摘要

智能化是视频监控的必然趋势，而在智能视频监控当中，人脸检测与识别又是非常重要而又具有挑战性的课题。人脸包含了非常重要的生物信息，其提供了一种可靠、方便的身份鉴别途径，有着巨大的实用价值和广阔的应用前景。然而，在监控视频当中，场景复杂多变、光照不受控制、图像质量低，这给人脸检测和识别带来了很大的挑战。本文在大量研究者工作的基础上，主要考虑监控场景下多姿态人脸的检测与识别问题。

针对监控环境下人脸形变、尺度、摄像头角度等问题，本文采用了一种基于级联可变形部件模型（DPM）的人脸检测方法。可变形部件模型有效地解决了因人脸姿态变化导致的漏检和因场景差异带来的误检，而级联机制有效的加快了检测速度。实验结果表明，该方法不但提高了对多人脸多个姿态的检测率，而且降低了人脸检测的误检率。并经过多线程和指令集优化后，在i5 CPU的PC上，对720p监控视频取得每秒25帧的速率。

针对人脸检测后的归一化问题，需要对多姿态的人脸进行特征点检测，也称人脸对齐。针对特征点问题当中的形变、遮挡及实时性的要求，本文提出了一种两级级联回归的人脸特征点定位方法，并通过自适应初始化的方式提高了定位的速度。方法采用了两级形状回归、局部坐标系下的形状索引特征和基于相关性的特征选择机制，实现了由粗到细的特征点定位。实验表明，该方法能够对多姿态的人脸实现精准的定位，且本文提出的自适应初始化的策略使得回归过程快速收敛，实现了人脸特征点的实时定位。

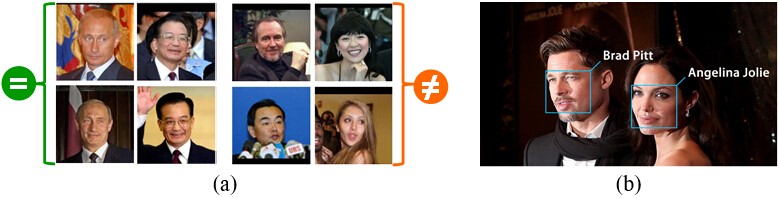
针对监控环境下人脸识别中人脸确认和辨识两个任务，本文提出了两级多任务度量学习方法用于人脸辨识，并通过简化的单任务度量学习方法应用于人脸确认任务。相比人脸确认这个二分类问题，人脸辨识问题可以被看作一个多分类问题。传统度量学习方法忽视了多分类问题当中共享的类间相似性，如：不同人的人脸有很强的视觉相似性。本文通过等价约束的方式学习一个度量矩阵。在第一级当中，采用了多种特征。在第二级当中，将每个人当作一个特定的子任务，而每个任务的标签信息代表着身份信息，最终通过多任务投票机制进行分类。该方法充分利用了成对的标签信息，并将该信息迁移到多分类任务当中。实验结果表明，该方法比传统多任务学习方法和多任务支持向量机有着更好的效果。

关键词：人脸检测，人脸对齐，人脸辨识，度量学习，可变形部件模型

# 第一章 绪论

## **1.1** 研究背景与意义

马克思说：“如果事物的表现形式和事物的本质会直接合而为一，一切科学就成为多余的了”。科学研究本质上就是对未知对象进行观测，而后根据事物之间的联系揭示背后隐藏的本质。处在“信息爆炸”的今天，大量的信息资源丰富了人类活动的同时也带来了前所未有的挑战：海量数量，有价值的信息隐藏在大量冗余的数据中，很难直接获取，多数据源，不同特性的采集设备带来的不同形式信息很难进行比较和处理。如何有效分析和利用这些作为人类感知外界信息主要来源的视觉信息显得尤为重要。分析和处理高维图像数据正是机器学习、模式识别、计算机视觉、大数据等领域一个主要研究方向和任务。面向智能视频监控环境下的人脸识别技术就是基于监控视频数据进行分析处理的过程。



#### 图 1.1 人脸确认（1:1）与人脸辨识 (1:N)

Fig 1.1 Faec Verification (1:1) vs. Face Identification (1:N)

人脸识别技术主要分为两类：人脸确认（FaceVerification）和人脸辨识（FaceIdentification）。人脸确认是判断两张已知的人脸图像是不是同一个人的技术。人脸辨识是指鉴别某张人脸图像身份的技术，即确定一张给定人脸图片的身份类别标签。如图1.1所示，其中图1.1. (a)表示人脸确认，即判断两张图片是否是同一个人，即1:1的任务，本质上可以看成一个二分类问题。人脸确认技术可被用于登陆验证、身份识别等应用场景。帮助用户快速判定两张照片是否为同一个人、判定视频中的人脸是否为目标人脸并支持实时识别认证，还可以实现身份和人脸绑定等功能。而图1.1. (b)表示人脸辨识任务，即要判断其身份类别，其难度随着人数规模而增加，即1:N的任务，其本质上是一个多分类问题。通过人脸辨识可以实现安防检查、VIP识别、照片自动圈人、人脸登陆、出入控制等多种功能。研究人脸识别在理论技术和行业应用方面都有重要意义。

人脸识别是人工智能和模式识别的一个前沿研究领域。人脸识别的研究已经有几十年的历史，早在60年代初期就有学者开始人脸识别问题的研究。随着计算机技术和模式识别技术的发展在80年代左右不断得到提高，经过很多学者的不懈努力，与人脸识别相关的技术开始逐渐作为一个独立的研究方向得到了学者们的广泛关注，如人脸检测、人脸特征点检测、人脸属性（性别、年龄、种族、微笑等）的识别等。所以广义的人脸识别包括了上述及部分内容（本文人脸识别特指人脸确认和人脸辨识）。

在2001年，人脸识别相关的算法得到了突破性的进展，以级联Haar Adaboost 算法[1]为代表的人脸检测和以ASM/AAM[2] 为代表的人脸对齐方法开始走向实用，与人脸识别相关的技术也从实验室限制条件下向非限制条件下的人脸识别问题转移。而现在深度学习的方法使得人脸识别有了更大的突破，使得人脸识别技术逐步从理论研究走向更多的实际应用[3,4]。总结来说，人脸识别是计算机视觉、模式识别、人工智能等前沿领域的重要部分，人脸识别技术的不断进步展现了弱人工智能向强人工智能的转化，具有重要的学术研究价值。

人脸识别是新一代智能视频监控领域的一项重要技术。传统的视频监控系统通常是通过人员监控和录像来实现安全防护，由于显示屏数量有限、人工费时费力，对安全隐患无法实时监控和预警。另一方面，由于监控点过多，人员监控根本无法顾及所有监控场景。随着数字化进程的推进，设备智能化程度、系统复杂化程度越来越高、环节流程、设备数量也越来越多，传统的视频监控存在明显的不足，必将被数字化的智能视频监控所替代[5]。人脸识别作为监控领域的一项核心技术，可以有效解决目前数字监控系统中存在的某些难题，如：确定监控场景中是否有人、对监控对象跟踪、查找、确定监控对象身份、流量监控等。然而由于人脸自身的某些局限，在监控视频当中的人脸受光照强度、姿态变化、表情、饰物遮挡等影响，加上摄像头本身角度及分辨率等问题、使得监控场景下的人脸识别问题变得非常具有挑战性。因此，监控环境下的人脸识别是一项核心技术，而人脸识别和监控技术的有效结合，在安防领域具有十分广泛的应用前景。

人脸识别技术在机器人视觉、人机交互领域有着重要应用。机器人视觉(Robot Vision)是机器人视觉感知的重要组成部分。机器人技术目前发展十分迅速，有很多产品已经进入到了人的实际生活当中，而机器人与人的自然交互是人机友好共处的关键技术之一，也是人们追求的目标。人脸具有丰富的信息，利用人脸识别技术可以有效的提高机器人与人的友好交互能力，机器人不但可以通过人脸识别来进行身份的辨认，而且还可以对表情、年龄、性别、种族、情感等更加细化的人脸属性进行分析。其中，法国Aldebaran Robotics公司的NAO机器人和日本本田公司研制的ASIMO[6]机器人是人型机器人的两款经典产品。随着人脸识别技术的不断发展以及人机交互（Human Robot Interaction, HRI）的深入研究，必将使机器人领域不断有新的发展和突破，实现人与机器人更加自然的交互。

因此，人脸识别在人机交互[7]、智能机器人[8]、安防监控[3,4,9]等领域都具有重要的理论意义和实际应用价值。

## **1.2** 研究现状与趋势

### **1.2.1** 前沿难点与关键问题

本文主要关注智能监控环境下的人脸识别问题，而对智能监控系统的架构和工程问题不做细致的探讨。所以围绕人脸识别问题的前沿难点和关键问题，本文将从以下几个方面进行论述：

人脸识别研究内容。在智能监控系统中，对人脸问题的研究不仅仅是对视频当中的人脸进行确认和辨识。智能监控环境下对人脸识别的研究是一个广义的概念，经过众多研究者的努力，人脸识别技术在研究及应用方面得到了长足的发展，特别是最近几年来取得了突飞猛进的进步。总结来讲，人脸识别的研究范围大致可以分为如下几个方面的内容：

人脸检测(Face Detection)：即通过给定的图像或视频确定其人脸大小和位置。由于受光照、表情、人脸姿态、遮挡、人脸大小变化等因素的影响，使得人脸检测问题变得非常复杂。而更加精细的人脸检测需要确定人脸的朝向及姿态。一般来讲，都是通过滑动窗口的方式将人脸检测看成一个二分类问题。

人脸对齐(Face Alignment)：即从已经检测到的人脸进行特征点定位。因此，人脸对齐又称之为人脸特征点检测（Face Landmark Detection）或人脸特征点定位（Face Landmark Localization）。人脸特征点检测最重要的应用是将一张旋转、形变、未归一化的人脸变换成一张正脸图片，以便更好的进行人脸识别。除此之外，人脸特征点检测也能广泛应用于人脸替换、人脸渐变、虚拟化妆、表情分析等领域。

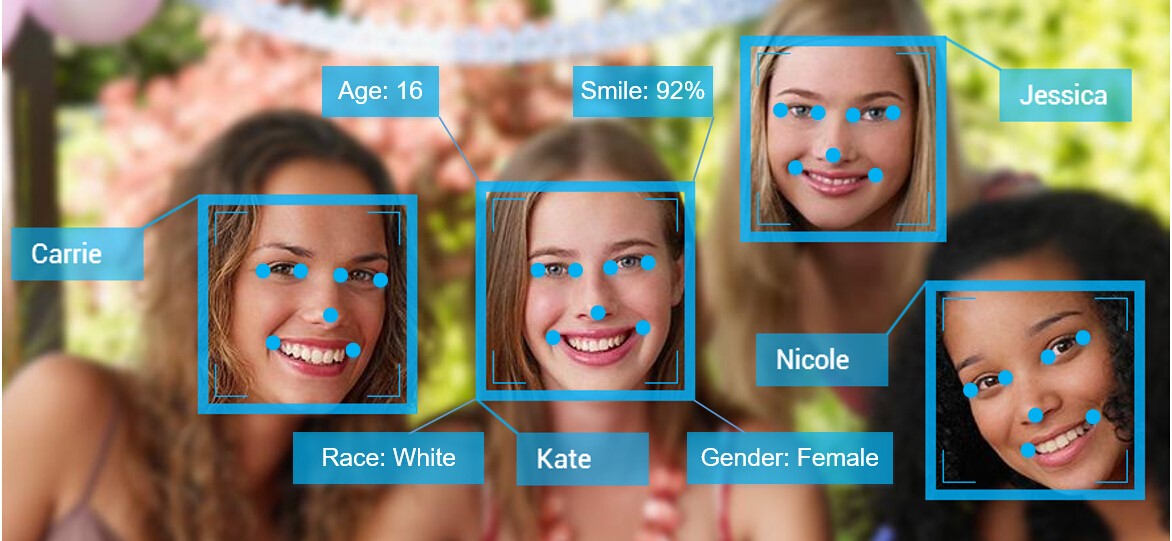
人脸表征(Face Representation)：即对人脸进行更有区分力的表达。传统方法采用了手工设计的描述子进行特征提取，然后再将其变换到一个低维的判别空间，如LBP、

SIFT、HOG等。而深度学习与其最本质的区别在于直接从原始数据当中学习到更加完备的人脸表征。

人脸识别(Face Recognition)：人脸识别包括两个任务。分别是人脸确认（Face Verification）和人脸辨识（Face Identification）。人脸确认任务即判断给定的两张人脸图片是否是同一个人，是一个1：1的任务，本质上是一个二分类问题。而人脸辨识是将待识别的人脸与数据库中的已知人脸进行比较匹配，判断其身份的任务，是1：N的任务，本质上是一个多分类问题。这两个过程的核心是选择适当的人脸的表征方式与相似度度量策略。

人脸属性(Face Attribute)：即对人脸的基于面部的属性进行识别与分析。常见的面部属性有：性别、年龄、种族、微笑分析、是否带眼镜、是否化妆等。对微笑的分析又包括对微笑的检测、微笑强度估计、真假笑识别这三个方面。而对表情的分析又包括常见的对六种人脸表情的分类与微表情的分类。显然，完成这一操作需要大量的知识并且通常是非常困难和复杂的。

结合上述几个方面的内容，即可得到一个完整的人脸分析。如下图1.2所示。一个完整的人脸分析系统包括人脸检测、人脸对齐、特征表达、人脸识别及人脸属性识别五个部分。



##### 图 1.2 人脸分析示例

Fig 1.2 Example of Face Analysis

人脸识别技术优势。人脸识别作为一种新兴的生物特征识别技术，与虹膜识别、指纹扫描、掌形扫描等技术相比，人脸识别技术在应用方面具有独到的优势如下所示：

* 使用方便，无接触，用户接受度高；
* 低成本，直观性突出；
* 识别精确度高，速度快；
* 不易仿冒，可获取其它信息，如：表情，人脸属性等；
* 使用通用性设备；
* 基础资料易于获得。

智能监控系统人脸识别关键问题。(1) 光照问题。光照变化是影响人脸识别性能的最关键因素，对该问题，需要将人脸图像中固有的人脸属性和光源、遮挡及高光等非人脸固有属性分离开来，在人脸图像预处理或者归一化阶段进行针对性的光照补偿，以便消除非均匀正面光照造成的阴影、高光等对识别性能的影响。(2) 遮挡问题。对于非限制条件下的人脸识别，遮挡是一个非常严重的问题。特别是在监控环境下，往往彼监控对象都会带着眼镜，帽子等饰物，使得被采集出来的人脸图像有可能不完整，从而影响了后面的特征提取与识别。(3) 去冗问题。要求人脸识别监控系统能对视频捕捉中的画面能够快速的检测单个和多个人脸图像，并自动去冗余，减除重复的画像，并提取相应的人脸图像特征实现人脸的快速比对，并输出相应的结果信息。(4) 姿态问题。姿态问题涉及头部在三维垂直坐标系中绕三个轴的旋转造成的面部变化，目前多数的人脸识别算法主要针列正面、一定角度的侧脸的人脸图像，当发生俯仰或者左右侧而比较厉害的情况下，即垂直于图像平面的两个方向的深度旋转会造成面部信息的部分缺失，从而人脸识别算法的识别率也将会急剧下降。(5) 样本缺乏。基于统计学习的人脸识别算法是目前人脸识别领域中的主流算法，但是统计学习方法需要大量的训练。由于人脸图像在高维空间中的分布是一个不规则的流形分布，能得到的样本只是对人脸图像空间中的一个极小部分的采样，如何解决小样本下的统计学习问题有待进一步的研究。另外，还有其它比如年龄、图像质量、大规模人脸图像检索都是一个实际的智能监控系统需要考虑的问题。下图1.3为室外典型监控场景，其面临着上述提到的这四类关键问题，即光照、遮挡、姿态、冗余等对检测和识别有非常大的影响。



图 1.3 室外典型监控场景

Fig 1.3 The typical outdoor surveillance scenario

### **1.2.2** 研究团队与发展现状

首先，智能视频监控在最近几点越来越受到国内外学者的关注。而在智能监控环境下的人脸识别问题是一个研究热点。在计算机视觉、模式识别领域有两大顶级权威期刊和三大视觉会议。其中两大期刊为国际计算机视觉期刊(International Journal of Computer Vision, IJCV)和模式分析和机器智能汇刊(IEEE Transaction On Pattern Analysis and Machine Intelligence, PAMI)。三大会分别为国际计算机视觉会议(IEEE International Conference of Computer Vision, ICCV)、计算机视觉和模式识别会议(IEEE Conference on Computer Vision and Pattern Recognition, CVPR)和欧洲计算机视觉会议(European Conference on Computer Vision, ECCV)。在这些国际主流的期刊和会议当中都出版了针对智能视频监控的专刊，并且针对非限制条件下的人脸识别问题都取得了非常大的突破。人脸识别的最新理论和应用成果都会发表在上述期刊和会议当中。值得一提的是，随着工业界的参与以及深度学习的发展，很多高科技公司都成立了人工智能相关的研究院或研究小组，如谷歌、facebook的人工智能研究院、百度的深度学习研究院等等。

随着人脸识别的发展，在国际上也涌现出了一大批专门从事人脸识别研究的优秀团队。如微软亚洲研究院(MSRA)的计算机视觉小组、美国麻省理工(MIT)情感计算实验室、美国卡耐基梅隆大学(CMU)的计算机视觉小组、加州大学伯克利分校(UCBerkeley)计算机视觉小组、法国国家研究院(INRIA)的人脸和人体研究组等。在国内，有目前在国际主流期刊和会议上取得突破性进展的香港中文大学汤晓鸥、王晓刚带领的多媒体实验室、中科院自动化所模式识别国家重点实验室的谭铁牛、李子青研究组，以及中科院计算所专门研究人脸识别的山世光、清华大学艾海舟等。值得一提的是，国内的一些互联网公司及创业公司在国际公开的数据库上也取得了顶尖的效果。如腾讯优图团队、SenseTime、Linkface、Face++等。

### **1.2.3** 相关典型应用与系统

未来智能家居环境中智能安防是必不可少的一大模块，而智能安防监控系统匹配视频监控必然会搭配先进的人脸识别技术，通过与全国人脸数据库共享数据可以通过计算分析在最短时间内得出分析结果，让任何犯罪分子都无处可藏。在国际上，也出现了一些针对智能监控系统的商业系统，如IBM的“智能监控系统”(S3：Smart Surveillance System)，简称“S3”[[[1]](#footnote-1)0]系统。随着技术的发展视频监控逐渐向网络化、高清化、智能化方向发展，这也给人脸识别带来了好处，使得图像质量对人脸识别的影响越来越小[9,11]。以下是全球面向智能监控环境下人脸识别典型系统。

1. 美国Animetrics公司ForensicaGPS人脸识别工具1. 美国Animetrics是一家专注3D人脸识别技术的公司。其推出的一款新的人脸识别工具——ForensicaGPS，该产品最大的优势在于能够对低分辨率的监控图像或视频人脸识别。在监控场景中，面临着图像质量不清晰、分辨率低、人脸姿态变化多样等特点，Animetrics则有效地解决了上述场景下的人脸识别。ForensicaGPS核心技术在于能将低分辨率的2D人脸图像转化为经人脸姿态校正处理后的3D图像，这样可以更加容易地识别嫌犯。

此外，ForensicaGPS系统还能帮助执法人员将人脸图像与嫌犯数据库做比对，即人脸确认，并最终确定匹配结果。ForensicaGPS能同时分析、比对多达5个有关某嫌犯的人脸图像，能对面部结构、几何特征与皮肤纹理进行2D、3D可视化处理，能识别脸部疤痕、痣、纹身等特征，并进行脸部特征距离计算等。据Animetrics称，目前该公司已经与DataworksPlus和BI2 Technologies等执法技术方案提供商合作，向美国国内执法机构提供ForensicaGPS产品。此外，Animetrics还与欧洲执法技术经销商德国UNIDAS公

司、瑞士UMCForensicsGmBH公司等一道共同推广该产品的应用。目前，ForensicaGPS已经被瑞士、卡塔尔、阿联酋和沙特阿拉伯等国家的警察部门采用，如下图1.4所示。

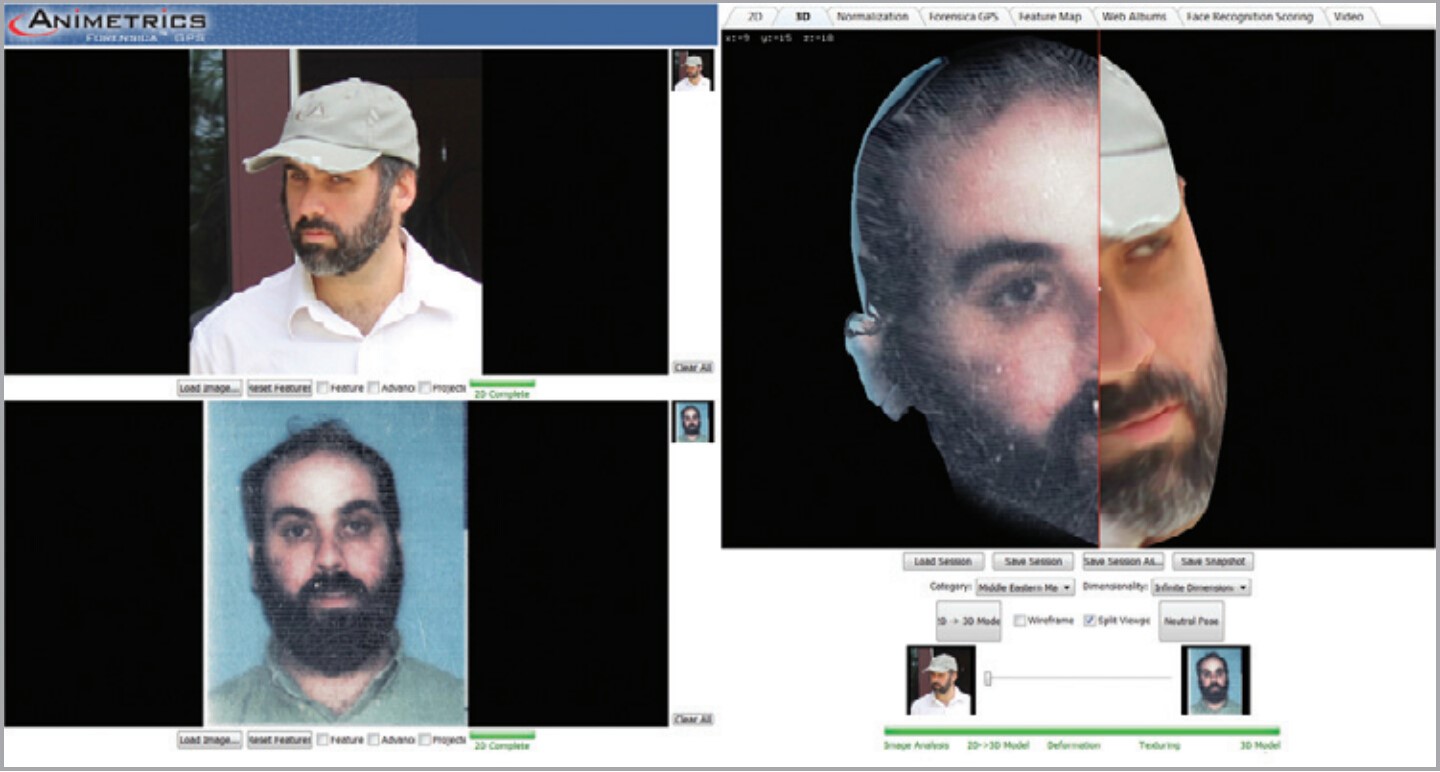


图 1.4 美国ForensicaGPS智能监控人脸识别系统

Fig 1.4 American ForensicaGPS Facial Analysis System

1. 美国ISS公司SecurOS Face人脸采集与识别产品[[2]](#footnote-2).

