1

4.3.2 2

4.3.3 3

### 4.4 总结



## 第五章 系统设计

5.1 需求分析

5.2 功能设计

5.2.1 功能图

5.2.2 时序图

5.2.3 等

5.3 界面设计

5.4 小节



## 第六章 总结与展望

The current ME studies can be continued and improved from four aspects in future work. First, about the ME database: more spontaneous ME data are still needed in order to develop more sophisticated computational models. Compared to ordinary FE databases, the size of current ME databases are not big enough. Future collection of ME data can be improved from three ways: the first is to increase the sample size; the second is to involve AU labelling; the third is to include depth information to build 3D ME models. A large 3D ME database is now under construction with collaboration of a group of UK researchers. 在今后的工作中，可以从四个方面继续和完善目前ME的研究。首先，关于ME数据库：为了开发更复杂的计算模型，仍然需要更多自发的ME数据。与普通FE数据库相比，目前ME数据库的规模还不够大。未来ME数据的收集可以通过三种方式进行改进：一是增加样本量；二是涉及AU标签；第三个是包含深度信息来构建3D ME模型。一组英国研究人员正在合作建立一个大型3D ME数据库。

Second, about ME spotting: the framework using feature difference analysis for ME spotting described in Section 2.5 was the first method proposed for spotting MEs from spontaneous long videos. One challenge of the current spotting framework is that there are other brief but non-emotional movements (e.g., eye blinks) that need to be ruled out from MEs. In future, more refined spotting method will be developed on the AU level, so that non-emotional brief movements can be ruled out to reduce the false positive rate. Besides, future ME spotting method will also try to target at providing more precise temporal information of the ME including the onset, apex and offset frames. 第二，关于ME点测：2.5节中描述的ME点测特征差异分析框架是第一个从自发长视频中提取MEs的方法。当前的识别框架的一个挑战是，需要排除MEs中其他短暂但非情绪的动作(例如眨眼)。未来将在AU水平上开发更精细的点样方法，排除非情绪短暂运动，降低假阳性率。此外，未来的ME定位方法也将致力于提供更精确的ME的时间信息，包括起始帧、顶点帧和偏移帧。

Third, about ME recognition: the latest method proposed in paper III showed advantage over previous methods by employing one extra step to magnify the subtle motions. Other video processing methods will be explored and added to the framework, if they are demonstrated to be helpful for the ME recognition task. More sophisticated machine learning models will be studied including deep learning models. It is also planned to use 3D information for ME recognition

when the new 3D ME database is finished. 第三，关于ME识别:第三篇论文中提出的最新方法比之前的方法有优势，多了一步放大了细微的运动。如果其他视频处理方法被证明对ME识别任务有帮助，我们将探索并添加到该框架中。将研究更复杂的机器学习模型，包括深度学习模型。也计划在新的3D ME数据库完成后使用3D信息进行ME识别。

Fourth, about integrated ME spotting and recognition systems: after progresses are made for both ME recognition methods and ME spotting methods, it is also planned to build advanced integrated systems for more accurate ME spotting and recognition. 第四，ME点测与识别集成系统:在ME点测与识别方法取得进展后，计划构建先进的ME点测与识别集成系统，提高ME点测与识别的准确率。

## 6.1 总结

## 6.2 存在的问题与展望

### 6.2.1 存在的问题

### 6.2.2 展望

## 附录 A 附录

学位论文是研究生科研工作成果的集中体现，是评判学位申请者学术水平、授予其学位的主要依据，是科研领域重要的文献资料。根据《科学技术报告、学位论文和学术论文的编写格式》(GB/T 7713-1987)、《学位论文编写规则》(GB/T 7713.1-2006)和《文后参考文献著录规则》(GB7714—87)等国家有关标准，结合中国科学院大学（以下简称“国科大”）的实际情况，特制订本规定。

### A.1 论文无附录者无需附录部分

### A.2 测试公式编号

$$\begin{cases} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0 \text{ times font test} \\ \frac{\partial(\rho \mathbf{V})}{\partial t} + \nabla \cdot (\rho \mathbf{V} \mathbf{V}) = \nabla \cdot \boldsymbol{\sigma} \text{ times font test} \\ \frac{\partial(\rho E)}{\partial t} + \nabla \cdot (\rho E \mathbf{V}) = \nabla \cdot (k \nabla T) + \nabla \cdot (\boldsymbol{\sigma} \cdot \mathbf{V}) \end{cases} \quad (A.1)$$

$$\frac{\partial}{\partial t} \int_{\Omega} u \, d\Omega + \int_S \mathbf{n} \cdot (u \mathbf{V}) \, dS = \dot{\phi} \quad (A.2)$$