SecurOS Face是世界上正确识别率最高的人脸识别系统之一，它既是一款行业领先的视频分析产品，又是一套先进的人脸采集与识别解决方案。SecurOS Face可进行主动式/被动式目标人脸识别，不仅能作为监控模块进行独立工作，还可以集成至第三方监控系统中(如门禁系统)综合使用。本系统在合理光照条件下(不低于250勒克斯)的正确识别率不低于85%(一般为95%)，在同类产品中位列前茅。SecurOS Face人脸识别数据库容量不低于500万人，人脸识别时间(在10万个人脸档案的情况下)不超过1秒;人脸识别最快速度(单处理器配置)每秒7张人脸。 SecurOS Face功能齐备，无需采用特殊

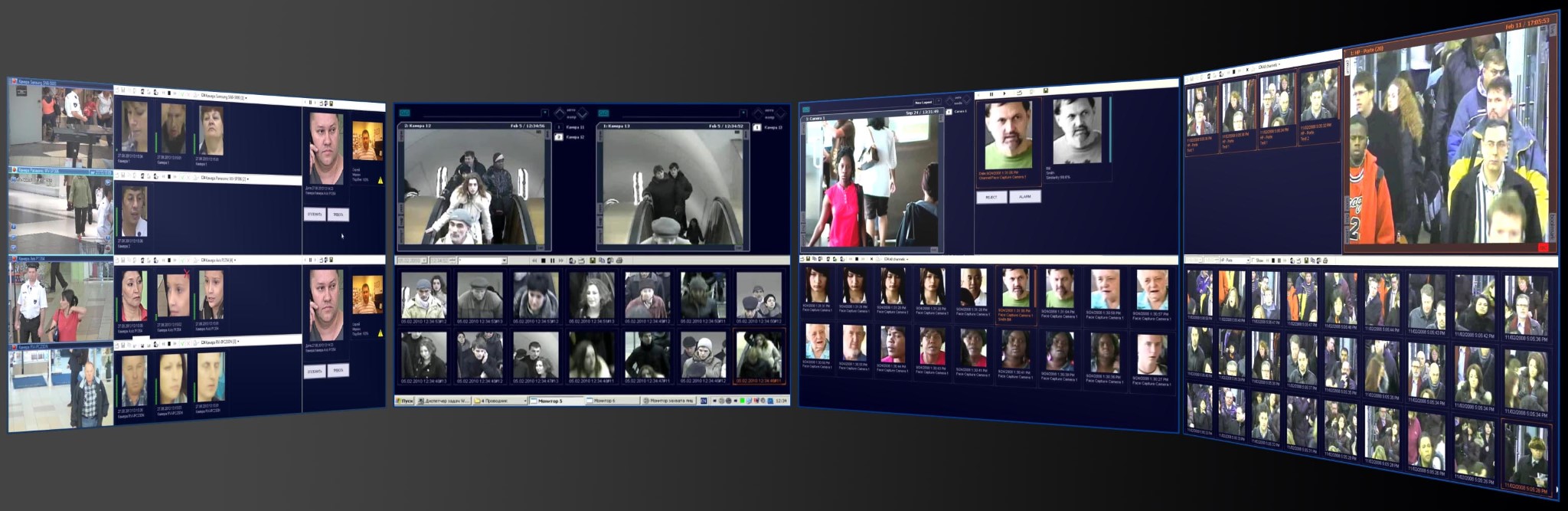


图 1.5 美国ISS公司SecurOS Face智能监控人脸识别系统

Fig 1.5 American ISS SecurOS Face Facial Analysis System

摄像头，能同时检测多个人脸图像，采集高速运动中的人脸图像，快捷访问所有与某人脸相关的视频，在目标区域进行快速搜索与采集，自动确定人脸最佳定位角度，自动传输与比对人脸图像，远程查询视频监控记录，在视频档案中进行快速信息搜索，通过电话、电子邮件、短信息等发出自动报警提示等。此外，SecurOS Face具有分布式模块化结构，每台服务器的人脸采集设备不受限制，能采集多个人脸图像，且都对一个中央人脸识别数据库进行支持。

SecurOS Face特点突出：第一，使用方便，图形用户接口(GUI)操作简单，可以很方便地使用AVI和JPEG视频格式输出、数字变焦、颜色调整、对比度调整、远程视频摄像头调整、基于识别结果的自动通知、事件分析、详细数据输出与报警门限设定等多项功能；第二，集成度高，能与安全管理系统、访问控制系统、软件信息系统、外接数据库等实现联动，此外，本产品的远程视频查看、视频管理和视频存储等功能可同时运行；第三，方式灵活，可调整事件响应或报警方式(如通过声音、电话、电子邮件等报警)，还能根据用户需求，实现自动人脸定位、自动最佳人脸选择等功能调整；第四，稳定可靠，在不同光照环境下的人脸识别正确率稳定，可靠性强，还能与任意一款高分辨率的模拟摄像头结合使用。图1.5展示了SecureOS Face 人脸识别系统。

3、以色列NICE公司NiceVision智能视频解决方案[[3]](#footnote-3).

以色列NICE公司NiceVision智能视频解决方案能够实时探测和识别威胁，以应对最严峻的安全挑战。该方案包括智能摄像机、ControlCenter解决方案、NiceVision视频分析模块、NiceVision嫌疑人搜索模块，同时还带有扩展性SDK，可使第三方安全系统很容易地集成在NiceVision上。如下图1.6所示。

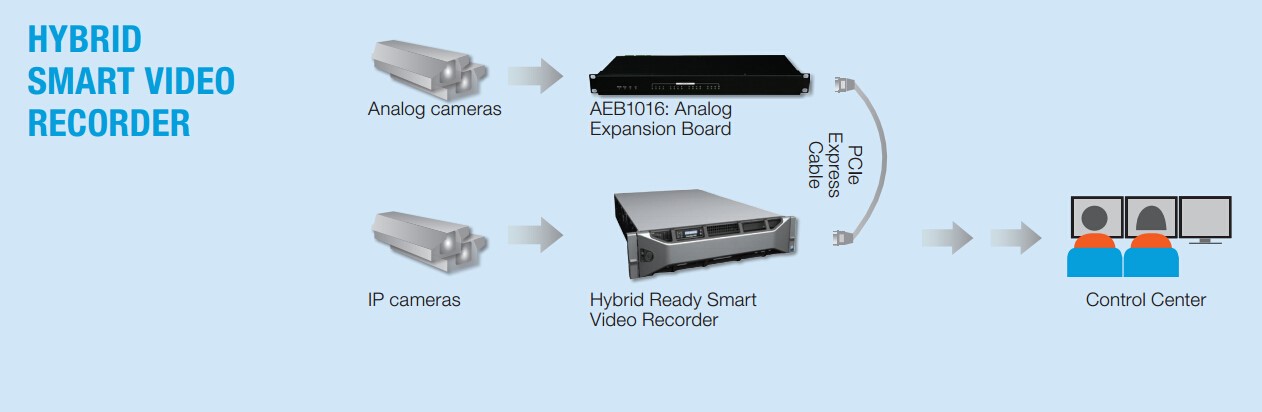


图 1.6 以色列NICE公司NiceVision智能监控系统

Fig 1.6 Israel NiceVision Smart Video Surveillance System

控制中心解决方案的主要特点是：多显示器、多选项卡视频管理；高级用户配置文件和特权管理；高级事件管理；简单视频导出和归档；视觉导航搜索；与智能手机和网络客户无缝连接；简化管理和减少拥有替代传统模拟矩阵的NiceVision虚拟矩阵(VMX)的成本；高级和视觉查询功能和同步视频。

NiceVision视频分析模块通过经现场验证的算法，对即将发生的事件发出主动提醒，可用于入侵检测、人群管理，此外还能够控制变量，如天气状况、移动车辆或照明的变化，对反向活动的物体或人进行识别并发出警报，在同一视频通道上使用多个分析应用程序，利用云台控制(PTZ)跟踪能力以获得更高的探测率。

NICE嫌疑人搜索模块是下一代视频分析系统，具有以下特点：在几秒钟内搜索图片，迅速滤除90%的无关图片，3至5秒显示匹配情况，准确度高，可用地图定位和跟踪，非常快地找到可疑人物。

4、德国Cognitec公司FaceVACS-VideoSan智能视频监控系统[[4]](#footnote-4).

德国Cognitec是一家专门从事生物特征识别领域技术研发的公司，其推出了的新一代人脸识别视频监控系统FaceVACS-VideoSan，该系统能够接入多路视频流，进行多人脸检测，并与监视名单中的人脸图像进行匹配。如下图1.7所示。

FaceVACS-VideoSan系统不仅仅针对监控系统进行人脸比对，还可应用于特定场景下的人脸识别，如VIP客户识别、目标客户识别等。即该系统不但实现了人脸确认的任务，而且实现了人脸辨识的任务。此外，FaceVACS-VideoSan系统还能实时追踪多路视频流中的人脸，实时比对监视名单中的人脸图像，若匹配成功或检测到异常事件

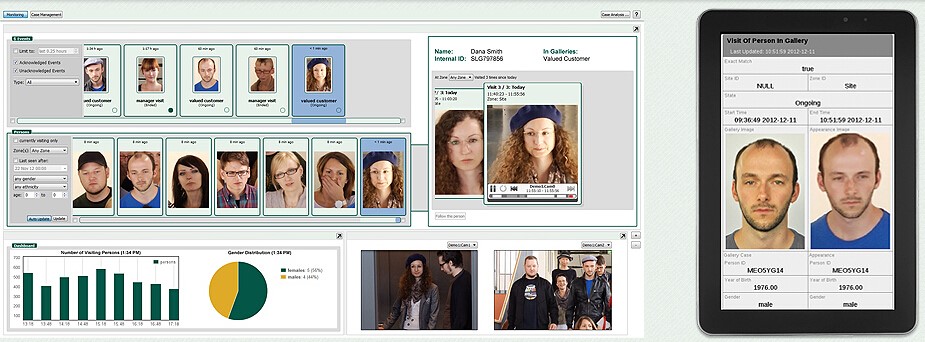


图 1.7 德国Cognitec公司FaceVACS-VideoSan智能监控系统

Fig 1.7 Germany Cognitec FaceVACS-VideoSan Smart Video Surveillance System

便发出警告并记录。另外该系统可根据监视名单大小、视频流数量和可视人脸数量的不同进行拓展。

5、德国博世公司IVA 4.0智能视频分析系统[[5]](#footnote-5).

IVA 4.0是博世公司推出的第4代智能视频分析系统，它能用于室内外移动目标的检测、跟踪与分析，在提高准确率的同时也能降低虚报警率。IVA4.0可与博世DinionIP、

Extreme IP、FlexiDome IP、AutoDome IP等摄像机，以及VIP X1600、VideoJet X系列编码器配合使用以提高综合性能。如图1.8所示。



图 1.8 德国BOSCH公司IVA 4.0智能监控系统

Fig 1.8 Germany Cognitec FaceVACS-VideoSan Video Surveillance System

IVA 4.0具有多种性能：第一，运动检测，先进的内置算法能使其适应不同的光照条件和复杂多变的天气环境。图像稳定技术可确保摄像头在摇晃的时候也能正常工作，不断改进的背景学习算法能使其具备最新的智能化处理功能，内置的篡改检测功能可以对摄像机掩盖、散焦、移位等情况发出警报；第二，滤波处理，用户可以选择图像敏感区域，检测感兴趣的运动目标，并创建具体的检测规则，检测结果统计数据可以反馈作用于滤波功能调整；第三，智能分析，嵌入式智能分析模块能够决定何时开始采集视频数据，其占用带宽更少，占用空间更小，该功能可以产生结构化数据以描述所记录场景内容，可以为司法调查提供证据，还能够在数秒之内扫描海量数据；第四，图形接口管理，可提供所有必要的工具来轻松设置IVA 4.0，图形化显示的任务管理器便于用户设置敏感区域、拍摄尺寸、拍摄速度和拍摄角度等。

6、印度Inttelix公司Suspect Detection人脸识别监控系统[[6]](#footnote-6).

Inttelix是一家专门研究准确生物特征识别系统的公司，公司总部位于印度。该公司推出了一系列人脸识别相关的产品，并在校园、安防、零售业、智能交通、访客管理等领域都有相关的应用。Inttelix公司推出的Suspect Detection System(Inttelix® SDS)系统可实时检测视频中的人脸图像，并针对特定的场景进行人脸确认和辨识任务。如下图1.9所示。

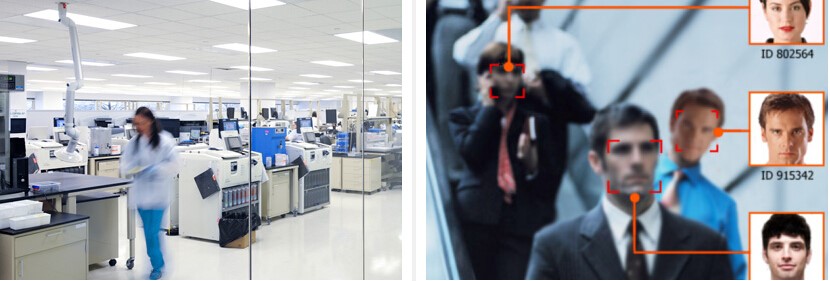


图 1.9 印度Inttelix公司Suspect Detection智能监控系统应用实例

Fig 1.9 An Application for Indian Inttelix Suspect Detection Video Surveillance System

Suspect Detection作为监控系统具有视频记录的功能，同时自动检测监控区域中出现的人脸图像。系统的工作流程是：首先将嫌疑人图像加入比对数据库，然后在全球各个数据库当中共享，最后在监控视频中采集图像检测人脸，并对视频中检测到的人脸和数据库中的人脸图像进行比对。实现1:N的人脸比对，即人类辨识任务。

## **1.3** 本文主要内容及结构

智能视频监控当中的人脸识别问题是一个非常有挑战性的课题。从某一方面讲，

视频监控是人脸识别任务当中的非常重要的应用。所以，本文主要研究的内容是人脸识别问题。对一个人脸识别系统来拆分，主要包括人脸检测、人脸特征点检测（人脸配准）、人脸识别三个部分。而每个部分都会对一个人脸识别系统的性能有非常重要的影响。因此，本课题按照人脸检测、人脸特征点定位、人脸识别“三步走”的路线展开研究，最终得到一个鲁棒的人脸识别系统。论文的结构及内容也按照“三步走” 的路线分别进行撰写，并且最后给出每个部分的实验结果和分析。下面是文章的主要结构及研究思路：

第一章主要介绍了人脸识别在智能监控应用的大背景。其中主要描述了人脸识别问题的前沿难点问题，以及国内外研究现状。针对人脸识别在监控领域这个特殊领域，介绍了主流的人脸识别智能监控应用系统。

第二章为“三步走”战略的第一步，即人脸检测。监控系统中人脸姿态、光照、形变、遮挡现象比较常见，针对智能监控环境下人脸检测面临的挑战，采用了一种基于级联可变形部件模型，并与基于Haar Adaboost人脸检测算法进行了比较，实现了针对监控环境下多姿态人脸在接近零误检率下的实时检测。

第三章主要介绍了人脸特征点的主要工作。人脸检测可以看作人脸特征点检测的预处理，而人脸特征点检测（人脸配准）可看作后续人脸识别的预处理。由人脸检测部分可以知道，人脸特征点仍旧面临着人脸检测所面临的难点，因为对特征点的检测与定位都在人脸检测的基础上进行。针对人脸特征点定位当中的人脸姿态、形变等问题，我们首先介绍了传统的基于AAM模型的人脸特征点定位方法，然后介绍了本文采用的两级级联回归模型用于特征点定位，并提出了一种自适应的人脸形状初始化策略，提高了人脸特征点定位的效率，实现了实时的人脸特征点定位。

第四章主要介绍了人脸识别的相关工作。人脸识别包含两个任务：人脸确认（Face Verification）和人脸辨识（Face Identification）。人脸确认即判断两张人脸图片是否为一个人，是一个二分类问题。在此部分通过单任务度量学习的方法实现了人脸确认任务。人脸辨识是给定一张人脸照片，判断其身份的问题，其本质是一个多分类问题。针对人脸辨识这个任务，本文提出了一种两级的多任务度量学习方法，将多任务标签的信息引入到了人脸辨识任务中，实现了准确高效的人脸辨识任务。

第五章是实验验证和结果分析部分。此部分对人脸检测、特征点定位、人脸识别

（人脸确认和人脸辨识）三部分都进行了有效的实验验证。

最后，对本课题研究的人脸识别相关的问题进行了总结，分析了其中存在的不足之处，并提出进一步的改进方案，为今后的研究确定方向。

# 第二章 复杂姿态下人脸检测

人脸检测(Face Detection)的目的是为了确定图像中人脸的位置、大小和数量，更加精细的人脸检测还包括人脸的朝向及姿态。人脸检测是所有人脸图像分析的预处理过程，经过多年发展，人脸检测在检测准确率和实时性方面都取得了很大的突破，研究者们针对各种复杂场景也提出了大量的人脸检测算法。本章将从人脸检测研究概述、人脸检测和识别当中的特征、人脸检测方法三个方法进行论述。在人脸检测方法当中，主要介绍经典的基于Haar Adaboost的方法和本文采用的级联可变形部件模型人脸检测方法，并在后续的实验章节对两种方法进行了对比。

## **2.1** 人脸检测研究概述

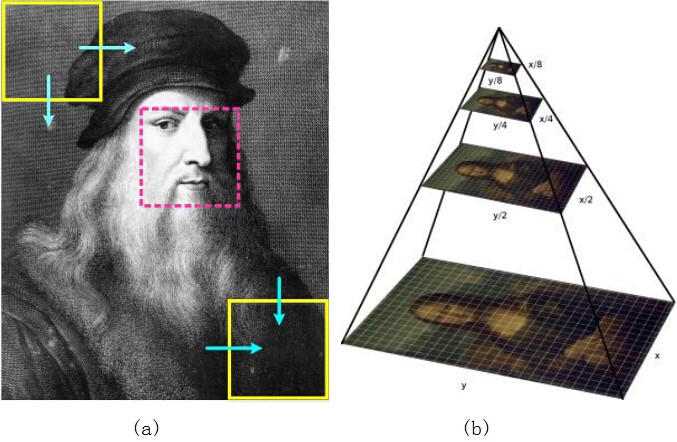
人脸检测是人脸识别的基础，如果人脸检测出现错误，后续的人脸识别基本无法进行。衡量人脸检测算法的好坏主要有以下三个指标：（1）召回率，即正确检测到的人脸数量与数据库中总的人脸数量之比。这个指标反映了检测器可以正确找到多少人脸。（2）精确度，即在所有输出的人脸中，正确的人脸所占的比例。这个指标反映了检测器检出结果的可靠程度。（3）稳定度，即当人脸在图片中的位置和大小发生变化或者人脸保持一定姿态不动时，输出的人脸位置和大小也需要相对于真实人脸保持不变，即检测的窗口不能发生闪烁现象。越稳定的人脸检测结果对于后序的人脸校准操作越有利。这三个指标越高，人脸检测算法越好。到目前为止，最具有里程碑意义的人脸检测算法是文献[1]，它奠定了现代人脸检测算法的基础。为了解决文献[1]不能很好的处理多角度的问题，随后也出现了许多针对多角度人脸检测的经典算法[12–16]。有关人脸检测技术的发展，读者可以参考综述文献 [17]。

人脸检测早在上世纪70年代就已经有人开始研究，然而受当时落后的技术条件和有限的需求所影响，直到上世纪90年代，人脸检测技术才开始加快向前发展的脚步，在新世纪到来前的最后十年间，涌现出了大量关于人脸检测的研究工作，这时期设计的很多人脸检测器已经有了现代人脸检测技术的影子，例如可变形模板的设计（将人脸按照五官和轮廓划分成多个相互连接的局部块）、神经网络的引入（作为判断输入是否为人脸的分类模型）等。这些早期的工作主要关注于检测正面的人脸，基于简单的底层特征如物体边缘、图像灰度值等来对图像进行分析，结合关于人脸的先验知识来设计模型和算法（如五官、肤色），并开始引入一些当时已有的的模式识别方法。虽然早期关于人脸检测的研究工作离实际应用的要求还有很远，但其中进行检测的流程已经和现代的人脸检测方法没有本质区别。给定一张输入图像，要完成人脸检测这个任务，我们通常分成三步来进行：

* 选择图像上的某个（矩形）区域作为一个观察窗口；
* 在选定的窗口中提取一些特征对其包含的图像区域进行描述；
* 根据特征描述来判断这个窗口是不是正好框住了一张人脸。

检测人脸的过程就是不断地执行上面三步，直到遍历所有需要观察的窗口。如果所有的窗口都被判断为不包含人脸，那么就认为所给的图像上不存在人脸，否则就根据判断为包含人脸的窗口来给出人脸所在的位置及其大小。

那么，如何来选择我们要观察的窗口呢？所谓眼见为实，要判断图像上的某个位置是不是一张人脸，必须要观察了这个位置之后才知道，因此，选择的窗口应该覆盖图像上的所有位置。显然，最直接的方式就是让观察的窗口在图像上从左至右、从上往下一步一步地滑动，从图像的左上角滑动到右下角——这就是所谓的滑动窗口范式，你可以将它想象成是福尔摩斯（检测器）在拿着放大镜（观察窗口）仔细观察案发现场（输入图像）每一个角落（滑动）的过程。其滑动窗口的机制如下图2.1. (a)所示。



#### 图 2.1 (a) 滑动窗口 (b) 图像金字塔

Fig 2.1 (a) Sliding window (b) Image pyramid

对于观察窗口，还有一个重要的问题就是：窗口应该多大？我们认为一个窗口是一个人脸窗口当且仅当其恰好框住了一张人脸，即窗口的大小和人脸的大小是一致的，窗口基本贴合人脸的外轮廓。即使是同一张图像上，人脸的大小不仅不固定，而且可以是任意的，这样怎么才能让观察窗口适应不同大小的人脸呢？一种做法当然是采用多种不同大小的窗口，分别去扫描图像，但是这种做法并不高效。换一个角度来看，其实也可以将图像缩放到不同的大小，然后用相同大小的窗口去扫描——这就是所谓的构造图像金字塔的方式。图像金字塔这一名字非常生动形象，将缩放成不同大小的图像按照从大到小的顺序依次往上堆叠，正好就组成了一个金字塔的形状。如图2.1. (b)所示。通过构建图像金字塔，同时允许窗口和人脸的贴合程度在小范围内变动，就能够检测到不同位置、不同大小的人脸了。另外需要一提的是，对于人脸而言，通常只用正方形的观察窗口，因此不需要考虑窗口的长宽比等问题了。

然而，统计表明75%的家庭照片中的人脸并非正面人脸，并且在实际应用中由于人体的运动或摄像设备的移动，使得正面人脸图像的获取比较困难。另外在智能监控视频当中，由于摄像头位置和角度问题，检测的人脸大多为非正面。为此多姿态人脸检测算法受到越来越多的关注。Rowley将Viola的方法推广到多视角人脸检测，针对不同角度训练多个网络，并采用姿态预估计加快检测速度；Schneiderman等[18]将人脸划分为左侧面、正面和右侧面，分别训练了三个不同视图的检测器来分别处理这三种情况；Li等[19]则用SVM进行多姿态人脸的检测和识别；Feraud等[20]提出了一种受限生成模型（Constrained Generative Model），并利用该模型构造了四个人脸检测器分别针对左右偏转[0◦, 20◦]和[20◦, 40◦]的人脸进行检测。Li等[21]手动的将人脸训练样本分成几个较均匀的子视角类，提出了Floatboost学习方法，训练得到“由粗到细”的金字塔型结构人脸检测器，用于近乎实时的多视角人脸检测。

另外，人脸检测作为一种特定类型目标的检测任务，一方面具有其自己鲜明的特点，需要考虑人脸这一目标的特殊性，另一方面其也和其它类型目标的检测任务具有一定的共性，能够直接借鉴在通用目标检测方法上的研究经验。目标检测任务作为一个分类问题，其不仅受益于计算机视觉领域相关技术的不断发展，在机器学习领域的研究进展同样也对目标检测任务具有推波助澜的作用。事实上，从2006年开始逐步蔓延开的深度学习大爆发给目标检测的研究带来了强劲的助推力，使得通用的目标检测以及各种特定类型目标的检测任务得到了跨越式地发展。

经过几十年的研究和发展，人脸检测方法正日趋成熟，在现实场景中也已经得到了比较广泛的应用，但是人脸检测问题还并没有被完全解决，复杂多样的姿态变化，千奇百怪的遮挡情况，捉摸不定的光照条件，不同的分辨率，迥异的清晰度，微妙的肤色差，各种内外因素的共同作用让人脸的变化模式变得极其丰富，而目前还没有检测器可以同时对所有的变化模式都足够鲁棒。

总结人脸检测多年的发展历程，我们可以发现人脸检测的发展某一方面也是人脸特征表示的发展。人脸检测的核心问题在于对人脸和非人脸的图片进行有区分力的分类，因而人脸特征表示至关重要。因此，在下文当中，本章首先介绍人脸检测过程当中常见的特征，然后介绍人脸检测领域经典的人脸检测方法以及我们自己采用的方法，并在后续的章节中进行实验对比及验证。

## **2.2** 人脸检测和识别中的特征

人脸特征表示的好坏，直接关系到了人脸检测与识别的精度的高低。因此，如何得到具有表达力的人脸特征，就是人脸检测与识别最关键的问题。纵观30多年的人脸识别的研究历史，我们发现人脸识别技术的发展本质上也就是人脸特征的发展。特征提取的作用就是根据已知的人脸的位置信息，提取出对于人脸识别有利的信息,如面部区域、纹理结构等，而去掉和人脸识别无关的，如背景、噪声等信息。一个优秀的人脸特征需要满足以下两个要求：（1）该特征需要对不同人之间的细微区别十分敏感。（2）同时，该特征需要对人脸角度、光照、表情等变化具有一定的不变性。目前

人脸识别技术中，运用的最多的特征有以下几种LBP[22]，SIFT[23]，HOG[24]，Gabor[25]，

LE[26]等。这些特征在广泛的应用中均取得了不错的识别精度。目前，由于深度学习（deep learning）的发展，也有许多基于深度学习的算法将特征提取和特征学习融合到同一个框架中，其中包括：DeepFace[27]、DeepID[28]、FaceNet[29]等。由于深度神经网络强大的拟合能力，这类方法在各种人脸识别问题中均取得了十分优秀的结果。但是由于这类方法网络结构复杂，模型参数一般都在百万量级。模型大小一般为几百兆，在实际使用中不太方便。另外，这类方法需要硬件支持并行计算，不适合手机或嵌入式系统上的人脸识别。因此，深度学习算法一般运用在云计算中。目前有许多公司提供基于深度学习的人脸云计算服务，例如:Face++、Microsoft、SenseTime、Tencent等。

### **2.2.1 Haar-Like**特征

（1）Haar-Like特征简介

Haar-like特征最早是由Papageorgiou等人提出并应用于人脸表示，Viola和Jones在此基础上使用了3种类型4种形式的特征提出了一种新的人脸检测器，检测VJ人脸检测器。在人脸检测的过程中，需要AdaBoost级联分类器，分类器的训练则要在图像中选取一些简单特征，这些抽取的特征由各个矩形特征组成，称为Haar特征。Haar特征是计算机视觉领域一种常用的特征描述算子，基本Haar特征原型只有四种，而扩展的类Haar特征分为三类：边缘特征、线性特征和中心特征，它们组合成为特征模板。如图2.2所示[30]。

所谓Haar特征，其实就是在窗口的某个位置取一个矩形的小块，然后将这个矩形小块划分为黑色和白色两部分，并分别对两部分所覆盖的像素点（图像上的每个点称为一个像素）的灰度值求和，最后用白色部分像素点灰度值的和减去黑色部分像素点

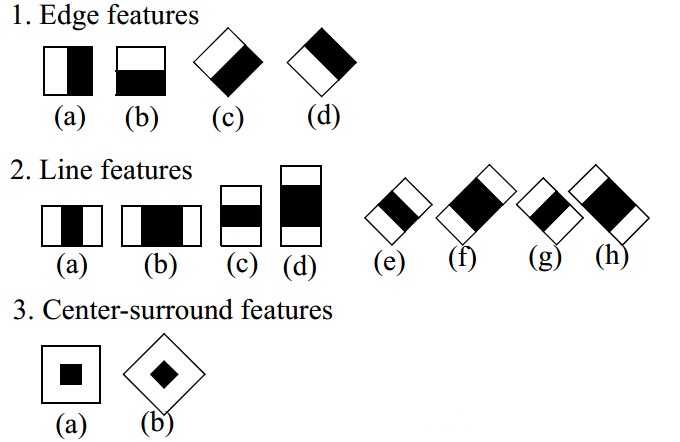


图 2.2 类Haar特征

Fig 2.2 Haar-like features

灰度值的和，得到一个Haar特征的值。Haar特征反映了局部区域之间的相对明暗关系，能够为人脸和非人脸的区分提供有效的信息，例如眼睛区域比周围的皮肤区域要暗，通过Haar特征就可以将这一特点表示出来。但是由于提取Haar特征时每次都需要计算局部区域内多个像素点灰度值之和，因此在速度上其并不快，为此VJ人脸检测器引入了积分图来加速Haar特征的提取。

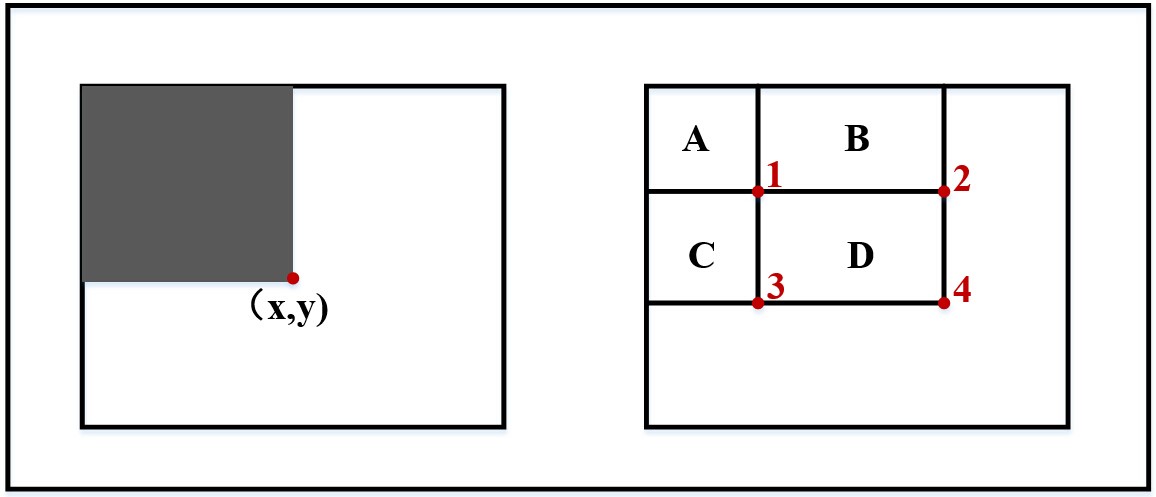
（2）积分图的快速计算

积分图是一张和输入图像一样大的图，但其每个点上不再是存放这个点的灰度值，而是存放从图像左上角到该点所确定的矩形区域内全部点的灰度值之和。积分图所带来的好处是两方面的，一方面它使得每次计算局部区域像素点的灰度值之和仅需要做4次加减法，与局部区域的大小无关；另一方面它避免了在相同像素点上重复求和，只在最开始计算一次——相邻的窗口有很大的重叠部分，对应的Haar特征也会重叠，如果每次都重新计算像素点的灰度值之和，则重叠部分的计算是重复的。积分图极大地加速了Haar特征的提取，向快速的检测器迈出了第一步。积分图的表示方法对Haar特征的快速计算有重要意义。积分图表示方法如图2.3所示。

在这个积分图像中，每一点的值表示图像中该点位置左上部分区域内的所有像素的灰度值积分。如公式(2.1)所示。

*ii*(*x*, y) = ∑ *i*(*x*′, y′) (2.1) x′≤x,y′≤y

其中*ii*(*x*, y)是图像中(*x*, y)点的像素灰度值积分，*i*(*x*′, y′)是图像中(*x*′, y′)点的像素灰度值。如果计算积分图像中某一区域的像素灰度值积分，则由该区域四个顶点对应的积分值作相应的加减运算得到。例如，区域D的积分*S*D可由图像中1，2，3，4点的积分



##### 图 2.3 积分图

Fig 2.3 Integral image

值*S*1，*S*2，*S*3，*S*4计算得到，即：

*S*D = *S*4 + *S*1 − *S*2 − *S*3 (2.2)

图像中每点的积分值保存在内存中，计算某区域积分值变成了简单的加减运算，不必重新计算该区域像素和，计算消耗的时间是常量，所以不管图像尺寸如何，积分图表示方法可以降低特征值的计算复杂度。

### **2.2.2 LBP**特征

（1）基本LBP

局部二元模式（LocalBinaryPattern, LBP）是一种用来描述图像局部纹理特征的算子。1996年，Ojala[22]等人提出了原始LBP算子，并用来提取图像的纹理特征，并且验证了它在纹理分类中的强大能力。由于其在纹理分类中优良的有效性及其计算简单，且对单调的灰度变化具有不变性等特点，由于LBP是整数特征，计算速度很快，常被普遍应用于人脸检测、人脸识别以及表情分析和背景建模等领域。一般来讲，基本的LBP算子定义为在3 × 3的窗口内，以窗口中心像素为基准点，将相邻的8个像素的灰度值与其进行比较，若周围像素值大于中心像素值，则该像素点的位置被标记为1，否则为0。这样，3 × 3领域内的8个点可产生8比特（bit）的无符号数，即得到该窗口的LBP值，经过遍历每个像素点后，即可得到经过LBP算子之后的效果图，这种图纹理信息。如图2.4所示。

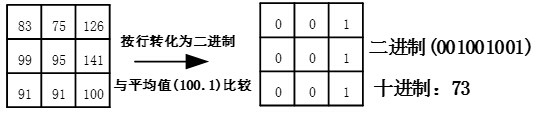
上述十进制的形式可以用一个8比特的一个字节来表示，其表达式为：

7

*LBP*(*x*c, yc) = ∑ *s*(*i*n − *i*c)2n (2.3)

n=0

其中，对于像素点位置(*x*c, yc)，其对应的像素的值为*i*c，并作为中心点位置，将有8个



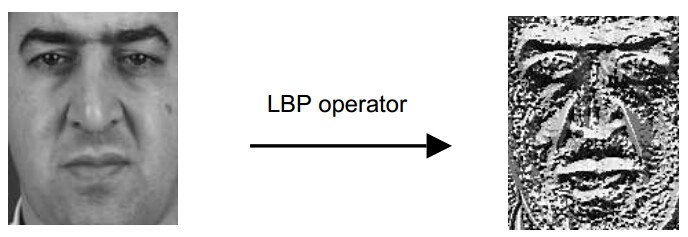
##### 图 2.4 基本LBP算子示意图

Fig 2.4 Illustration of basic LBP

像素点*i*n(*n* = 0, 1, 2...7)对其依次环绕，左上角为第一个像素点。二值化处理函数定义为：

|  |  |
| --- | --- |
| 1 *i f x* ≥ 0  *s*(*x*) =   0 *i f x* < 0 | (2.4) |

下图2.5展示了一张人脸图片经过LBP算子 运算之后得到的图像。



##### 图 2.5 原始图像（左）经LBP算子处理后的图像（右）

Fig 2.5 Original image (left) processed by the LBP operator (right).

1. 多尺度LBP

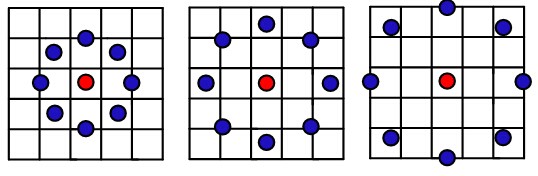
Ojala等人在后续工作中又对基本的LBP算子进行了拓展，使得邻域的范围不单局限在3 × 3 的邻域，如图2.6所示。为了避免图像旋转中LBP值变化的缺陷，提出了旋转不变模式的LBP，更进一步得到了等价模式的LBP。使用圆形邻域以及双线性差值的方法就可以将LBP算子扩展到任意圆半径和任意邻域像素数目，使用表示*LBP*P,R方法来表示领域参数，*P* 表示邻域像素个数，*R*表示领域的圆周半径，图2.6分别依次给出

了*LBP*8,1、*LBP*8,2和*LBP*8,2.5三种模式。

1. 旋转不变LBP

邻域有*P*个点的LBP算子可以产生2P种不同的模式，每个模式对应着*P*个相邻像素所组成的2P个不同的二进制值。当图像发生旋转时，*P*个相邻像素也将绕着中心点沿圆形邻域的圆周旋转一定的角度，因此会使得*LBP*P,R值发生改变。为了避免这个缺陷，Ojala又对*LBP*P,R值进行重新定义，提出了具有旋转不变性的局部二值模式为：

*LBP*Pri,R = *min* {*ROR*(*LBP*P,R,*i*)|*i* = 0, 1, · · · , *p* − 1} (2.5)



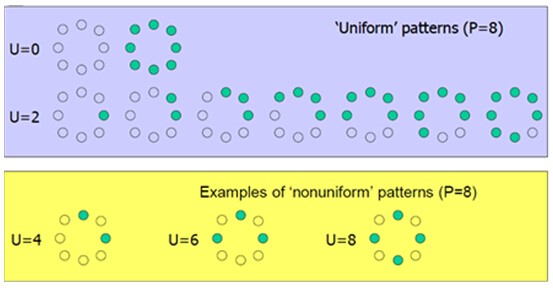
##### 图 2.6 三种多尺度LBP

Fig 2.6 Three type of multi-scale LBP

上式中，*LBP*Pri,R代表旋转不变 LBP 算子，*ROR*(*x*,*i*)为定义在*P*位数值*x*上的*i*位循环右移。如果对应图像采样的像素点来说，也就是对这*P*个像素点按照顺时针方向转动*i*次。由公式(2.5)看出，对于那些不断旋转得到的LBP模式取其最小值就是LBP的旋转不变模式。引入旋转不变模式定义后，使得LBP纹理特征对于图像旋转表现得更加鲁棒并且也减少了LBP模式的种类。

（4）LBP等价模式

尽管*LBP*Pri,R在理论上旋转不变性的性质，但在实际的应用中，并不具有很强的分类能力。这主要是因为具有旋转不变性的*LBP*Pri,R的不同模式的出现频率差异非常大造成的，其中有些特定的LBP模式出现频率达90%以上，大大超过其它模式的出现频率。即模式中对应二进制串0和1的跳变次数为0或者只有两次跳变，这些特定的 LBP 模式被称为等价模式(Uniform LBP)。而对应的非等价模式，即为模式中对应二进制串0和1的跳变次数大于两次[31]。如图2.7，给出了等价模式和非等价模式的例子。



##### 图 2.7 等价模式与非等价模式示例

Fig 2.7 Examples of uniform and nonuniform patterns

由图2.7可以看出，采用了等价模式后，二进制模式大大减少，模式的数量从最开始的2P种减少到了*P*(*P* − 1) + 2种。实验表明，等价模式的LBP算子不受图像旋转和平移的影响，能够有效地表达纹理特征，又能起到降维的作用，并且对暗点、亮点、平滑区域、边缘等也具有较好的描述能力。

### **2.2.3 HOG**特征

HOG(Histogram of Oriented Gradient)特征由Navneet Dalal和Bill Triggs于2005年提出[24]，应用于行人检测问题。在该特征被提出后，同样在物体检测(Object Detection)研究中得到了应用。由于HOG特征在目标检测领域取得了巨大的成功，它逐渐被扩展用于图像处理和模式识别的其他领域，比如：行为分析，场景分类，图像检索等等。

HOG特征有很多优点，其本质在于在一幅图像中，局部目标的表象和形状能够被梯度或边缘的方向密度分布很好地描述。首先，由于HOG是在图像的局部方格单元上操作，所以它对图像几何和光学的形变都能保持很好的不变性，这两种形变只会出现在更大的空间领域上。其次，在粗的空域抽样、精细的方向抽样以及较强的局部光学归一化等条件下，只要行人大体上能够保持直立的姿势，可以容许行人有一些细微的肢体动作，这些细微的动作可以被忽略而不影响检测效果。因此HOG特征是特别适合于做图像中的人体检测的。目前，学术界将HOG特征分为HOG和FHOG这两种，下面将分别介绍。

（1）原始HOG特征

对图片计算原始的HOG特征包括了梯度的计算，块（Block）和细胞（Cell）的分割，归一化特征权重和全局扫描图像这四个部分，下面将分别介绍这四个部分。

梯度计算。某像素点(*x*, y)点的梯度包括水平方向梯度*G*x和垂直方向梯度*G*y。其由公式(2.6) 和公式(2.7) 分别确定。最常用的方法是：首先用[−1, 0, 1]梯度算子对原图像做卷积运算，得到*x*方向（水平方向，以向右为正方向）的梯度分量gradscalx，然后用[1, 0, −1]T梯度算子对原图像做卷积运算，得到y方向（竖直方向，以向上为正方向）的梯度分量gradscaly。然后再用以下公式计算该像素点的梯度大小和方向。

|  |  |
| --- | --- |
| *G*x(*x*, y) = *H*(*x* + 1, y) − *H*(*x* − 1, y) | (2.6) |
| *G*y(*x*, y) = *H*(*x*, y+ 1) − *H*(*x*, y − 1) | (2.7) |

|  |  |
| --- | --- |
| *G*(*x*, y) = √*G*x(*x*, y)[[7]](#footnote-7) + *G*y(*x*, y)2 | (2.8) |
| α(*x*, y) = *tan*−1(*G*y(*x*, y)) *G*x(*x*, y) | (2.9) |

其中，*H*(*x*, y)表示在像素点(*x*, y)处的像素值。求导操作不仅能够捕获轮廓，人影和一些纹理信息，还能进一步弱化光照的影响。根据梯度值的计算，可以得到图像中像素点(x,y)的梯度大小和方向分别为：