### A.3 测试生僻字

霜蟾盟薇曜灵霜飈妙鬘虛霏凌澌蕘枯菡萏汎寥宵冥瑤濩落雪雪便嬛岩峣灑澣姽婳  
婳愔嫕飒纏梦俪縕冤孽甲摛藻卮言倥偬椒觞期颐夜阑彬蔚倥偬澄廓簪缨陟遐迤逦缥缃  
鶡蝶憮懔闺闼璀璨错媖婀嚧咤湧洞闔闊覩缕玓瓅逡巡譏譏琭琭濂濂跼跼叆叇氤氲瓠犀流  
眄蹀躞瀼瀼暭暭瓔珞螭首衡皋琳愴缱绻皵皵顰頷愀然菡萏卑陬纯懿犇羣崕崕  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇墇墇  
墇墇墇墇墇墇墇  
墇墇墇墇墇  
墇墇墇  
墇墇  
墇  
墇



## 参考文献

- [1] Haggard E A, Isaacs K S. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy[M/OL]. Springer US, 1966: 154-165. [https://doi.org/10.1007/978-1-4684-6045-2\\_14](https://doi.org/10.1007/978-1-4684-6045-2_14).
- [2] Ekman P, Friesen W V. Nonverbal leakage and clues to deception[J]. Psychiatry, 1969, 32(1):88-106.
- [3] Mehrabian A, Ferris S R. Inference of attitudes from nonverbal communication in two channels[J]. Journal of Consulting Psychology, 1967, 31(3):248.
- [4] Kazlev M. The triune brain[J/OL]. KHEPER, 1999, 5(19)[2003-11-19]. <http://www.kheper.net/topics/intelligence/MacLean.htm>.
- [5] Chiu M H, Chou C C, Wu W L, et al. The role of facial microexpression state (fmes) change in the process of conceptual conflict[J]. British Journal of Educational Technology, 2014, 45(3):471-486.
- [6] Yan W J, Wu Q, Liang J, et al. How fast are the leaked facial expressions: The duration of micro-expressions[J]. Journal of Nonverbal Behavior, 2013, 37(4):217-230.
- [7] Matsumoto D, Hwang H S. Evidence for training the ability to read microexpressions of emotion[J]. Motivation and Emotion, 2011, 35(2):181-191.
- [8] Porter S, Brinke L T. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions[J]. Psychological Science, 2008, 19(5):508-514.
- [9] 周凯莉. 别对我撒谎, 我懂“微表情” [N/OL]. 中国青年报, 2010, 11(10). [http://zqb.cyol.com/content/2010-11/10/content\\_3441411.htm](http://zqb.cyol.com/content/2010-11/10/content_3441411.htm).
- [10] Ekman, Paul, Sullivan O, et al. A few can catch a liar[J]. Psychological Science, 1999, 10(3):263-266.
- [11] Li X, Hong X, Moilanen A, et al. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods[J]. IEEE Transactions on Affective Computing, 2017, PP(99):1-1.
- [12] Li X. Reading subtle information from human faces[D/OL]. Doctoral Dissertation: Acta Universitatis Ouluensis. C, Technica, 2017[2017-09-08]. <http://urn.fi/urn:isbn:9789526216386>.
- [13] Ekman P, Friesen W V. Detecting deception from the body or face[J]. Journal of Personality & Social Psychology, 1974, 29(3):288-298.
- [14] Wikibook. <http://en.wikibooks.org/wiki/Latex>[M]. On-line Resources, 2014.
- [15] Leslie L. Document preparation system[M]. Addison-Wesley Reading, MA, 1986.
- [16] 初景利. 图书馆数字参考咨询服务研究[M]. 北京: 北京图书馆出版社, 2004.
- [17] 陈浩元. 著录文后参考文献的规则及注意事项[J]. 编辑学报, 2005, 17(6):413-415.
- [18] Walls S C, Barichivich W J, Brown M. Drought, deluge and declines: the impact of precipitation extremes on amphibians in a changing climate[J/OL]. Biology, 2013, 2(1):399-418[2013-11-04]. <http://www.mdpi.com/2079-7737/2/1/399>. DOI: 10.3390/biology2010399.
- [19] 陈晋镳, 张惠民, 朱士兴, 等. 蓟县震旦亚界研究[M]//中国地质科学院天津地质矿产研究所. 中国震旦亚界. 天津: 天津科学技术出版社, 1980: 56-114.

- [20] 袁训来, 陈哲, 肖书海. 蓝田生物群: 一个认识多细胞生物起源和早期演化的新窗口 – 篇三[J]. 科学通报, 2012, 57(34):3219.
- [21] Heinrich S, Deyan G, Catalin D, et al. MathML-aware article conversion from LaTeX[J]. Towards a Digital Mathematics Library, 2009, 16(2):109-120.
- [22] 哈里森·沃尔德伦. 经济数学与金融数学[M]. 谢远涛, 译. 北京: 中国人民大学出版社, 2012: 235-236.
- [23] 牛志明, 斯温兰德, 雷光春. 综合湿地管理国际研讨会论文集[C]. 北京: 海洋出版社, 2013.
- [24] ボハンデ. 過去及び現在に於ける英國と会[J]. 日本時報, 1928, 17:5-9.
- [25] Ду бровина. И. Открытое письмо Председателя Гла вного Совета Союза Русского Народа Свят-Петрбургскому Антонию, Первенствующему Члену Священного Синода [J]. Вече, 1906:1-3.
- [26] Pfister T, Li X, Zhao G, et al. Recognising spontaneous facial micro-expressions[C]//Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.
- [27] Ruiz-Hernandez J A, Pietikäinen M. Encoding local binary patterns using the re-parametrization of the second order gaussian jet[C]//Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013.
- [28] Wang S J, Yan W J, Zhao G, et al. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features[C]//Workshop at the European conference on computer vision. Springer, 2014.
- [29] Wang Y, See J, Phan R C W, et al. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition[C]//Asian Conference on Computer Vision. Springer, 2014.
- [30] Wang Y, See J, Phan R C W, et al. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition[C]//PloS one. Public Library of Science, 2015.
- [31] Wang S J, Yan W J, Li X, et al. Micro-expression recognition using dynamic textures on tensor independent color space[C]//2014 22nd International Conference on Pattern Recognition (ICPR). IEEE, 2014.
- [32] Le Ngo A C, Phan R C W, See J. Spontaneous subtle expression recognition: Imbalanced databases and solutions[C]//Asian conference on computer vision. Springer, 2014.
- [33] Liang S T, See J, Phan R C W, et al. Subtle expression recognition using optical strain weighted features [C]//Asian Conference on Computer Vision. Springer, 2014: 644-657.
- [34] Oh Y H, Le Ngo A C, See J, et al. Monogenic riesz wavelet representation for micro-expression recognition[C]//Digital Signal Processing (DSP), 2015 IEEE International Conference on. IEEE, 2015.
- [35] Xiaohua H, Wang S J, Liu X, et al. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition[C]//IEEE Transactions on Affective Computing. IEEE, 2017.
- [36] Liu Y J, Zhang J K, Yan W J, et al. A main directional mean optical flow feature for spontaneous micro-expression recognition[C]//IEEE Transactions on Affective Computing. IEEE, 2016.

- 
- [37] Lu Z, Luo Z, Zheng H, et al. A delaunay-based temporal coding model for micro-expression recognition [C]//Asian conference on computer vision. Springer, 2014.
  - [38] Xu F, Zhang J, Wang J Z. [C].
  - [39] He J, Hu J F, Lu X, et al. Multi-task mid-level feature learning for micro-expression recognition[C]//Pattern Recognition. Elsevier, 2017.
  - [40] Patel D, Hong X, Zhao G. Selective deep features for micro-expression recognition[C]//Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016.
  - [41] Lei Z, Ahonen T, Pietikäinen M, et al. Local frequency descriptor for low-resolution face recognition [C]//Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. IEEE, 2011.
  - [42] Wang Z, Miao Z, Wu Q J, et al. Low-resolution face recognition: a review[C]//The Visual Computer. Springer, 2014.
  - [43] Li X, Pfister T, Huang X, et al. A spontaneous micro-expression database: Inducement, collection and baseline[C]//Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on. IEEE, 2013.
  - [44] Yan W J, Li X, Wang S J, et al. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation[C]//PloS one. Public Library of Science, 2014.
  - [45] Shi J, Liu X, Zong Y, et al. Hallucinating face image by regularization models in high-resolution feature space[C]//IEEE Transactions on Image Processing. IEEE, 2018.
  - [46] Zhou Z, Zhao G, Pietikäinen M. Towards a practical lipreading system[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.
  - [47] Cootes T F, Taylor C J, Cooper D H, et al. Active shape models-their training and application[C]//Computer vision and image understanding. Elsevier, 1995.
  - [48] Goshtasby A. Image registration by local approximation methods[C]//Image and Vision Computing. Elsevier, 1988.
  - [49] Zhao G, Pietikäinen M. Dynamic texture recognition using local binary patterns with an application to facial expressions[C]//IEEE transactions on pattern analysis and machine intelligence. IEEE, 2007.
  - [50] Chang C C, Lin C J. Libsvm: a library for support vector machines[C]//ACM transactions on intelligent systems and technology (TIST). ACM, 2011.
  - [51] Hong X, Xu Y, Zhao G. Lbp-top: a tensor unfolding revisit[J]. 2016.



## 致谢

### 1.基本情况

吴凌云，福建省屏南县人，中国科学院数学与系统科学研究院博士研究生。

### 2.教育背景

2008.08~2012.07西北大学，本科，专业：

2012.09~西北大学，硕士研究生，专业：

### 3.攻读硕士学位期间的其它奖励

可以随意添加新的条目或是结构。



## 攻读硕士学位期间取得的科研成果

### 1.发表学术论文

[1] ucasthesis: A LaTeX Thesis Template for the University of Chinese Academy of Sciences, 2014.

### 2.申请（授权）专利

(无专利时此项不必列出)

### 3.参与科研项目及科研获奖

可以随意添加新的条目或是结构。