Block和Cell。计算出来图像所有像素点的梯度之后，需要做进一步的梯度统计，用到的统计单元就是Block和Cell，它们之间的对应关系如图2.8所示。

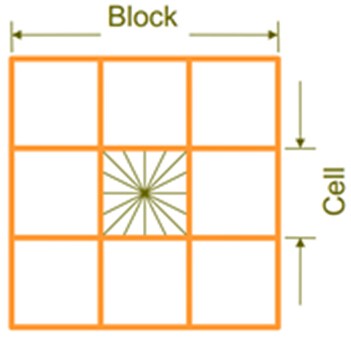


图 2.8 Block和Cell的关系

Fig 2.8 Relationship between Block and Cell

可以看出，一个Block实际上是由若干个Cell组成的，每个Cell其实是统计出来的

基于方向的梯度直方图。梯度直方图又可以分为无向直方图*B*1（Contrast Insensitive）和有向直方图*B*2（Contrast Sensitive）：

|  |  |
| --- | --- |
| *B*1(*x*, y) = *round*( )*mod p* π | (2.10) |
| *p*θ(*x*, y)  *B*2(*x*, y) = *round*( )*mod p* | (2.11) |

*p*θ(*x*, y)

于(*x*, y)处的方向梯度特征向量为：



##### *G*(*x*, y), *i f b* = *B*(*x*, y)  ∈ 0{, −*other*} w*ise*



*F*(*x*, y)b =(2.12)

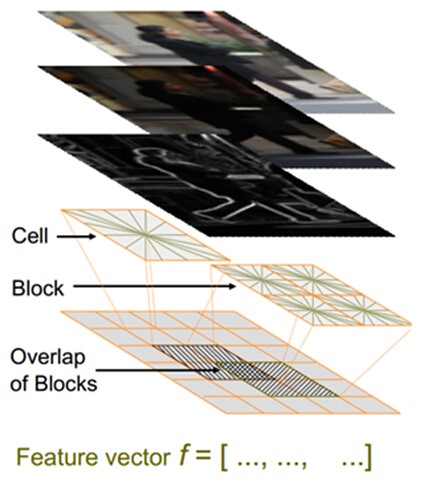
其中，*b*为直方图任意一个通道，即*b* 0, ..., *p* 1 。*p*为总通道数。Cell中每个像素点的梯度方向落入某个通道区域，则它对应的梯度大小被加入直方图的对应块。即对Cell中所有(*x*, y)的方向梯度特征向量*F*(*x*, y)求和。在实际应用中，一般采用9个通道的无向直方图统计(*p* = 9)，并且每一个Block包含2 × 2 = 4个Cell。

Block归一化及全局扫描。Block中4个Cell统计得到直方图存在尺度不一致的情况，所以需要根据每个Cell各自的权重进行归一化，常用的归一化手段有：*L*1−*norm*，*L*2− *norm*，*L*1 − *sqrt*和*L*2 − *H*y*s*。由于局部光照的变化以及前景-背景对比度的变化，使得梯度强度的变化范围非常大。归一化能够进一步地对光照、阴影和边缘进行压缩。通过将各个细胞单元组合成大的、空间上连通的区间（Blocks），一个Block内所有Cell的特征向量串联起来便得到该Block的HOG特征。这些区间是互有重叠的，这就意味着：每一个单元格的特征会以不同的结果多次出现在最后的特征向量中。我们将归一化之后的块描述符（向量）就称之为HOG描述符。区间有两个主要的几何形状——矩形区间（R-HOG）和环形区间（C-HOG）。R-HOG区间大体上是一些方形的格子，它可以有三个参数来表征：每个区间中细胞单元的数目、每个细胞单元中像素点的数目、每个细胞的直方图通道数目。那么最终的特征维数计算如下：最终的HOG特征，实际上是由很多个以Block为计算单位得到的特征向量串联而成，Block以一个Cell的大小为跨度在Cell-Map上全局扫描，得到的特征向量加入末尾，如图2.9所示。

样本图像分割为若干个像素的单元（Cell），把梯度方向平均划分为9个区间（bin），在每个单元里面对所有像素的梯度方向在各个方向区间进行直方图统计，得到一个9维的特征向量，每相邻的4个单元构成一个块（Block），把一个块内的特征向量联起来得到36维的特征向量，用块对样本图像进行扫描，扫描步长为一个单元，最后将所有块的特征串联起来，就得到了人体的特征。例如，对于64 × 128的图像而言，每8 × 8的像素组成一个Cell，每2 × 2个Cell组成一个块，因为每个cell有9个特征，所以每个块内有4 × 9 = 36个特征，以8个像素为步长，那么水平方向将有7个扫描窗口，垂直方向将有15个扫描窗口。也就是说，64 × 128的图片，总共有36 × 7 × 15 = 3780个特征。

（2）FHOG特征[32,33]

FHOG由美国芝加哥大学的Felzenszwalb等人[32]提出，是原始HOG特征的一个重要变形，它被验证在目标检测任务中某些类别的检测效果比原始HOG更好，并在目标检测，特别是行人、人脸检测当中得到广泛应用。FHOG和HOG最大的区别在于前



###### 图 2.9 通过全局扫描得到HOG特征向量

Fig 2.9 HOG vector by Global Scan

者取消了Block的概念，对于一张输入大小为w×h的图片，直接进行Cell的切分，得到一张基于Cell的图*C*，假设Cell的宽高都为*k*个像素，则*C*(*i*, *j*)满足： 0*i* ⌊ − ⌋ (2.13) w 1 *k*

≤

≤

⌊

##### *h* − 1⌋

≤

0*j* ≤(2.14) *k*

要将原始图像中的每个点(*x*, y)的方向梯度特征计算后相加到相应的*C*(*i*, *j*)中，具体的步骤和原始HOG梯度计算一致，其中(*i*, *j*)和(*x*, y)服从如下的关系： ⌊*x*⌋

1. = (2.15) *k* ⌊y⌋
2. = (2.16) *k*

在FHOG中，每个*C*(*i*, *j*)对应四种不同的归一化因子，每个归一化因子记做*N*δ,γ(*i*, *j*)，且δ, γ ∈ {−1, 1}，则有：

*N*δ,γ(*i*, *j*) = √||*C*(*i*, *j*)||2 + ||*C*(*i* + δ, *j*)||2 + ||*C*(*i*, *j* + γ)||2 + ||*C*(*i* + δ, *j* + γ)||2 (2.17) 每一个归一化因子的含义是*C*(*i*, *j*)所对应的Block的“梯度能量”，因为一个Cell对应四个Block，所以也对应四个“梯度能量”（归一化因子），每个因子都是对包含(*i*, *j*)在内的4个Cell组成的块Block的梯度能量的度量。在FHOG中，除了Cell图*C*之外，增加了另外一个Cell图，记作图*D*。图*C*和图*D*的区别在于图*C*统计无向的直方图(Contrast Insensitive)，图*D*统计有向的直方图（Contrast Sensitive）。经过4个不同方向的归一化因子归一化之后，图*C*可以计算得到*H*1(*i*, *j*)，它是一个4 × 9的特征矩阵，图D可以计算得到*H*2(*i*, *j*)，它是一个4 × 18的特征矩阵，这里将*H*1(*i*, *j*)和*H*2(*i*, *j*)进行组合得到*H*(*i*, *j*):



|  |  |
| --- | --- |
| *H*(*i*, *j*) = [*H*1(*i*, *j*), *H*2(*i*, *j*)]  可见，H是4 × 27维的矩阵，令: | (2.18) |
| = {*u*1, ...,*u*9} ∪ {*u*10, ...,*u*27} ∪ {v1, ..., v4}  其中： | (2.19) |

###### 1, *i f j* = *k* 0, *other*w*ise*



*u*k (*i*, *j*) =(2.20)





###### 1, *i f i* = *k*

vk (*i*, *j*) =(2.21)

k k × 0, *other*w*ise*

如上述定义，*u* 和v 也都是4 27维的矩阵，FHOG则为*V*中所有的元素和*H*点乘求和的结果。*V*中有31个矩阵元素，所以得到的FHOG是一个31维的特征向量。它的前27维对应不同的方向通道（9个Contrast Insensitive和18个Contrast Sensitive），剩余4个通道则

是当前Cell在4个不同Block中的能量和。在PASCALVOC数据库上实验表明，用FHOG做特征提取的算法在很多应用中的效果比原始HOG特征效果更好。

### **2.2.4** 非手工设计特征

上述几种特征都是人类手工设计的特征，是人类智慧的结晶。深度学习是近年来人工智能领域取得的重要突破。深度学习的本质是通过多层非线性变换从大数据中自动学习不同抽象层次的特征，从而替代手工设计。深层的结构使深度学习具有极强的表达能力和学习能力，尤其擅长提取复杂的全局特征和上下文信息，而这是浅层模型难以做到的。传统的机器学习技术在处理未加工过的数据时，体现出来的能力是有限的。几十年来，想要构建一个模式识别系统或者机器学习系统，需要一个精致的引擎和相当专业的知识来设计一个特征提取器，把原始数据（如图像的像素值）转换成一个适当的内部特征表示或特征向量，子学习系统通常是一个分类器，对输入的样本进行检测或分类。特征表示学习是一套给机器灌入原始数据，然后能自动发现需要进行检测和分类的表达的方法。深度学习就是一种特征学习方法，把原始数据通过一些简单的但是非线性的模型转变成为更高层次的，更加抽象的表达。通过足够多的转换的组合，非常复杂的函数也可以被学习。对于分类任务，高层次的表达能够强化输入数据的区分能力，同时削弱不相关因素。比如，一张图像的原始格式是一个像素数组，那么在第一层上的学习特征表达通常指的是在图像的特定位置和方向上有没有边的存在。第二层通常会根据那些边的某些排放而来检测图案，这时候会忽略掉一些边上的一些小的干扰。第三层或许会把那些图案进行组合，从而使其对应于目标的某部分。随后的一些层会将这些部分再组合，从而构成待检测目标。深度学习的核心方面是，上述各层的特征都不是利用人工工程来设计的，而是使用一种通用的学习过程从数据中学到的。

深度学习正在取得重大进展，解决了人工智能界手工设计的特征所带来的局限性的问题。它已经被证明，它能够擅长发现高维数据中的复杂结构，因此它能够被应用于科学、商业和政府等领域。除了在图像识别、语音识别等领域打破了纪录外，它还在另外的领域击败了其他机器学习技术，包括预测潜在的药物分子的活性、分析粒子加速器数据、重建大脑回路、预测在非编码DNA突变对基因表达和疾病的影响。也许更令人惊讶的是，深度学习在自然语言理解的各项任务中产生了非常可喜的成果，特别是主题分类、情感分析、自动问答和语言翻译。我们认为，在不久的将来，深度学习将会取得更多的成功，因为它是基于原始数据端到端的映射学习，需要很少的手工设计，并很容易受益于可用计算能力和数据量的增加。目前正在为深度神经网络开发的新的学习算法和架构更会加速这一进程。

## **2.3** 基于类**Haar**特征的**AdaBoost**人脸检测

### **2.3.1** 基于**Adaboost**学习算法

人脸检测技术的突破发生在2001年，两位杰出的科研工作者Paul Viola和Michael

Jones设计了出了一个快速而准确的人脸检测器：在获得相同甚至更好准确度的同时，速度提升了几十上百倍——在当时的硬件条件下达到了每秒处理15张图像的速度，已经接近实时速度25fps（即25帧每秒）。这不仅是人脸检测技术发展的一个里程碑，也标志着计算机视觉领域的研究成果开始具备投入实际应用的能力。为了纪念这一工作，人们将这个人脸检测器用两位科研工作者的名字命名，称之为Viola-Jones人脸检测器，或者简称为VJ人脸检测器。基于AdaBoost的学习算法有两个目的，一是从大量类Haar特征中选择出最优的一部分特征，二是将挑选出的特征用于训练有效的分类器。由于并不是每一个类Haar特征都能很好地表现人脸灰度分布特点，因此如何挑选出少量的最优的特征来组成有效的分类器成为我们要解决的主要问题。

每一个分类器对应一个矩形特征，通过将一个复杂的分类器拆分成简化的弱分类器，然后将弱分类器集成层叠组合形成有效的强分类器，从而将候选图像子窗口的人脸和非人脸区分开。具体的训练步骤如下所示：

* 给定样本图像(*x*1, y1), ..., (*x*n, yn)，其中yi = 0, 1分别代表负样本和正样本；
* 初始化权重： w1,i =2ml ii (2.22)





1

y = 0

1

2 y = 1

其中*m*表示负样本的数目，*l*表示正样本的数目；

* 对于*t* = 1, ...,*T*：
  1. 归一化权重：

wt,i

wt,i ← t (2.23) w ,j 其中wt是概率分布；

* 1. 对于每个特征j训练分类器*h*j，每个分类器有且只对应一种特征，分类器错

|  |  |
| --- | --- |
| 误率是 |  |
| εj = ∑wi|*h*j(*x*i) − yi|  i   1. 选择错误率εt最小的分类器*h*t； 2. 更新权重： | (2.24) |
| wt+1,i =wt,i βt1−ei | (2.25) |

其中，βt = 1−εεt t 如果样本*x*i 被正确分类*e*i = 0，反之，*e*i = 1；

* 最后得到强分类器为：

*h*(*x*) =  t (2.26)

0 Ùę 其中αt = lg β1t

### **2.3.2** 级联分类器

基于AbaBoost级联分类器如图2.10所示。层叠的分类器每一层都是由若干弱分类器组成的强分类器，首先采用较少的特征放在前级分类器，即弱分类器较少，以排除大部分非人脸区域，能够通过几乎全部的人脸，后级分类器采用更多的特征，即强分类器由更多的弱分类器组成，检测上一层通过的子窗口，加强检测人脸的区域。随着层数的增加，弱分类器的数量也在增加，同时从上一层过滤下来的子窗口数量在减少，从而能够保证较高的检测速度。这里的待测图像，一般将其按比例缩放进行归一化，每个图像的尺寸为24 × 24像素的子窗口，依次判别是人脸还是非人脸。

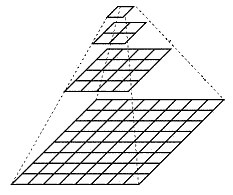
图 2.10 基于AdaBoost级联分类器结构

Fig 2.10 Structure of AdaBoost-based cascade classifiers

## **2.4** 基于级联可变形部件模型的人脸检测

### **2.4.1** 特征金字塔的构建

在自然环境下，目标图像可能具有不同的尺寸，这就要求滤波器能够在不同的分辨率下对图像进行扫描检测，从而得到最完备的检测结果，如图2.11所示。



###### 图 2.11 图像金字塔示例

Fig 2.11 Illustration of Image Pyramid

因此，在特征提取过程中，采用了特征金字塔的形式。对待检测图片进行尺寸的变换，构造图像的特征金字塔，然后到于每一层的图像使用检测器进行检测。通过对图像金字塔的分层检测，可以保证检测结果不会因为人脸的尺寸不同而产生错误，也可避免人脸大小不同导致的漏检。由于局部图像的边缘和光照变化能够使用梯度方向直方图来描述，这样，特征金字塔的结构就可以对各尺度下的特征结构进行编码。也就是说一幅图像的 HOG 特征是由多层的特征金字塔组成的。特征金字塔的使用保证了可变形部件模型对多角度目标的检测。上一章所述FHOG特征能够很好的勾勒出物体对象的轮廓，对微小形变和光照变化有很好的鲁棒性，不仅适合所涉及到的非生物目标的特征表示，也很适合来表示人脸。如下图2.12所示：

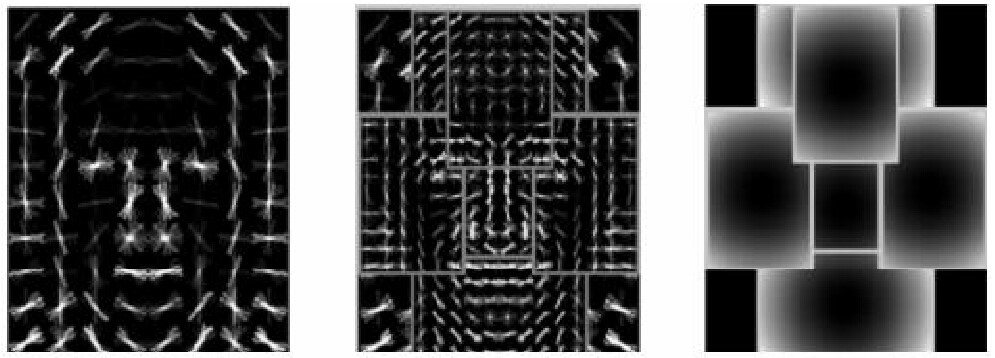


图 2.12 以FHOG 特征表示的人脸图像模型

Fig 2.12 Face model represented by FHOG descriptor

### **2.4.2 DPM**人脸检测模型

可变形部件模型（Deformable Part Model, DPM）是由Felzenszwalb[33]等人提出的基于滑动窗口中各部件以及各部件之间关系的检测模型。该方法对于通过可变形部件进行建模的配置提供了一个优良的框架。可变形部件模型基于Dalal-Triggs检测器，使用HOG特征和HOG多尺度特征金字塔，引入了可变形部件，提高了模型对于物体类内多样性变化的适用性。DPM模型使用多尺度HOG特征金字塔来描述物体的特征，在各组金字塔的每一层中提取特征，也就是提取了多尺度的待检测物体特征，这样就适应了待检测目标在一定程度上的尺度伸缩变化。可变形部件模型除了包括一个粗粒度的模板来覆盖整个对象，还包括高分辨率的部件模板来对各部件进行匹配。在训练过程中，将框定待检测对象的图片作为正样本，将每一个对象的各部件的位置作为一个潜在变量。可变形部件模型(Deformable Part Model)由三部分组成：

* (1) 一个较为粗糙的，覆盖整个目标的全局根模板(或叫做根滤波器)。
* (2) 几个高分辨率的部件模板(或叫做部件滤波器)。
* (3) 部件模板相对于根模板的空间位置。

首先要计算一个HOG金字塔：通过计算标准图像金字塔中每层图像的特征得到特征金字塔，HOG金字塔中每一层的最小单位是细胞单元(cell)。滤波器(模板)就是一个权重向量，一个w × *h*大小的滤波器*F*是一个含w × *h* × 9 × 4个权重的向量(9 × 4是一个HOG细胞单元的特征向量的维数)。所谓滤波器的得分就是此权重向量与HOG金字塔中w × *h*大小子窗口的HOG特征向量的点积(DotProduct)。而检测窗口的得分是根滤波器的分数加上各个部件的分数的总和，每个部件的分数是此部件的各个空间位置得分的最大值，每个部件的空间位置得分是部件在该子窗口上滤波器的得分减去变形花费。

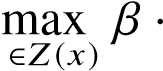
假设H是HOG金字塔，*p* = (*x*, y, *l*) 表示金字塔第一层 (*x*, y) 位置的一个细胞单元。

φ(*H*, *p*, w, *h*)是将金字塔*H*中以*p*为左上角点的w × *h*大小子窗口的HOG特征串接起来得到的向量。所以，滤波器*F*在此检测窗口上的得分为：*F*φ(H,p,w,h)。此后，在不引起歧义的情况下，我们使用φ(*H*, *p*)代表φ(*H*, *p*, w, *h*)。所以，含*n*个部件的模型可以通过根滤波器*F*0和一系列部件模型(*P*1, ..., *P*n)来定义，其中*P*i = (*F*i, vi, *s*i, *a*i, *b*i)。*F*i是第*i*个部件的滤波器；vi和*s*i都是二维向量，都以细胞单元为单位，vi指明第*i*个部件位置的矩形中心点相对于根位置的坐标，*s*i是此矩形的大小，*a*i和*b*i也都是二维向量，指明一个二次函数的参数，此二次函数用来对第*i*个部件的每个可能位置进行评分。

这里，使用β表示模型的参数，*F*0表示根滤波器的参数，*F*1, ..., *F*7为各部件的参数， *d*1, ..., *d*n为各部件的形变值，*b*为偏移值，则模型的参数向量可以表示为：

β = (*F*0, *F*1, ...*F*n, *d*1, ...*d*n, *b*) (2.27)

在进行检测时，使用了滑动窗口的方法，检测对象模型包括一个全局的根滤波器和很多部件模型。每一个部件模型都制定了一个具体的空间模型和部件滤波器。空间模型定义了对于一个部件相对于检测窗口允许变化的范围空间以及在该位置的变形代价。待检测窗口的得分是根滤波器的得分加上各部分最大重合部分的得分减去变形代价。若检测窗口的得分高于阈值，则说明检测窗口中包含待检测物体，否则不包含。令φ(*x*, y)为检测窗口对应的图像特征，*x*表示检测窗口的位置，*z*表示各部件的位置，那么该图像的最终得分如下:

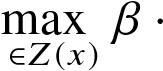
*score*(*x*) =φ(*x*, *z*) (2.28) z

其中，参数β由训练得出。待检测窗口的得分是根滤波器的得分加上各部分最大重合部分的得分减去变形代价，这与传统的基于部件的模型相似。根滤波器和部件滤波器都是由计算HOG特征与权值之间的点积得到。跟滤波器等同于Dalal-Triggs模型，计算各部件的特征的分辨率是根滤波器的两倍。本文提到的模型是固定规模的，通过在图像金字塔上检测对象，从而达到对不同分辨率进行检测的效果。高得分的根位置定义了一次检测，产生高得分根位置的部件位置定义了一个完整的目标假设。通过定义每个根位置的综合得分(overall score)，可以检测目标的多个实例(假设每个根位置上最多一个实例)。这种方法与滑动窗口检测器有关联，因为可以认为*score*(*p*0) 是检测窗口在指定根位置的得分。通过使用动态规划和广义距离变换(min-convolution)来计算部件的最优位置(是根位置的函数)，此方法非常高效，花费*O*(*nk*)时间计算一个滤波器的响应值，*n*是模型中部件的个数，*k*是特征金字塔中位置的总数，文献 [34]和文献[35]中有详细的解释。

### **2.4.3 Structured SVM**模型训练

（1）隐藏变量SVM——Latent SVM

在模型的训练过程中，引入了专门为弱监督设计的隐支持向量机(Latent Support Vector Machine, LSVM)。考虑一个分类器，在 LSVM 中每个样本*x*的得分可以使用下面的形式表示：

*f*β(*x*) =φ(*x*, *z*) (2.29) z

这里β是模型参数向量，*z*是隐藏变量。集合*Z*(*x*)定义了样本*x*所有可能的隐藏变量值。通过对此得分值进行阈值化，可以获得样本*x*的二分类类标。类比经典SVM算法，我们使用带类标的样本集*D* = (< *x*1, y1 >, ..., < *xn*, y*n* >), y*i* ∈ −1, 1来训练参数β，优化目标函数为：

*L*D(β) = *a* || β||2 + *C* ∑max(0, 1 − yi *f*β(*x*i)) (2.30) n

*b*

i=1

其中max(0, 1 − yi *f*β(*x*i))是标准铰链损失函数（hinge loss），常数*C*控制正则项的相对权重。如果每个样本的隐藏变量有唯一可能值|*Z*(*x*i)| = 1，则*f*β是β的线性函数，此时变为线性SVM问题，这是LSVM的一个特例。

1. 半凸规划

LSVM最终是一个非凸规划(non-convex)问题。然而，下面所讨论的情况下LSVM是半凸规划(semi-convexity)问题，一旦将隐藏信息指定给正样本则训练问题变为凸规划问题。几个凸函数的最大值问题是凸规划问题。在线性SVM中，有:

*f*β(*x*) = β · φ(*x*) (2.31)

*f*是β的线性函数，此时铰链损失函数对每个样本都是凸的，因为它是两个凸函数的最大值。注意到公式(2.29)中定义的*f*β是一系列函数的最大值，而这些函数都是β的线性函数，因此*f*β是β的凸函数。所以当yi = −1时，两个函数*f* (*x*) = 0和*f* (*x*) = 1 − yi *f*β(*x*i)都是β的凸函数，所以铰链损失函数max(0, 1 − yi *f*β(*x*i))是β的凸函数。也就是说，只有当样本为负样本时，损失函数是β的凸函数。我们将损失函数的这一性质叫做半凸(semi-convexity)。对于正样本(y*i* = 1)来说，LSVM的铰链损失函数不是凸函数，因为它是一个凸函数*f* (*x*) = 0和一个凹函数*f* (*x*) = 1 − yi *f*β(*x*i)的最大值。但是，当LSVM的正样本的隐藏变量具有唯一可能的取值时，*f*β(*x*i)是β的线性函数，因此损失函数是β的凸函数，再加上半凸性质，公式(2.30)变为凸规划问题。

1. 模型训练步骤

模型的训练过程主要包括以下几个步骤：

1. 初始化根滤波器。初始化根滤波器的时候首先根据训练图片的标记确定根滤波器的规模，然后使用支持向量机(SVM)来训练根滤波器*F*0，在这个过程中没有使用到隐变量。
2. 更新根滤波器。这个过程中，对于训练集中的每一个样本使用上一步确定的根滤波器寻找响应最大并且和原有标记有明显覆盖的位置，并以此位置更新训练图片的框标记，然后使用新位置标记的样本以及随机选取的负样本来更新根滤波器*F*0。
3. 初始化部件滤波器。在确定的根滤波器的位置上，贪婪的选择得分最大的区域作为部件滤波器的位置。
4. 更新滤波器。使用不断更新的正负训练样本来训练、更新滤波器。在这里使用的正负训练样本具有如下的特点：正样本选择在训练中的高分区域占到标记的50%以上的样本，负样本选择在旧滤波器的检测中得分很高的负样本。使用上述样本来训练新的模型，并不断地更新训练样本，如此多次迭代(10 次)得到最终的模型。

以上是使用隐支持向量机训练可变形部件测的过程，通过不断地迭代以及样本更新，最终可以得到我们需要的可变形部件模型。

## **2.5** 本章小结

本章主要介绍了人脸检测的研究概述和本文所采用的人脸检测的方法，另外介绍了在人脸检测和识别当中用到的主流的特征。人脸检测是非常重要的预处理部分，其不仅考虑到多视角下的检测率和误检率，而且要保证检测的实时性。基于Haar特征的Adaboost人脸检测器因其特征简单，所以速度很快，但是检测率低，误检率却很高，且在处理多视角人脸中显得不足，而基于DPM模型的人脸检测因其本身模型的适应性以及FHOG特征金字塔模型的有效性，使得多姿态人脸检测能够保证很高的检测率和很低的误检率，但是速度较Haar-Adaboost模型较慢，这部分可通过工程上的实现，比如并行加速、指令集优化来弥补。详情在实验结果与分析部分会给出介绍。

# 第三章 复杂姿态下人脸特征点定位

## **3.1** 人脸特征点定位研究概述

人脸有复杂的表面结构，因而人脸的描述或刻画相对简单的物体来说较为复杂。同时面部肌肉的运动使得人脸成为一种非刚体目标，与刚体目标识别相比，非刚体目标的检测和定位更加困难。此外，由于光照、姿态、饰物、遮挡以及采集和成像条件的变化，使得面部特征点的局部外观和人脸形状千差万别。对于计算机而言，要分辨这些变化和其间的差别是一件十分困难的事情，其主要面临着人脸姿态变化、遮挡、表情变化、光照变化等挑战，如下图3.1所示。



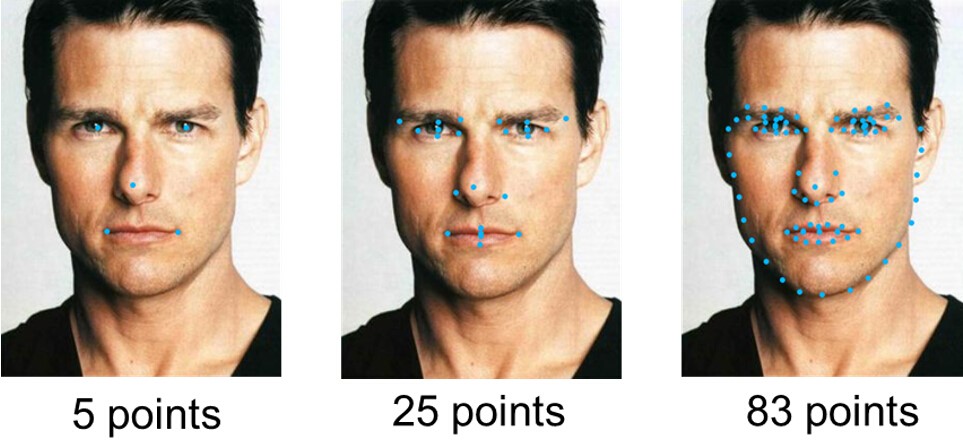
#### 图 3.1 人脸图像的复杂变化

Fig 3.1 The complex variance of face image

从图3.1中的人脸图像示例中可以看到要让计算机自动完成这项任务，研究人员面临巨大的挑战。因而，研究者必须设计各种算法和模型，并参照人类对这些模式的认知来指导计算机对人脸图像进行分析，从而完成特征点的定位和人脸形状的提取。

人脸特征点（Face landmark）定位是在人脸检测基础上，找到具有语义的人脸特征点，如眼角、眉端、鼻尖、嘴巴轮廓等，如图3.2所示。

人脸特征点对人脸识别、建模、跟踪、动画制作等是极为关键的步骤。对于一幅人脸图像，通过定位特征点可以方便地抽取人脸的各部分特征，若以这些特征点为基准对抽取的各个特征值进行归一化，则这些特征值具有平移、旋转和尺度上的不变性，通过这样对人脸进行规范化处理，可以将不同大小、方向和水平旋转等情况的人脸统一起来，扩大人脸库的入库范围，提高人脸识别的速度。从某种意义上讲，面部关键特征点的定位，亦即人脸图像像素之间高层语义的对齐，是人脸识别最核心的问题



#### 图 3.2 人脸特征点图示

Fig 3.2 Illustrations of Face Landmarks

之一。错误的特征定位会导致提取的人脸描述特征的严重变形，不精确的对齐也会带来识别性能的快速下降，所以准确的人脸特征点定位对于后续的人脸识别有很大的影响。

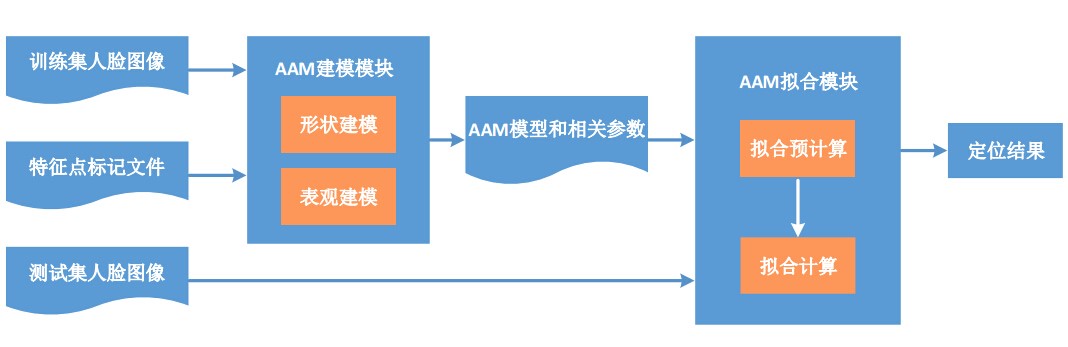
在早期的研究方法中，进行人脸面部特征定位的方法主要有主动形状模型（Active

Shape Models, ASM）[36]以及主动外观模型（Active Appearance Models, AAM）[37,38]。现有的主流方法大致分为两类，一种是基于参数模型的方法，另外一种是基于回归的方法。参数形状模型（Parametric Shape Model）存在着如下缺点，一是参数误差和对齐误差不一定对等，通过最小化参数化误差得到的结果是次优的。此外，参数化模型通常很难具体估计模型的大小，通常是启发式的，或者使用固定数量的参数，因此在迭代对齐中不能灵活应变。基于参数模型的方法以Cootes在1995年提出的ASM方法为代表，ASM方法将多个脸部特征点的纹理和位置关系作为约束条件，计算出一个参数模型。从局部特征中检测到所求的关键点，但是这种方法对噪声敏感。AAM则易受初始化值和外观变化的影响。在后续的工作中，研究者们虽然对ASM和AAM做出了改进，但是这些方法普遍存在泛化能力差、训练耗时严重等现象，对于真实场景下的人脸特征点定位很难取得很好的效果。另外一种流行的方法是基于局部部分可变模型的方法（Deformable Part Models，DPM）[39]，该方法虽然在物体检测方面表现良好，但是对于准确度较高的人脸标记估计，其表现一般。

现有方法中取得比较好的效果的是P. Dollár等人提出的级联姿态回归（Cascaded Pose Regression，CPR）[40]方法，这种方法在估计人脸标记当中准确率和效率都有很大提高，并成为一种主流方法，然而CPR方法仍不能很好的解决遮挡和形变过大的问题。基于此，研究者提出了一些改进的算法，如在2012年CVPR中，Cao等人受形状回归模型（Shape Regression Model, SRM）[41]的启发提出了Explicit Shape Regression （ESR）模型，该方法直接通过回归估计形状，克服高维输出等挑战，且保留了形状约束。在2013年的ICCV中，Burgos等人提出了鲁棒的级联位姿回归（Robust Cascaded Pose Regression, RCPR）[42]，专注于解决人脸特征点定位过程中的遮挡和大的姿态形变问题。RCPR方法与ESR方法相比，RCPR当预检测失败的时，添加了一个智能重启方法。基于这些方法，并针对监控环境下人脸姿态变化及遮挡的问题，本文研究并寻找一个对于遮挡和形变的鲁棒人脸特征点定位方法，方法要求能够精确、快速、鲁棒地估计出人脸特征点。为了跟传统的AAM方法比较，本章首先介绍在人脸特征点定位经典的AAM方法，然后再介绍基于级联回归模型的多角度人脸特征点定位方法，并从方法和实验上进行比较。

## **3.2** 基于主动外观模型的人脸特征点定位

人脸特征点定位的目标是能够对人脸进行全自动精确定位。主动形状模型 (Active shape modal, ASM) 和主动外观模型 (Active appearance modal, AAM) 的提出为全自动人脸特征点定位工作提供了很好的思路和解决框架。之后很多研究工作也都在 ASM 和 AAM 的框架下进行了改进。AAM是近年来在特征点定位领域得到广泛应用的一种快速有效的方法。基于AAM的人脸特征点定位方法不但考虑了局部特征信息，而且考虑了全局的形状和纹理信息，通过对人脸的形状特征和纹理特征进行统计分析，建立对应的AAM模型，对被测人脸对象进行特征点定位时，结合“合成→比较→调整再合成→再比较”这样一个图像拟合(image fitting)策略，往往能够实现快速而准确的人脸特征点定位。图3.3表示了AAM特征点定位的基本框架。



#### 图 3.3 AAM 特征点定位框架

Fig 3.3 AAM framework

AAM模型是对象的动态表观模型，表观是形状和纹理的组合，表观模型是在形状模型的基础上结合对象的纹理而建立的，AAM模型实例就是将AAM的表观模型通过仿射变换的形式映射到对应的形状实例中去，得到描述当前对象的当前模型。动态就体现在AAM通过相应的拟合算法不断调整生成新的AAM模型实例与待定位的对象进行匹配，直到生成的模型实例能和该对象真正吻合，可见生成AAM模型实例是AAM中比较重要的一个部分。

AAM模型也是一种基于统计信息建模来实现对目标图像的搜索与解释的一种方法。在主动表观模型中，不仅对形状进行建模，同时又建立了反映目标物体灰度变化的纹理模型。并将纹理模型与形状模型合理地结合起来，建立起了反应整个目标图像形状及纹理共同合理变化的统计表观模型。该方法采用了基于优化算法的搜索机制，使合成的模型与目标图像不断接近，最终得到能够反应目标图像纹理及形状的合成模型。本节通过形状建模、表观建模和模型生成中拟合计算三个部分来介绍AAM模型的建立过程。

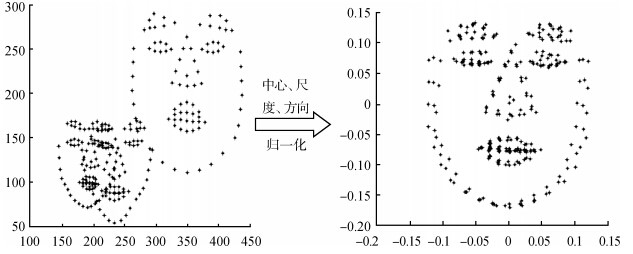
### **3.2.1** 形状建模

统计模型的建立需要分三个步骤，首先是获取样本图像中的信息；其次是样本图像集的归一化处理；然后是对归一化处理的数据进行统计分析，建立统计模型。因此不论AAM对形状还是对表观建模都是遵循统计模型建立的基本步骤。 AAM形状建模实现步骤如下：

1. 选择一些适合的学习样本，在训练集中的每一个训练样本上手工标定*k*个特征点(landmark)，这*k*个特征点代表了人脸上一些比较显著的特征点，那么人脸形状可由人脸的*k*个标定点组成的向量表示为：

Si = [*x*i0, yi0, *x*i1, yi1, · · · , *x*ik, yik]T (3.1)

1. 对训练集合中的每个形状向量，进行中心、尺度和方向的归一化。归一化是指把所有用于学习的人脸形状去除旋转、缩放和平移等全局变换，如图3.4所示。



##### 图 3.4 人脸形状归一化示例

Fig 3.4 Normalization of facial features

为了对齐所有的训练样本，一般采用 Procrustes方法[43]。该方法实际上是一种几何学的方法，它把一系列的点分布模型通过适当的平移、旋转、放缩变换，在不改变点分布模型的形状的基础之上对齐到同一个点分布模型，从而改变获取的原始数据的杂乱无章的状态，减少“非形状”因素对人脸点分布模型数据分析的影响，为人脸特征提取做好准备。其基本思想是最小化所有形状到平均形状的距离和，即：

*D* = ∑ |*S*i − *S*¯|2 (3.2)

其迭代过程为：

* 对每个样本*x*i做旋转、缩放并与样本*x*1对齐，因此得到*x*¯；
* 重复计算变换后的形状平均值*x*¯，将所有样本旋转、缩放、并与平均形状对齐；
* 如果平均形状*x*¯变化不大，即收敛，则算法停止。否则，算法继续重复。

1. 对归一化的形状进行主成分分析(PCA)变换，得到对应训练集的平均形状*S*0和前*n*个特征值对应的形状特征向量*S*i；
2. 任意人脸形状*S*就可以用线性方程进行表达：

*S* = *S*0 +∑ *p*i*S*i (3.3) n

i=1

在公式(3.3)中，*S*0为平均形状，*p*i为形状参数，反映了任意形状在平均形状基础上的形变。*S*i为保留的主成分特征向量所构成的变换矩阵，可通过计算协方差矩阵的特征值得到。*S*i反映了形状的主要模式。这样就完成了对形状的建模。

### **3.2.2** 纹理建模

AAM的纹理是训练集中已经标记好的形状区域内的对象纹理信息，它是对象形状和纹理的组合。获取表观的方法是，利用计算机图形学中纹理映射的方法，对每个表观样本通过已经标记好的形状来获得其纹理的有效区域，建立一个可逆的映射方程式，然后将这个表观区域映射到一个己设定的基准形状网格，在该形状网格内进行一致的参数采样，再将它们映射回各自的纹理区域，这样就可以获得归一化的纹理样本，再对归一化后的纹理进行PCA即得到了所需要的表观模型的参数，实现了纹理建模。在获得纹理向量前，需要得到纹理的标准模板。将人脸区域划分为多个三角形区域，通过三角形仿射将图像样本的纹理信息投影到标准模板。以*S*0 作为纹理模板的基础进行适当的调整后得到模板，采用三角分段仿射的方法将各种姿态，各种表情的人脸图片映射到标准的人脸，得到标准纹理图，如图3.5所示。总结来看，对标准的建模要以下几个步骤：

（1）将*S*0和训练集中的人脸形状，分别Delaunay三角化；



##### 图 3.5 分段线性仿射效果图

Fig 3.5 Illustration of piecewise linear affine

1. 通过分段线性仿射将样本集人脸形状中的纹理信息映射到平均形状*S*0中去，实现对纹理归一化；
2. 对归一化后的纹理信息进行PCA变换，得到平均纹理*A*0和前*m*个特征值对应的纹理特征向量*A*i；
3. 纹理与形状非常相似，任意人脸的纹理信息也可以用线性表达式表示：

*A*(*x*) = *A*0(*x*) +∑ λi *A*i(*x*) ∀*x* ∈ *S*0 (3.4) n

i=1 按上述步骤操作即可完成对纹理的建模。

### **3.2.3 AAM**拟合计算

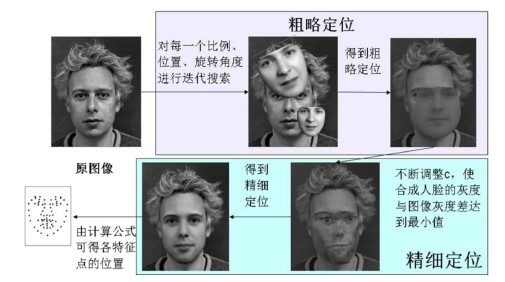
AAM分为模型建立和模型实例生成这两部分内容，而AAM模型实例的生成需要如下步骤：先得到任意一组形状参数*p*后，用形状模型进行线性表示，就能够得到一个对应的形状*S*，接着得到一组纹理参数λ后，用纹理模型进行线性表示，得到一个对应纹理实例A(x)。最后将平均形状*S*0中的纹理信息*A*(*x*)映射到当前的形状*S*中去，这样就生成了一个AAM的模型实例。

当模型建立后，要实现当前AAM模型实例与待定对象的匹配就离不开AAM拟合计算。因此AAM拟合计算是AAM的核心部分，早期的AAM拟合算法是由T.F.Cootes等人[37,44]提出，都属于标准梯度递减最优化算法。这些早期的AAM拟合算法的优势是它们用来进行最优化收敛的理论和思想都容易让人理解，但它们总是简单的假设在误差图像与参数的增加量上有一种固定的线性关系，而这种假设往往是不成立的，所以这样的拟合算法往往会引起拟合的失败，而且这些算法在迭代过程中计算量相当大，每次变量更新后都需要对 Hessian、梯度等根据新变量进行重新计算，因此它们已经不能够满足快速准确地进行拟合计算的需要。2001 年卡耐基梅隆大学的Simon Baker等人[2]提出了反向组合算法，它是一种有效的图像对齐算法，之后又扩展原反向组合算法，并且将它引入到AAM拟合计算中，提出反向组合的AAM拟合算法。反向组合算法由Lucas-Kanade[45]算法演化而来，利用反向组合进行AAM拟合计算，使得在AAM拟合计算过程中的 Hessian 和梯度都能够提前在预处理中进行计算，并且避免了原AAM拟合计算中需要对这些值在迭代过程中重复计算的问题，从而大大减少了拟合的时间。主动表观模型AAM中主动体现在AAM拟合计算中，形状建模可用PCA方法来描述形状控制点的运动变化，形状控制点表征了特征点的位置，再以模型实例与输入图像差的平方和来定义一个能量函数。具体拟合过程如下：

定义(*x*, y)T = *W*(*x*, *p*)是标准模板中像素点*x*在形状特征系数为*p*时映射到图像*I*中的坐标，从图像*I*中坐标*W*(*x*, *p*)处采得纹理值*I*(*W*(*x*, *p*))。当形状特征系数为*p*时，标准模板中像素点*x*在图像*I*上采得的像素值为*I*(*W*(*x*, *p*))。

*E*(*X*; *p*) = ∑[*A*0(*x*) − *I*(*W*(*x*, *p*))]2 (3.5) x

利用该能量函数来评价拟合程度，在定位拟合的过程中，根据模型的线性表达式，通过调整形状参数*p*，以及纹理参数λ的值，利用有效的拟合算法变化模型参数组，从而控制形状控制点的位置变化生成当前AAM模型实例，得到当前能量函数的值，再更新模型的参数。如此迭代反复以实现能量函数的最小化，达到模型实例与输入图像相拟合的目的，而最终得到的形状控制点的位置就描述了当前对象的特征点位置。具体的拟合过程如下图所示：



##### 图 3.6 人脸拟合技术示例

Fig 3.6 Illustrations of face fitting by AAM

基于AAM的人脸特征提取方法不但考虑了局部特征信息，而且考虑了全局的形状和纹理信息，在复杂的环境中，可以实现快速而准确的脸部特征点定位。基于 AAM 的人脸特征点定位方法充分利用了训练人脸样本和目标人脸的形状和纹理信息，其定位能力较强，具有很高的准确性，然而，该方法也存在一定的局限性。首先，基于 AAM 的人脸特征点定位方法是一种局部最优化的方法，因此模型的初始位置对迭代结果有一定的影响，如果初始位置不能满足局部最优的条件，会导致定位算法难以收敛，从而使得定位结果出现很大的偏差，导致定位失败。另外一个很重要的缺陷是在定位过程中，实时得到的 AAM 人脸形状会出现非法变形的情况，即出现不符合自然中人脸的形状规则。当出现非法变形时，AAM 没有考虑如何实时的解决这样的问题，仍然认为是正常的拟合并行，在不进行修正的情况下继续人脸特征点定位，最终导致人脸特征点定位的失败。这些问题，都需要得到解决。

## **3.3** 基于级联回归模型的多角度人脸特征点定位

近年来，有不少其它新颖的人脸特征点定位方法被提出，并取得了比AAM更好的效果。其中包括基于回归的方法[41]和基于深度神经网络的方法[46]等。在这些方法中，基于级联的姿态回归( Cascaded Pose Regression，CPR)方法[40]在特征点邻域的纹理特征与特征点位置的调整量之间建立回归关系，从而能够从特征点初始位置出发经过多次迭代回归和调整后收敛于特征点的正确位置。显性形状回归的人脸对齐方法使用矢量回归函数来完成人脸面部形状推断并且显式的最小化训练集数据对齐过程中产生的错误。CPR方法是将形状估计视为一个回归的问题来解决。利用训练所得的回归器就可以直接预测物体的形状。CPR方法的关键在于计算形状索引特征和训练回归器。文献[40, 41]运用了深度为5的随机蕨回归器，并且采用形状索引来刻画特征点的特征。基于级联的姿态回归方法易于实施，且检测效果也明显优于传统的基于ASM和AAM的方法。然而，所有这些方法都只考虑了正面或者接近正面的情形，并对遮挡其检测精度在人脸姿态角度大和有遮挡时明显下降。

针对上述问题，我们采用基于回归方法，并针对形变等问题做出了改进，并在准确率和速度上取得平衡，设计一种鲁棒的人脸特征点定位方法。级联回归量学习框架包含固有的形状约束条件，得到从粗到细的特征点位置确认。方法使用两级回归模型，包含了形状索引和基于相关性的两种特征选择方法。为了加快速度，通过采用像素级的形状索引特征和自适应的初始化策略。方法详情由以下小节分别介绍。

### **3.3.1** 级联回归模型

在人脸检测算法中，最具有开创性的工作是2001年由Viola和Jones[1]提出来的检测框架。他们为实用的人脸检测器确定了两个基本的准则：1、使用级联结构，在早期就拒绝掉一些简单的负样本。2、使用简单的特征，加快分类速度。现在几乎所有能达到实时的人脸检测器都遵循着这两个准则。受这个框架启发，本文采用了一种级联的回归模型和可快速计算的特征进行人脸的特征点定位。下面首先分析人脸特征点的级联回归模型。

定义一个人脸形状*S*i = [*x*i0, yi0, *x*i1, yi1, · · · , *x*ik, yik]T，其中*i*表示第*i*个样本，*x*, y为特征点的坐标位置，总共包含*k*个特征点。对于给定一个人脸图像，人脸对齐的目标是估计一个形状*S*使得与标定的真实形状*S*ˆ的误差最小，即：

min ||*S* − *S*ˆ|| (3.6)

通常，公式3.6所表示的对齐误差被用来引导训练和评估性能。在P. Dollár等人提出的级联姿态回归模型（Cascaded Pose Regression, CPR）方法[40]中，用基于回归的方法取代传统的基于优化的方法。他们提出一种级联回归的模型用于形状回归，其步骤如下：首先，基于boosted regression[47]的方法联合*T*个弱回归器，给定人脸图像*I*和初始形状*S*0 ，每个回归器通过图像特征计算一个形状增益δ*S* ，然后更新当前的形状，即：

*S*t = *S*t−1 + *R*t (*I*, *S*t−1),*t* = 1, ...,*T* (3.7)

给定*N*个训练样本(*I*i, *S*ˆi)}iN ，逐步学习各个回归器直到训练误差不再减小，回归

=1 器Rt通过公式3.8得到：

{

N

*R*t = arg *min*R ∑ *S*ˆi −  (3.8)

i=1

其中*S*tt−1是在上一阶段估计的形状。通过不断地更新和学习，最终得到人脸形状。原始级联回归模型如算法1所示。

通过利用全面的训练数据明确最小化对齐误差来训练回归器，即所有的面部标记联合回归。回归器以一种非参数化的方式理解形状的限制，回归的形状总是训练形状的一个线性组合，而且为所有标记使用全局特征比使用局部特征更有区分性。这些性能使我们能够学习到一个对于大量训练数据拥有较强表达能力的灵活的模型，这种方法被称为“显式形状回归”。对于整个形状采用联合回归的主要挑战在于图像外形的大幅变化。算法使用了一个增强的回归器来推断形状。即前期的回归器控制形状的大

**Algorithm 1:** 级联姿态回归算法

1. 输入：图像*I*，初始位置*S*0，回归器*R*1, *R*2, · · · , *R*T，特征计算函数*h*t (*S*t, *I*)
2. 输出：估计的形状*S*T
3. **for** *k* ← 1 **to** *T* **do**

**4**计算特征*x*t = *h*t (*S*t−1, *I*)

**5**计算形状增量δ*S* = *R*t (*x*t)

**6**更新形状估计*S*t = *S*t−1 + *R*t (*I*, *S*t−1)

**7 end**

##### **8** 返回：*S*t

幅变化并且保证健壮，然后后期的回归器控制细小的形状变化并且保证准确度。这样形状约束自动的由粗到细的执行，如下图3.7所示。

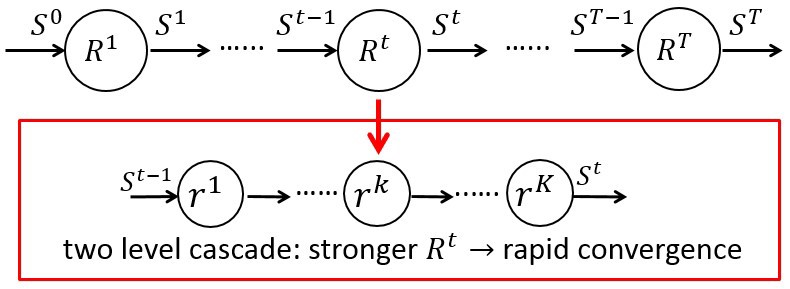


###### 图 3.7 显式形状回归示例

Fig 3.7 Illustrations of explicit shape regression

学习一个好的弱分类器至关重要，并且能够减小误差。boosted regression与以前的方法相似，使用简单的弱回归器如决策树。然而这些回归器太弱导致在训练过程中收敛速度很慢，且在测试过程中效果不佳。当存在大的外观变化和粗略的初始化的时候，这些回归器将很难回归出整个的形状。在这种情况下，一个简单的回归器能降低误差但不具备很好的泛化能力。因此本文采用两层的级联分类器，如图3.8所示，这样的二层级联方式能够克服收敛速度慢和泛化能力差的缺点。

算法采用两级形状回归方法计算形状回归量*R*t。在第一级回归当中，主要用于产生特征，并更新形状。在第二级回归中，主要进行特征选择，并计算形状增量。二级



###### 图 3.8 两层的级联回归

Fig 3.8 Two-level cascade regression model

回归是由若干个弱回归量*R*1, *R*2, · · · , *R*T组成，经过一级回归和二级回归的交叉迭代后产生一系列强回归量。一级回归中，采用形状特征索引方法获取特征，这种方式能准确定位特征位置，提取有效人脸特征。二级回归特征选择中，采用相关性特征选择方法，从特征池中选择与本次迭代相关性最大的几个特征作为本次迭代的人脸特征。这种方法能选择具有代表性的人脸特征，学习到较好的形状回归量。通过二级形状回归迭代后，输出对齐的人脸形状，该形状是由初始人脸形状和训练样本形状线性和表示，即：

N

*S* = *S*0 +∑ωi*S*ˆi (3.9)

i=1

其中，ω为回归系数，*S*ˆi为样本标准人脸形状。因为初始形状和样本形状都服从形状约束，所以输出的人脸形状也服从形状约束，这种线性表示方式能较好控制人脸形状。

为了解决复杂条件下人脸对齐的效率和准确率等问题，将人脸形状空间融合到级联回归框架中产生二级形状参数回归人脸对齐算法，以形状空间控制人脸形状，以低维形状参数进行形状迭代，较好解决对齐效率和对齐准确度问题。为使算法更高效，在二级形状参数回归中，给出结合旋转和尺度因素的明确形状特征索引方法和一种多重随机特征选择方法，能更准确定位和提取人脸特征，提高对齐精度。总体框架如图

3.9所示。

由图3.9可看出，在学习阶段，使用人脸形状空间将标准形状和初始形状转换为低维的人脸形状参数，然后通过二级形状参数回归进行迭代更新，最后输出对齐的人脸形状，并生成一系列形状参数回归量(*R*1, *R*2, · · · , *R*T )。在测试阶段，将初始人脸形状转换成人脸参数，根据学习好的参数回归量(*R*1, *R*2, · · · , *R*T )进行二级形状参数回归，不断迭代更新人脸形状，最终输出对齐的人脸形状。

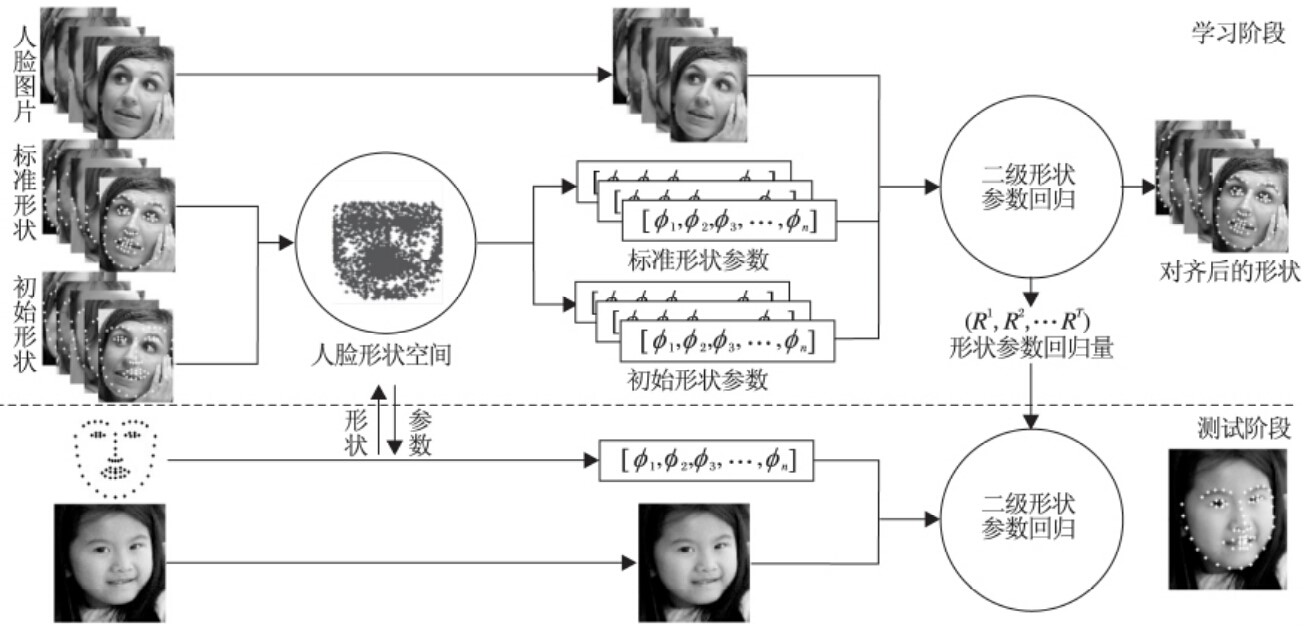


图 3.9 形状参数回归算法框架图

Fig 3.9 Framework of shape parameter regression algorithm

### **3.3.2** 形状索引特征

要实现快速和精确的特征点检测，离不开有效并可快速计算的特征。之前的一些解决人脸角度和表情变化的算法，例如：PEP[48]，DPM[33]等，一般由于计算量太大，限制了它们的应用范围。因此，在这章中，我们将介绍本文的重点之一：形状索引特征(Shape Index Feature)。这个特征有如下两个优点：

1. 形状索引特征的其中一个优势就是对齐的精度很高。这个特征对于人脸角度和表情的变化具有不变性。
2. 形状索引特征的另一个优势就是提取速度极快。这个特征提取过程简单，可以用在快速的人脸检测和特征点定位的有关算法中。形状索引特征配合像素灰度值特征使用，可以获得十分迅速的提取速度。我们甚至都不需要得到相似变换后的人脸图像就可以直接得到形状索引像素特征。

本文从以下两个方面对形状索引特征进行详细介绍：

（1）消除全局变换

为了消除图片中的因为人脸在图片中的大小、位置和左右歪头而产生的影响。这里我们采用的方法是使用相似变换(similarity transformation)将人脸的*p*个基准点的位置映射到固定的位置上。这里我们只用其中5个基准点说明。相似变换的具体做法如下：首先，我们假设图片中左右瞳孔、鼻尖、以及左右嘴角这五个基准点的位置分别为(*x*1, y1, · · · , *x*5, y5)，我们将这些坐标排列成如下形式:

##### \**x x*5

*P* = ....,y11 · · ·· · ·· · · y5+////- (3.10)

###### 1 1

假设我们需要将这5个映射到的固定的位置分别为*x*¯1, y¯1, · · · , *x*¯5, y¯5。这里需要注意的是，这些固定点的位置不可随意选取，如果选择不合适，会造成变换时的误差比较大，有时无法把人脸旋转到我们希望的正脸的效果。在我们的实验中，一般选择数据库上统计出来的这5个点的坐标的平均值。同样，我们将这5个固定位置也排列成如下形式:

*P*¯ =\*..,y¯11 · · · y¯5+////- (3.11) *x*¯ · · · *x*¯5

.

.

###### 1 · · · 1

接着，我们需要估计一个相似变换矩阵*T*，这个变化将*P*映射到*P*¯，即*P*¯ = *TP*。因为*T*是相似变换矩阵，所以*T*应该满足如下形式:

*T* =\*..,−*b a* δy+////- (3.12) *a b* δ*x*

.

.

###### 0 0 1

这里我们使用相似变换，而不是仿射变换，或者投影变换等是因为我们不希望扭曲人脸。因此我们希望变换中只包含旋转，平移和尺度变化。下面我们需要根据*P*和*P*¯去估计*T*中的4个参数 (*a*, *b*, δ*x*, δy)。这里我们使用最小二乘去估计，因为*P*¯ = *TP*，重新排列其中的元素的位置，可以得到如下表达式：

*Q*¯ =Θ*Q* (3.13)

其中：

Θ= (*a*, *b*, δ*x*, δy) (3.14)

*Q*¯ = *x*¯1, y¯1, · · · , *x*¯5, y¯5 (3.15)

##### \*.....*x*11 y11 · · · *x*55 y55//////-+// 1

.

,

*Q* =y −*x* · · · y −*x* (3.16)

.1 0 · · · 1 0

###### .0 1 · · · 0 1

这样可以利用最小二乘得到相似变换参数Θ= *QQ*¯ T (*QQ*T )− 。知道了相似变换矩阵*T*之后，我们就可以生成相似变换后的正脸图像。在实际操作中，一般釆用逆向变换的方式去产生相似变换后的图像。即：对于相似变换后的图像中的每个位置*x*¯, y¯，我们去寻找它在原图中的坐标(x,y)，其满足：

\*..,y+////- = *T* −1 .,\*...y¯///-+/ (3.17) *x x*¯

.

.

1 1

相似变换之后的人脸图像和人脸基准点的结果如图所示，可以看到，这个图像已经消除了原图像中因为左右歪头而产生的影响，得到了一个正面的人脸图像，并且所有的基准点都在正确的位置上。在人脸特征点检测的过程当中，通过每一个回归器得到预估计的人脸形状，然后通过逆向变换的方式得到该形状变换到原始图像的形状，在达到一定的错误率的阈值后即收敛，从而得到预测的人脸特征点。因为人脸具有一定的几何结构，各个面部器官的相对位置基本保持不变，因此，相似变换之后的基准点一般会排列成一个相对固定的形状，我们称这种在经过相似变换之后的基准点周围提取的特征叫做形状索引特征(Shape Index Feature)。

（2）建立局部坐标系

得到相似变换之后的人脸图像和基准点之后，就可以以相似变换后的各个基准点为中心，建立局部坐标系来提取特征。对于基于区域统计的特征，例如LBP[22,49]、

SIFT[23]、LE[48]等，以各个基准点为中心，在其周围先截取一个矩形区域。然后再将这个矩形区域均匀的划分成多个小的矩形区域。接着在每个小的矩形区域里提取这些特征。最后将所有小矩形区域里的特征拼接起来即可。对于像素灰度值特征，取两个像素点的灰度差异作为特征。

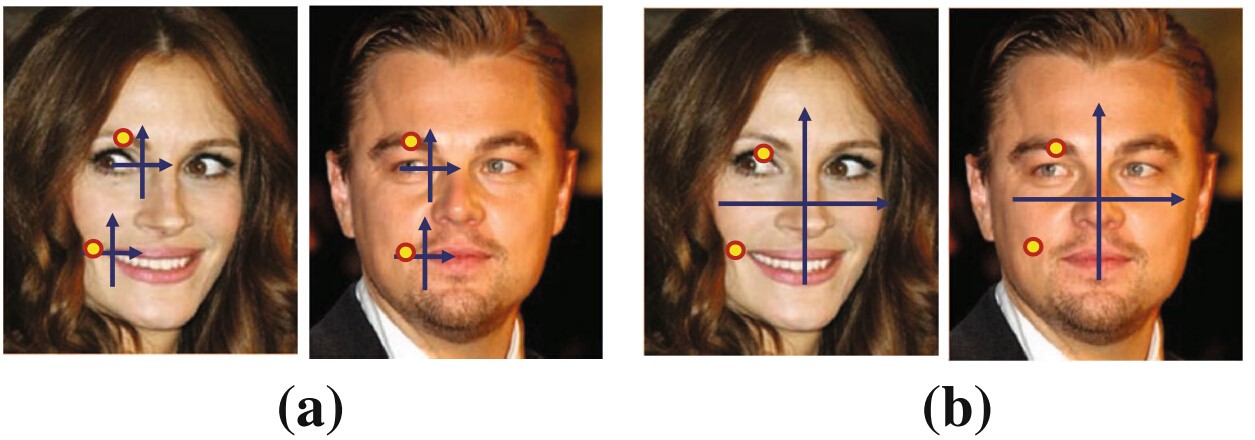


图 3.10 (a). 局部坐标系的像素索引具有相同的语义 (b). 全局坐标系下同位置的像素具有不同语义

Fig 3.10 (a). Pixels indexed by the same local coordinates have the same semantic meaning (b). pixels indexed by the same global coordinates have different semantic meanings due to the face shape variation

以各个基准点为原点，以水平方向为*x*轴，竖直方向为y轴建立局部坐标系，每个像素提取点都是相对于其中某个局部的坐标系而言的。在实际中，一般用两个像素点的差作为特征，以获得更好的光照不变性。我们称这种以各个基准点为原点提取特征的方式叫局部坐标系特征，它是相对于以图像左上角为原点提取特征的这种全局坐标系特征而言的。这特征计算量极低，并且足够用来提供训练数据。全局坐标系不能适用于非刚性的人脸形状，因大部分有用的特征都是分布在突出的部位四周，比方说眼睛、鼻子和嘴巴，而且标记点的位置在人脸不同表情、不同姿态的时候变化很大。而使用局部坐标索引像素，以最近的标记点作为参考。这样索引可以保持不变性，特征拥有很强的鲁棒性。

如图3.10 (a)所示，以左眼右侧和嘴巴左侧的特征点为原点，建立局部坐标系，在给定两个不同偏移量大小的两个点所表示的含义是一样的，都为眉毛和嘴角。而3.10 (b)所示的在全局坐标系下，同样位置的两个点却表示人脸的不同位置。因此局部坐标系能够使得特征对齐，更加鲁棒，可以有效解决因旋转和尺度变换带来的干扰。

### **3.3.3** 自适应初始化

为了解决简单模型中速度过慢的问题，提出了级联回归模型。这个模型基于以下观察：在以往的人脸检测算法中，例如在Viola & Jones的人脸检测算法中，为了提高算法的速度，一般采用的是级联结构(Cascade Structure)，这种结构的特点是将人脸检测器分成多个阶段，每个阶段都拒绝掉一定量的负样本。因为图片中的绝大多数背景区域都会在很早期的时候就被拒掉，因此可以加速人脸检测算法。而在目前的人脸校准算法中，例如：ESR[41]、 SDM[50]、LBF[51]等，这几种算法一般也分为多个阶段，在初始阶段一般解决的是大范围的变化，例如人脸角度的变化等。而在后面的阶段一般解决的是更加精细的变化，例如不同人的各个基准点的位置的微小差异等。

为了提高人脸特征点定位的准确度，级联回归模型一般采用多个初始化形状独立的运行多次，然后取最优的结果作为最终结果，这显然对速度有明显的影响。针对上述问题，本文提出一种自适应初始化策略。

上文提到，级联回归模型是一个由粗略到精细不断调整优化的过程，即使采用不同的初始化位置，在一定次数的迭代后都能达到相近的预测。如果不能达到相同或近似的值，那么很明显，初始化位置差异太大，也说明回归失败。为此我们的策略如下：给定一张图片以及一组不同的初始化值，最初只执行10% 的级联次数；然后检测预测值的变化，如果变化小于给定的阈值，那么剩下90% 的级联能够继续执行；否则，就选择一个新的初始值重新开始。后续实验证明，在保证正确率的前提下，本方法对人脸特征点定位的速度有很大提升。

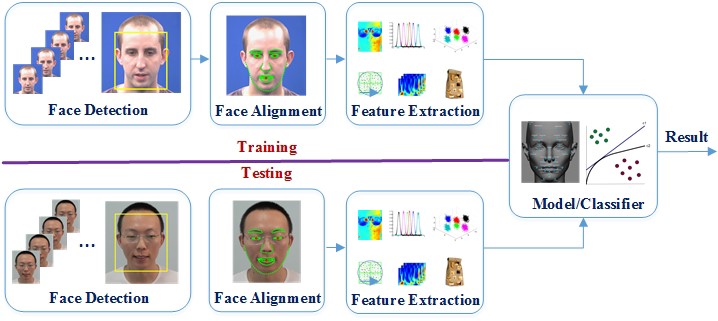
## **3.4** 本章小结

本章主要介绍了人脸特征点定位的相关工作。人脸特征点是人脸识别的重要的预处理部分，也是人脸对齐、特征提取等基础的工作。如何针对真实监控场景下的人脸形变问题而快速的得到精确的人脸特征点对后续的工作有非常大的作用。首先，介绍了经典的AAM框架，AAM作为一种经典的框架，一般作为基准算法进行对比。本章介绍此节目的也是为了跟本文提出的方法进行对比。其次，本章介绍了级联回归模型，本文所提出人脸特征点定位方法在经典的级联回归模型上进一步改进，通过采用局部坐标系下的形状索引特征、两级级联回归和自适应的初始化方法，实现了一种快速而又精确的人脸特征点定位方法，为后续的人脸识别工作提供了保障。

# 第四章 基于度量学习的人脸确认与辨识

## **4.1** 人脸识别研究概述

自动人脸识别的经典流程分为三个步骤：人脸检测、人脸特征点定位（又称Face Alignment人脸对齐）、特征提取与分类器设计，如图4.1所示。一般而言，狭义的人脸识别指的是“特征提取+分类器”两部分的算法研究。



#### 图 4.1 人脸识别框架

Fig 4.1 Framework of face recognition

对人脸识别等视觉任务而言，特征表示和模式分类是两个核心的步骤，其中又以特征提取最为关键。过去几十年，人脸识别的发展史在很大程度上是用来表示人脸特征方法的变迁史。最早的人脸识别文献大多采用直觉上“有效”的面部几何特征描述（如嘴巴大小等）来表示人脸，但实践很快表明了其区分力的不足。 1991 年之后，以EigenFaces[52]为代表的子空间分析方法在人脸识别领域几乎一统天下，衍生

出FisherFaces[53]，LaplacianFaces[54] 以及 2DPCA[55]，2DLDA[56] 等不计其数的子空间分析方法。这些方法直接采用人脸图像中所有像素的颜色或亮度值作为初始特征，然后对它们进行“变换”得到更具区分力的人脸表示。所采用的“变换”通常是在训练集合上学习而来的，最经典的优化目标是最大化 Fisher 判别准则，即类内差异小且类间差异大。为克服上述方法直接以“颜色或亮度值”作为原始特征的局限性，2000 年之后，涌现出很多对邻域像素亮度或颜色值进行局部特征提取的方法，其中既包括在图像分类领域取得极大成功的SIFT， HOG等局部特征，也包括尤其适用于人脸分析的LBP和Gabor特征。这类方法的共同特点是汇聚局部邻域像素区域的亮度值形成局部特征，再采用子空间分析等方法对局部特征进行特征变换。

一般而言，人脸识别的研究历史可以分为三个阶段。在第一阶段（1950s-1980s），人脸识别被当作一个一般性的模式识别问题，主流技术基于人脸的几何结构特征。在第二阶段（1990s）人脸识别迅速发展，出现了很多经典的方法，例如，EigenFace，

FisherFace和弹性图匹配，此时主流的技术路线为人脸表观建模。在第三阶段（1990s末期到现在），人脸识别的研究不断深入，研究者开始关注面向真实条件的人脸识别问题，主要包括以下四个方面的研究：1）提出不同的人脸空间模型，包括以线性判别分析为代表的线性建模方法，以Kernel方法为代表的非线性建模方法[57]和基于3D信息的3D人脸识别方法[58]。2）深入分析和研究影响人脸识别的因素，包括光照不变人脸识别、姿态不变人脸识别和表情不变人脸识别等。3）利用新的特征表示，包括局部描述子（Gabor Face, LBP Face等）和深度学习方法[28,29,59,60]。4）利用新的数据源，例如基于视频的人脸识别和基于素描、近红外图像的人脸识别。

在深度学习出现以前，人脸识别方法在特征表示部分一般采用高维人工提取特征（例如，LBP， Gabor等）并进行降维两个步骤，代表性的降维方法有PCA，LDA等子空间学习方法和LPP等流行学习方法。在深度学习方法流行之后，代表性方法为从原始的图像空间直接学习判别性的人脸表示。

2007年以来，LFW数据库成为事实上的真实条件下的人脸识别问题的测试基准。

LFW数据集包括来源于因特网的5749人的13233张人脸图像，其中有1680人有两张或以上的图像。LFW的标准测试协议包括6000对人脸的十折确认任务，每折包括300对正例和300对反例，采用十折平均精度作为性能评价指标。自从LFW发布以来，性能被不断刷新。2013年之前，主要技术路线为手工设计特征或基于学习的局部描述子加度量学习。2014年之后，主要技术路线为深度学习。

2014年以来，深度学习加大数据（海量的有标注人脸数据）成为人脸识别领域的主流技术路线，其中两个重要的趋势为：1）网络变大变深。例如：VGGFace[59]采用了16层的网络，而谷歌提出的FaceNet[29]采用了22 层）。2）数据量不断增大。例如：

Facebook提出的DeepFace[59]训练集达到400万，而谷歌提出的FaceNet[29] 则采用了2亿的数据用于训练，简言之，大数据成为提升人脸识别性能的关键。

在人脸识别领域，LFW[61]是著名的基准测试集。对LFW数据库而言， 2014年是其性能得以戏剧性提升的一年。在 2014年国际计算机视觉与模式识别会议 (Conference onComputerVisionandPatternRecognition2014, CVPR2014)上，Facebook和香港中文大学在允许利用有标签外部数据且非限定的测试条件下，分别报告了 97.35% 和 97.45% 的平均分类精度，比前述高维LBP特征方法的分类错误率降低了5%。上述两个团队均采用了卷积神经网络 (Convolutional Neural Network, CNN) 的变种架构。其中， Facebook的DeepFace 方法强调前端的人脸 3D对齐和虚拟正面化预处理，以削弱姿态变化的影响。而香港中文大学的 DeepID[28]方法则强调采用多个人脸区块分别训练卷积神经网络，并最终融合形成人脸特征表示。最近，该团队进一步开发了DeepID2+[60] 系统，在上述测试环境下取得了99.47% 的正确分类精度，错误率比 DeepID 降低约1.7%。需要特别指出的是，这两个系统能够取得优异性能的另一个重要原因是均采用了大规模的标注人脸数据进行训练，而且其训练图像的分布与 LFW 测试图像（名人图像）有一定的相似性。例如，DeepFace 采用了来自 4030 人的440 万幅人脸图像（均来自社交网络）；而 DeepID 则使用了来自10177 人的约 20万人脸图像（均为网络名人图像）。当然，在 LFW 上取得 99.47%的正确分类精度并不代表人脸识别技术已经成熟。实际上，LFW 数据集仅仅代表了人脸识别众多应用场景中的一种，即西方名人新闻照片识别。人脸识别还有很多其他应用场景，比如面向银行支付的人脸验证、面向智能视频监控的人脸识别等，尤其是后者，尚处于技术远远不能满足应用需求的状态。为了实现更为鲁棒和准确的识别，需要实现更为精确的面部特征定位，并处理好姿态、夸张表情和人脸老化等难题。

比较目前最好的上述几种方法，可以发现，Facebook的DeepFace将大数据（400万人脸数据）与深度卷积网络相结合，在LFW数据集上对于人脸确认任务逼近了人类的识别精度。其中DeepFace还引入了一个Local Connected卷积结构，在每个空间位置学习单独的卷积核，缺点是会导致参数膨胀，这个结构后来并没有流行起来。 DeepID家族可以看作是DL时代人脸识别领域的一组代表性工作。最早的DeepID网络包括四个卷积层，采用softmax损失函数。DeepID2在DeepID网络的基础上，同时考虑了分类损失（identity loss) 和确认损失（verification loss），这两种损失在Caffe深度学习框架中分别可以采用softmaxwithloss层和contrastive loss 层来实现。DeepID2+则是在DeepID2的基础上，增加了每一层的辅助损失函数（类似Deep Supervised Network)。

Google发表于CVPR2015的工作FaceNet采用了22层的深层卷积网络和海量的人脸数据（800 万人的2亿张图像）以及常用于图像检索任务的Triplet Loss损失函数。值得一提的是，由于人脸类别数达到800万类，如果使用softmax loss，输出层节点将达到800万个，需要至少32GB显存（假设上一个隐层节点1024个，采用单精度浮点数），而Triplet Loss则不需要额外占用显存。FaceNet在LFW数据集上人脸确认任务十折平均精度达到99.63%，这也是迄今为止正式发表的论文中的最好结果，几乎宣告了LFW数据库从2008年到2015年长达八年之久的性能竞赛的结束。

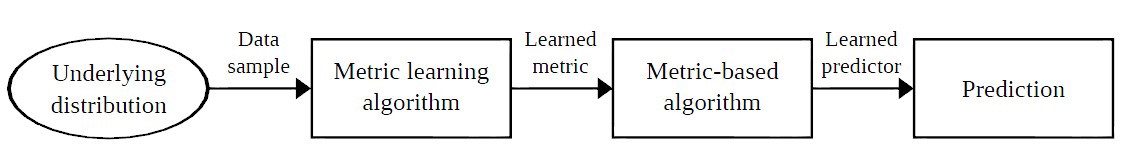
然而上述最新的识别率都是针对人脸确认的识别率。对于人类识别来讲，前文提到有两个任务。一个是人脸确认（Face Verification），另一个是人脸辨识（Face Identification）。人脸确认是一个二分类问题，其任务是比较两张图片是否是同一个人。而人脸辨识是判断一个人的身份，其问题难度随着人数规模而不断增加，因其本质上属于一个多分类问题。而这两个任务之间却有一些相似的地方，如何利用人脸确认当中成对的标签信息用于人脸辨识任务，是本文关注的问题。在以下内容中，本文将从人脸识别问题当中的人脸确认和人脸辨识两个任务分别展开介绍。

## **4.2** 基于单任务度量学习的人脸确认

### **4.2.1** 度量学习概述

相似或者非相似函数在机器学习、模式识别和数据挖掘领域发挥着重要的作用。

比如，K近邻分类器[62]通过一个度量来确定最近邻。在聚类算法中，K-means[63]则依赖于数据点间的距离测量，在信息检索中，文档通常根据与给出查询的相关性的相似值来排序。显然，这些方法的执行都依赖于度量的品质。俗语说：“同种的鸟儿合流飞”，同样，研究者们希望通过度量学习指出样本对的相似程度。通用的距离度量如欧式距离，余弦相似或者Levenshtein距离，往往不能反映数据的特质，所以，研究者们期望设计基于特别任务的度量。然而，人工调制的度量是一个困难且枯燥的过程，因此研究者们希望找到一种方法能够自动的基于数据学到度量的方法。度量学习处理的一般过程如下图4.2所示。



##### 图 4.2 度量学习处理的一般过程

Fig 4.2 The common process in metric learning

尽管可以追朔到更早的工作，如Friedman[64], Hastie等[65]，Baxter等[66]，度量学习正真出现在2002年，Xing等人[67]的先驱性工作中将其视作一个凸优化问题，并逐渐成

为一个研究热点。在各类顶级会议（ICML20101.、ECCV20102.、ICCV2011workshop3.、

NIPS20114.以及ICML20135.）中都有针对该主题的大会报告。

度量学习的目标是通过训练样本提供的信息学习到兴趣问题的度量。比如学习马氏距离（Mahalanobis）*d*M (*x*, *x*′) = √(*x* − *x*′)T *M*(*x* − *x*′)中的半正定矩阵*M*。绝大数的

1. <http://www.icml2010.org/tutorials.html>
2. <http://www.ics.forth.gr/eccv2010/tutorials.php>
3. <http://www.iccv2011.org/authors/workshops/>
4. <http://nips.cc/Conferences/2011/Program/schedule.php?Session=Workshops>
5. <http://icml.cc/2013/?page_id=41>

度量学习方法以如下表示成对或三元组约束弱监督的方式：

##### *S* = (*x*i, *x*j) : *x*i, *x*j *similar pair* *D* = (*x*i, *x*j) : *x*i, *x*j *dissimilar pair*





(4.1)

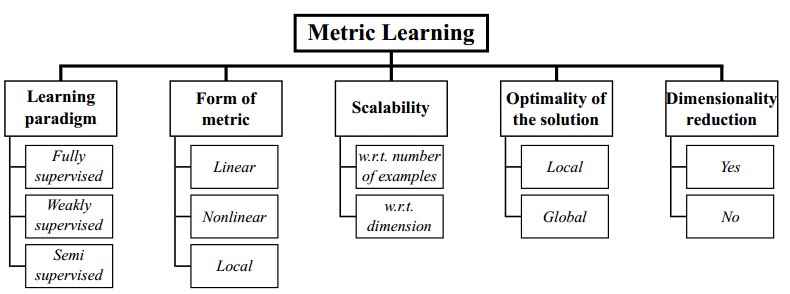
相对约束（训练三元组）为：*T* = (*x*i, *x*j, *x*k)。其中*x*i和*x*j相似，与*x*k不相似。那么一个度量学习算法的基本目标是找到一系列参数，使其最好地适应限制条件，从而最大程度地逼近底层的语义量度。通常可以写成具有以下形式的最优化问题：

min *L*(*M*, *S*, *D*,*T*) + λ*R*(*M*) (4.2)

M

其中*L*(*M*, *S*, *D*,*T*)为损失函数，当限制条件被违反时，则会触发惩罚，*R*(*M*)为正则项，调整学习到的度量矩阵*M*，λ为正则化参数。最新提出的度量学习方法，其本质上的不同大都来自度量的选择、约束条件、损失函数和正则项。度量学习后的阶段，可用于提升与度量相关的算法的性能，如K-NN，聚类算法K-Means，排序算法等。从训练数据学习一个度量并插入一个输出预测的算法（例如，一个分类器，一个回归器，一个推荐系统等），以此希望获得比基于标准度量的预测算法更好的性能。

度量学习算法可从5个关键属性来确定和描述[68]，如图4.3所示。



###### 图 4.3 度量学习算法的5个关键属性

Fig 4.3 Five key properties of metric learning algorithms

首先看学习模式（Learning paradigm）。度量学习主要有三种学习模式：

1. 全监督式（Fully supervised）：训练算法能够获取到训练实例{*z*i = (*x*i, yi)}in=1的标签，而每个训练样本*z*i ∈ *Z* = *X* × *Y*是由一个实例*x*i ∈ *X*和一个标签yi ∈ *Y*(所属的类别) 组成。*Y*是一个由标签|*Y* |组成的离散有限集合。通常，标签信息被用于产生特定的三元限制，如前面给出的*S*, *D*, *R*。
2. 弱监督式（Weakly supervised）：训练算法不能获取到训练实例的标签，只能通过集合约束*S*, *D*, *R*的形式给出侧面信息。这是很有用的设置，因为在很多应用中标记的数据的代价很昂贵，而提供侧面信息却很方便。这样的例子包括用户的隐反馈（例如，点击搜索引擎结果），在网络中对文章或链接间的引用。这都可以被看作仅在对三重约束情况下的标签信息。
3. 半监督式（Semi-supervised）：除了全监督和弱监督方式，半监督式算法能够访问未标记实例的样本（通常数据量很大）并且没有旁侧信息可用。当标记数据或旁侧信息很少的时候，半监督式可以避免过拟合。

度量的形式（Form of metric）：显然，度量形式的选择是一个关键，一般有三大主要度量家族[68]：1）线性度量(Linear metrics)，如Mahalanobis距离。由于这类距离度量容易优化，计算相对简单，且不容易过拟合（因为通常可表示为凸优化问题，因此可找到全局最优）。2）非线性度量（Nonlinear metrics），比如直方图距离，非线性度量的形式的缺点是容易引出非凸优化问题，得出的是局部最优，且可能过拟合。优点是能够捕获到数据的非线性特性。3）局部度量（Local metrics）：在局部度量中，多个（线性或者非线性的）局部度量被学习（通常是同步学习到的），从而更好的处理复杂的问题，如异构数据。然而这种形式相比于全局的方法更容易过拟合，因为所学习的参数的数量往往很多。

可扩展性（Scalability）：随着可用数据的迅猛增长，可扩展性成为机器学习领域所有问题需要关注的问题。首先，期望度量学习算法对训练样本的数目具有良好的扩展性。基于此，在线学习是解决方案之一。其次，度量学习方法应对数据的维数也具有可扩展性。然而，度量学习通常表达为学习一个矩阵的形式，设计基于如此规模的并能够合理扩展的算法依然是一个相当大的挑战。

解的最优性（Optimality of the solution）：这个属性涉及到算法的性能，即找到满足优化准则最优的参数。理想情况下，可通过全局最优来保证。这也是凸函数优化度量学习的关键。与此相反，非凸情况下，解可能仅是局部最优。

维数降低(Dimensionality reduction)：度量学习有时可设计为寻找一个到新的特征空间的映射。而这种方式的一个附加好处在于可以兼顾寻找一个低维度的映射空间，从而使得计算更快，表达更紧凑。而通常情况下，维数降低是通过正则项来驱使学到的度量矩阵是低秩的。

近年来，度量学习方法在人脸识别及Re-ID[69]等问题上得到了广泛的应用。如何为特定任务来学习适当的距离度量一直是度量学习研究的主要内容。经典的方法有基于最大间隔的多度量学习[70,71]以及融合了多种人脸区域描述子的多度量学习算法[72]。

Fu 等人[73]提出了一种学习关联度量（correlation metric）的方法，该度量学习模型在对样本降维后，可以保留样本之间的近邻关系，并针对关联度量提出了相关嵌入分析（CEA，Correlation Embedding Analysis）模型和相关主成分分析（CPCA， Correlational Principle Component Analysis）模型。Guillaumin等人[74]提出了MildML（Multiple Instance Logistic Discriminant Metric Learning）模型，并在多事例学习中得到了成功的

应用。Guillaumin 等人[75]还提出了LDML（Logistic Discriminant based Metric Learning）模型和MkNN（Marginalized k-nearest neighbor）模型来学习人脸识别中的距离度量。

LDML 模型通过将度量学习问题看作基于核的逻辑回归问题，通过极大似然估计来学习距离度量，MkNN 通过学习非线性度量学习模型，来提高模型的判别能力。为了改进已有度量模型的泛化能力，Nguyen 和 Bai 提出了余弦相似度度量学习（CSML）模型[76]，该模型使用余弦相似度来构造目标函数。文献[77]提出了一种Sub-SML的度量学习方法，通过采用马氏距离和余弦相似距离的学习，取得了在单任务学习中目前最高的识别率。

Köstinger等人[78]提出了KISS 模型用来从恒等约束中学习到距离度量，此外，该模型可以针对大规模的数据集学习距离度量。Huang 等人[79]提出了广义稀疏度量学习模型（Generalized Sparse Metric Learning, GSML），该方法为许多有代表性的稀疏度量学习模型提供了一个统一的角度，并且可以将现有的许多非稀疏度量学习模型扩展到稀疏度量学习形式。Shen等人[80]提出了推动度量（Boost Metric）方法，该方法针对传统度量学习方法中的半定规划问题不能有效地处理大规模数据的缺点，使用类似 boosting 的策略，通过将度量矩阵分解为一系列秩为1的矩阵来学习一组弱分类器，最后将它们组合。在上述方法当中，基于KISSME[78]的方法取得了在单任务学习下较好的效果，并且由于该方法不需要优化，只需要计算相似集和非相似集的协方差矩阵，所以运行效率相比其他方法快很多，可以处理大规模的数据的问题。以下我们先详细地介绍KISSME度量学习的原理，然后再给出我们提出的多任务度量学习方法。

### **4.2.2** 基于**KISSME**的度量学习

Mahalanobis距离常用于衡量两个特征向量的相似程度，对特征xi, xj ∈ Rd，其相对距离可计算如下：

*d*M (*x*, *x*′) = √(*x* − *x*′)T *M*(*x* − *x*′) (4.3)

其中M是正定的Mahalanobis矩阵。

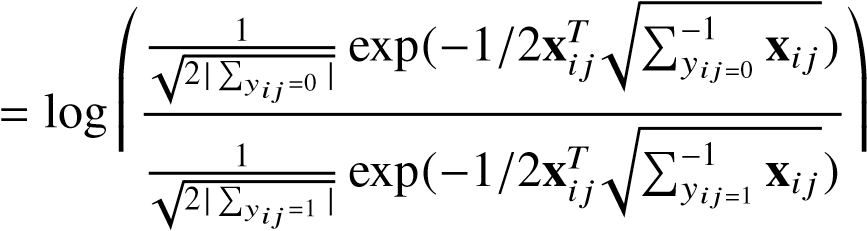
KISSME[78]是距离度量学习方法的一种，该方法通过以下函数计算两个特征是否相似，通过定义似然度来衡量输入对的相似程度：

δ(**x**i, **x**j) = *lo*g*p*(**x**ij||*H*0)) = *lo*g ( *f* (**x**ij||θ0)) (4.4) *p*(**x**ij *H*1) *f* (**x**ij θ1)

(

其中*H*0表示*x*i, *x*j是非相似的，当δ(*x*i, *x*j)值较大时成立，相反，*H*1表示*x*i, *x*j是相似的，当δ(*x*i, *x*j)值较小时成立。

假设特征差xij =xi − xj符合高斯分布，问题可转换为：

δ(**x**i, **x**j)  (4.5)

其中的方差矩阵通过下面公式计算, ： -

Σyi, j=0 = ∑(**x**i, **x**j)(**x**i, **x**j)T (4.6) yij=0

Σyi, j=1 = ∑(**x**i, **x**j)(**x**i, **x**j)T (4.7)

yij=1

上式中，当**x**i, **x**j相似时，yij = 1，反之yij = 0。因此，Mahalanobis度量矩阵*M*可以使用如下公式计算：

*M* =Σy−ij1=1 − Σy−ij1=0 (4.8)

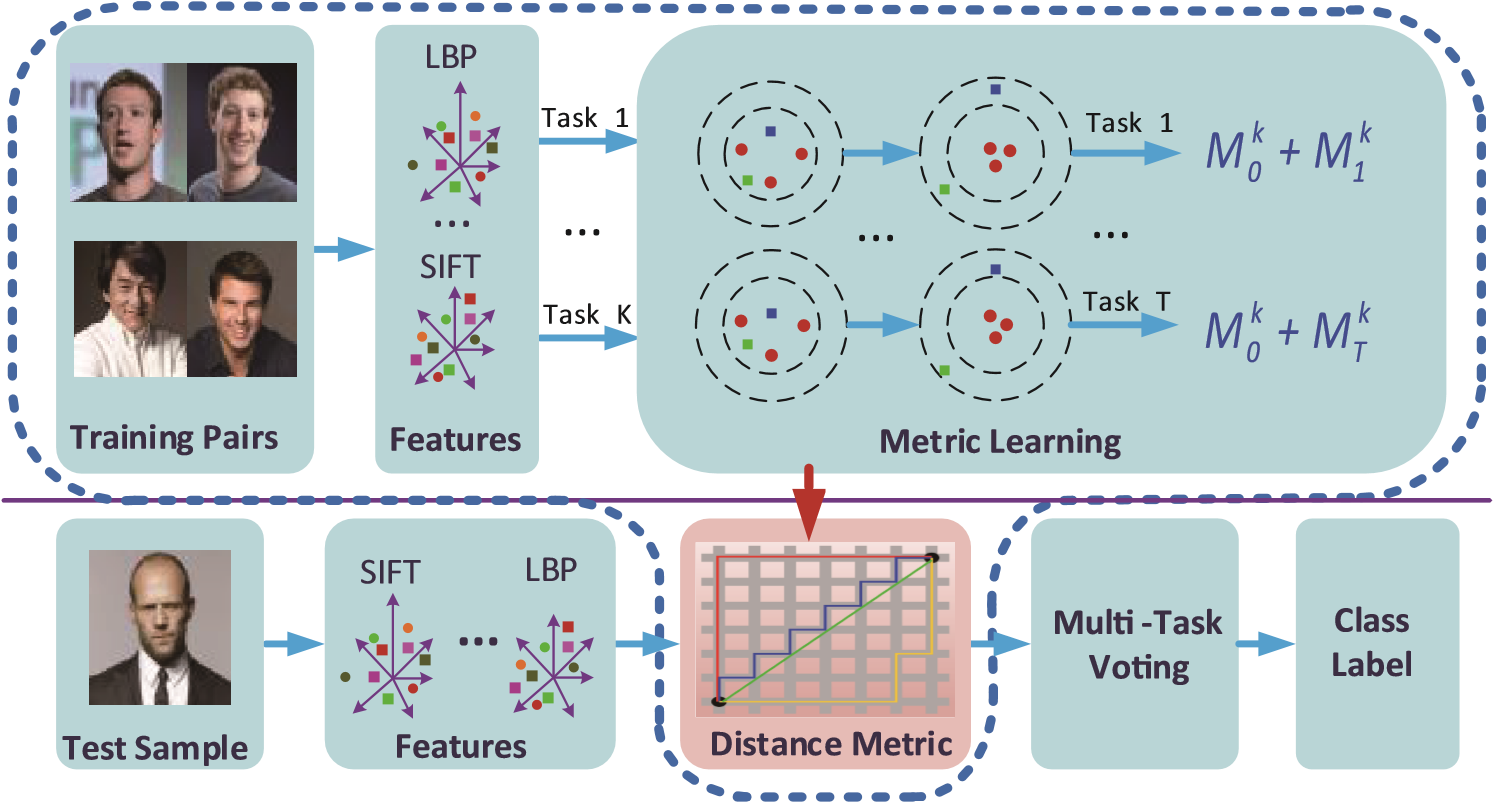
从而避免了迭代优化的过程。在KISSME度量学习实现当中，最耗时的操作是矩阵求逆，由于涉及矩阵求逆，其复杂度是矩阵维度的三次方，即*O*(*n*3)。但由于一般在使用 KISSME 的时候需要对特征进行降维，维数一般不会太高，且在整个过程中求逆操作只进行了两次，不需要耗时的迭代过程，因此KISSME的最大优点是训练速度非常快。

## **4.3** 基于两级多任务度量学习的人脸辨识

### **4.3.1** 两级多任务度量学习

KISSME算法的优点在于可以直接求解出度量矩阵，计算简单快速，不需要繁杂的迭代优化过程。KISSME解决单个特征的二分类问题，然而当面对多分类问题并且多种特征时，如何利用KISSME取得更好的分类效果且依旧保持其优点是我们想要解决的问题。因此，我们在KISSME的基础上，对其进行改进，提出双层多任务度量学习（Two-level Mult-Task Learning, TMTL），使其适用于多种特征多个任务的学习，提出的两层多任务学习框架如图4.4所示。第一层学习多特征，第二层计算不同学习任务之间的共享信息机制。

定义*k* ∈ {1, 2, ..., *K*}是第一层的第*k*种特征。对每个特征类型的每个子任务，以共享度量矩阵**M**0k和特定矩阵**M**kt 相结合的形式，计算任务*t* ∈ {1, 2, ...,*T*}的度量矩阵。即



##### 图 4.4 多任务度量学习人脸辨识框架

Fig 4.4 Framework of multi-task metric learning for face identification

第一层为选择对应的特征，第二层是在第一层选定的特征下进行多任务学习。因为我们考虑的是人脸辨识问题，将人脸辨识问题转化为多任务学习是基于这样的观察：大部分的人脸都包含共同的相似的区域，而每个人脸的表达是公共的部分和特有部分的叠加。因此，给出一对训练样本，其距离定义为：

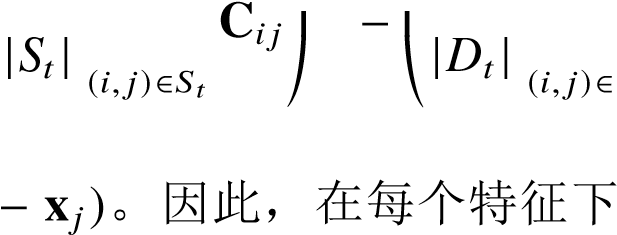
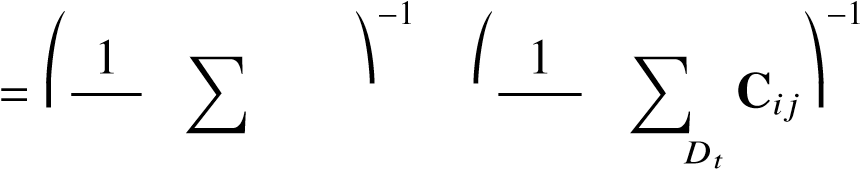
K

*d*t2(**x**i, **x**j) = ∑(**x**i  **x**j) (4.9)

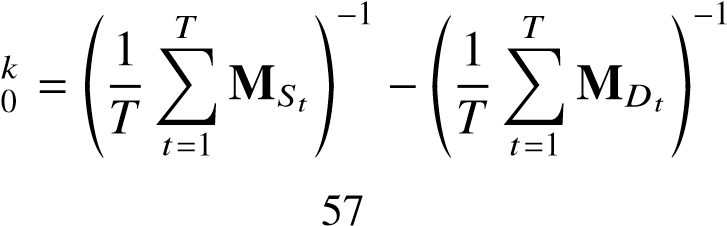
k=1

每一个任务定义了一个特定任务（task-specific）的相似对和非相似对的子集：*S*t = {(*i*, *j*) ∈ ψt |yi = yj} 和 *D*t = {(*i*, *j*) ∈ ψt |yi , yj}。根据公式4.7可以得到，第*k*个特征的第*t*个子任务的度量矩阵为：

**M**kt t

  (4.10)

其中，**C**ij = (**x**i − **x**j)T(**x**i,的，所- 有子任务公共的度量矩阵**M**0k可以通过所有子任务的平均估计得到，即：

**M** (4.11)

**Algorithm 2:** 基于两级多任务度量学习框架

**1** 输入：训练数据对{**x**ki , **x**kj } ∈ Rdk, *k* = 1, 2, ..., *K*, *K* 是特征总数. 子任务

*t* ∈ {1, 2, ...,*T*}. k k t

1. 输出：最终度量矩阵 **M**ˆ kt ∈ *R*(d ×d ) .
2. 定义一个基于特定任务的相似对 *S*t 和不相似对 *D*t子集.
3. 计算平衡因子 µ.
4. **for** *k* ← 1 **to** *K* **do**

**6for** *t* ← 1 **to** *T* **do**

**7**对*S*t 和 *D*t提取第*k*个特征.

**8**根据公式(4.10)计算第*k*个特征的特定任务的度量矩阵.

**9**根据公式(4.11)计算共享的度量矩阵.

**10**根据公式(4.12)得到最终度量学习矩阵**M**ˆ kt .

**11end**

1. **end**
2. 返回：**M**ˆ kt ∈ *R*(dk ×dk)t .

然后，最终的马氏（Mahalanobis）度量矩阵为：

**M**ˆ kt = **M**0k + µ**M**kt (4.12)

相应地，**M**0k表示了对于每个特征在所有子任务上的共同的特性，而**M**kt 聚焦于特定任务的特性，其中的平衡因子µ 在共享度量矩阵**M**0k和特定子任务度量矩阵**M**kt 取得平衡。在这里，µ 是子任务样本所占所有训练样本的比例。总结上述工作，本文提出的两级多任务度量学习算法如算法2所示。

### **4.3.2** 多任务投票

为了充分利用两级多任务度量学习的方法，所有的基于特定任务的度量矩阵组合起来转化为一个多分类的决策。类似多数投票机制，文献[81]采取了将类别标签赋予获得最多比较对票数的类别。

假设对任务*t*的正样本和类别标签一致，即**x**i : yi = *t*。给定一个测试样本**x**i，投票机制需要确定测试样本属于哪一个任务标签。为了提高比较的鲁棒性，我们不仅采用了任务*t*的单个的度量矩阵进行比较，而且采用了任务*u* ∈ {*u* , *t*|*u* = 1, ...,*T*}的度量矩阵进行辅助性的度量。至此，最终的类别标签为获得最多投票次数的子任务标签。为了简明的说明多任务投票机制，给出了算法3进行说明。

**Algorithm 3:** 多任务投票机制

1. 输入：距离度量矩阵**M**ˆ kt ∈ *R*(dk ×dk)t ；每个特定任务的索引集ψt；测试样本 **x**i. 输出：类别标签*c* ∈ {1, 2, ...,*T*}.
2. 初始化：*I*t ← 0, *Label*(*t*) ← 0, *c* ← 0, *A* ← 0, *B* ← 0.
3. **for** *t* ← 1 **to** *T* **do**
4. **while** *u* , *t, and u* ∈, ...,*T*} **do 5**计算 *A* ← min *d*t (**x**i, **x**j).
   1. ∈ψt ∧yj=t

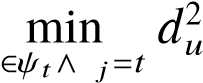
**6if** *A* ≤ min *d*t2(**x**i, **x**k) **then**

* 1. ∈ψu∧yk=u

**7***I*t ← *I*t + 1.

**8end**

**9**计算 *B* ←(**x**i, **x**j).

 j y

**10 if** *B*(**x**i, **x**k) **then** k ∈ψu∧yk=u

**11***I*t ← *I*t + 1.

**12end**

1. **end**
2. *Label*(*t*) ← *I*t.
3. **end**
4. *c* ← argmax = *Label*(*t*).
5. 返回：别标签 *c*.

### **4.3.3** 算法复杂度分析

该方法的运算复杂度主要有KISSME度量学习和两级多任务两部分组成，定义特征维度为*n*，两级多任务当中第一级特征个数为*K*，第二级类别个数为*T*，即总共包含*T*个人，类别规模为*T*，每个人有*m*张图片，即*m*个正样本。下面将从训练时间复杂度和测试时间复杂度进行分析。

训练时间复杂度。在KISSME的度量学习当中，可以通过公式(4.8)来近似计算马氏度量矩阵*M*。为了保证*M*的半正定性质，可以将矩阵*M*的负特征值变为0，保留非负特征值的方式，即先进行特征值分解：

*M*ˆ = *X*TΛ*X* (4.13)

其中*M*ˆ 是特征向量串联而成，对角阵Λ的对角元素都为特征值。令

Λ′ = *dia*g(*max*(0, λ1), ..., *max*(0, λn)) (4.14) 然后重新计算的*M*即满足了半正定的约束，即：

*M* = *X*TΛ′*X* (4.15)

上面的计算中，最耗时的操作是矩阵求逆，复杂度是矩阵维度的三次方，即*O*(*n*3)，但由于一般在使用 KISSME 的时候需要对特征进行降维，维数一般不会太高，且在整个过程中求逆操作只进行了两次，不需要耗时的迭代过程，因此 KISSME 的最大优点是训练速度非常快。

在两级多任务当中，对于每个特征，都要计算*T*个子任务当中的距离度量矩阵，尽管不同特征的特征维度不一致，但可通过PCA降维归一化到统一的大小。因此，整个两级多任务度量学习训练的时间负责都为：

*O*(*n*3)*O*(*KT*) = *O*(*n*3*KT*) (4.16)

由于n可以看作固定值，而一般特征的个数选择不会很大，所以该问题的复杂度随着第二级当中的子任务规模而增加。这意味着，当需要识别的人数越多，类别标签也越多，训练时间也越长。当*T*远大于*n*和*K*时，算法的复杂度近似于*O*(*T*)。

测试时间复杂度。对于新来的一个测试样本，其任务是确定该样本的类别标签，即身份信息。在两级多任务度量学习当中，类别标签即为属于最大投票数量的子任务标签。因为子任务*t*的正样本和类别标签是一致的。首先，对于测试样本，首先要与每个子任务*t*的每个正样本计算一次距离，为了保证投票的准确性，还与非当前任务的其它子任务的度量矩阵进行辅助性的度量，因此每个子任务的复杂度为：

|  |  |
| --- | --- |
| *O*(*m*) + *O*(*T* − 1)*O*(*m*) = *O*(*Tm*)  因其共有*T*个子任务，*K*个特征，所以最终的算法复杂度为： | (4.17) |
| *O*(*K*)*O*(*T*)*O*(*Tm*) = *O*(*KT*2*m*) | (4.18) |

可以看出，最终的测试时间与特征个数和子任务个数以及子任务的正样本都有关系，一般在测试的过程当中，*K*和*m*一般都不会很大，一般为常量。所以，测试时间与类别个数的规模有直接关系。上述算法复杂度的分析将一次距离的比较作为基本的操作，因为算法复杂度为*O*(*KT*2)，所以该算法在大规模的人脸辨识中测试将很耗时。一般在类别数目不大时，因其是多项式的复杂度，测试完全能保证实时性。

## **4.4** 本章小结

本章在人脸检测和人脸特征点定位的基础上进行了人脸识别相关的研究。人脸识别的问题其本质上还是距离度量的问题，其中人脸确认，即是对两张人脸图片相似度的度量，也是距离的度量。而人脸辨识是一个多分类问题，本章在基于KISSME方法之上提出一种多任务度量学习的框架，该方法不但很好的解决了人脸辨识的问题，而且可解决大部分多分类问题。另一方面，由于人脸辨识采用了基于度量学习的方法，本章还对度量学习方法进行了综述，与传统的分类方法相比，度量学习方法能够学习针对数据自身的特点，量身定做出相应的距离描述，从而更好地解决多分类问题。

# 第五章 实验结果与分析

## **5.1** 人脸检测结果与分析

### **5.1.1** 实验数据库与设置

**FDDB**[82]：FDDB（Face Detection Data Set and Benchmark）数据库是目前最大而且最难的公幵人脸检测数据库，它是由美国马萨诸塞大学（ University of Massachusetts）阿姆斯特分校（Amherst）计算机视觉实验室运营的人脸检测测试平台，所有团队可以通过其包含5171张人脸的2845张图片的测试集来测试检测技术的精度。在平台上，所有参与者都将面对学界、商界挑剔的目光，无论是百度、腾讯这样的互联网巨头，还是技术实力雄厚的技术团队，测试结果都将被公开展示，技术实力的排名被实时更新。它里面包含了各种困难的情况，包括遮挡、不同的角度、低分辨率和失焦的人脸等等。作为全世界最具权威的人脸检测评测平台之一，其公布的评测结果也代表了人脸检测的世界最高水平。

这个数据库包含了两种测试方案：离散的方案和连续的方案。在离散的方案中，只有当算法检测到的框和某一个标注的框的重合度大于0.5的时候，这个检测结果才算是正确的。这里的两个框的重合度的定义是两个矩形的交集的面积除以这两个矩形的并集的面积。这种离散的测试方案被广泛应用在各种物体检测算法评价中。而连续的测试方案指的是每个算法检测到的框和标注的框的重合度会作为权重叠加到每个检测结果上。这个测试方案更加严格，它要求检测出来的框不仅要正确，而且还要稳定，要尽量和标注结果保持一致。

FDDB是在无限制真实条件下获取的人脸图片，其中包含了遮挡、不同角度、低分辨率等照片，这与监控场景下人脸检测所带来的问题一致。因为FDDB是基准数据库，所以我们在该数据库上进行了测试。其测试数据为官方指定的2845张人脸图片。

图5.1是FDDB数据库的已经人工标记的人脸图片。

### **5.1.2** 性能评价准则

由于全世界研究人员的不懈努力，每年都有许多的人脸检测方法甚至是商用人脸检测系统的出现，为了便于统一管理和对这些检测方法进行测评，需要提供一个统一的测试集，并制定一套统一的指标来测试这些不同的方法。为此，世界上一些著名的大学和研究机构都做了自己的工作。人脸检测主要的评价标准有检测率、假阳



##### 图 5.1 FDDB多姿态人脸示例

Fig 5.1 Illustration of pose varied faces in FDDB

率、假阴率、及算法执行效率。而检测率、假阳率、假阴率统一在接收者操作特性

（Receiver Operating Characteristic, ROC）曲线[83]之内，简称“ROC曲线”。因此，本文主要以ROC曲线以及算法执行效率两个方面对人脸检测性能进行评价。

（1）ROC曲线

在认识ROC曲线之前，需要先了解以下几个概念：

1. 检测率

被正确检测到的人脸数与图像内包含的人脸数的比值。检测率越高，说明检测系统对人脸的接受能力越强。

1. 假阳率

也称误检率、虚警率、误报率，定义为负类样本被错误分类为正类的样本数与所有负类样本数的比例。设图像内被检测的所有非人脸子窗口数为*M* ，被误检为人脸的非人脸子窗口数为*N* ，则假阳率为*N*/*M* 。检测率无法反映系统对非人脸的排除能力，有可能出现这种情况所有人脸都被检测到，但同时很多非人脸区域被误认为是人脸。因此引入假阳率来衡量系统对非人脸的排除能力。假阳率越低，说明检测系统对非人脸的排除能力越强。

1. 假阴率

也称漏检率，即被误检为非人脸的人脸子窗口数与图像内被检测的所有人脸子窗口数的比值。设图像内被检测的所有人脸子窗口数为*M* ，被误检为非人脸的人脸子窗口数为*N* ，则假阴率为*N*/*M* 。

人脸检测本质上是一个二分类问题，基本框架是对一个滑动窗口内的图片进行分类。对一个二分问题来说，会出现四种情况。如果一个实例是正类并且也被预测成

##### 表 5.1 混淆矩阵参数

Table 5.1 The parameters of confusion matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | 预测 | |  |
| 人脸 | 非人脸 | 合计 |
| 实 人脸 | True Positive(TP) | False Negative(FN) | Actual Positive(TP+FN) |
| 际非人脸 | False Positive(FP) | True Negative(TN) | Actual Negative(FP+TN) |
| 合计 | Predicted Positive  (TP+FP) | Predicted Negative  (FN+TN) | TP+FP+FN+TN |

正类，即为真正类(True Positive, TP)。如果实例是负类被预测成正类，称之为假正类(False Positive, FP)。相应地，如果实例是负类被预测成负类，称之为真负类(True Positive, TP)，正类被预测成负类则为假负类(False Negative, FN)。如表 5.1所示，1代表正类，0代表负类。其给出了评价分类器好坏的性能指标，即混淆矩阵。

由表5.1可以得到总共有四个指标。其一是真正类率(True Positive Rate, TPR)，计算公式为*TPR* = *TP*/(*TP*+*FN*)，刻画的是分类器所识别出的正实例占所有正实例的比例。另外一个是负正类率(False Positive Rate, FPR)，计算公式为*FPR* = *FP*/(*FP*+*TN*)，计算的是分类器错认为正类的负实例占所有负实例的比例。还有一个真负类率(True

Negative Rate, TNR)，计算公式为*TNR* = *TN*/(*FP* +*TN*) = 1 − *FPR*。

在一个二分类模型中，对于所得到的连续结果，假设已确定一个阀值，比如说0.6，大于这个值的实例划归为正类，小于这个值则划到负类中。如果减小阀值，减到0.5，固然能识别出更多的正类，也就是提高了识别出的正例占所有正例的比类，即TPR，但同时也将更多的负实例当作了正实例，即提高了FPR。为了形象化这一变化，在此

引入接受者操作特性（Receiver Operating Characteristic, ROC）曲线，简称：ROC曲线。

曲线是由两个变量的组合，1-Specificity和Sensitivity。由于1-Specificity=FPR，即负正类率。Sensitivity即真正类率(True Positive Rate, TPR)，反映了正类覆盖程度。这个组合以1-Specificity对Sensitivity，即以代价(costs)对收益(benefits)。

用ROC曲线来表示分类器的性能很直观好用。可是，人们总是希望能有一个数值来标志分类器的好坏，于是Area Under roc Curve(AUC)就出现了。顾名思义，AUC的值就是处于ROC 曲线下方的那部分面积的大小。通常，AUC的值介于0.5到1.0之间，较大的AUC代表了较好的性能。

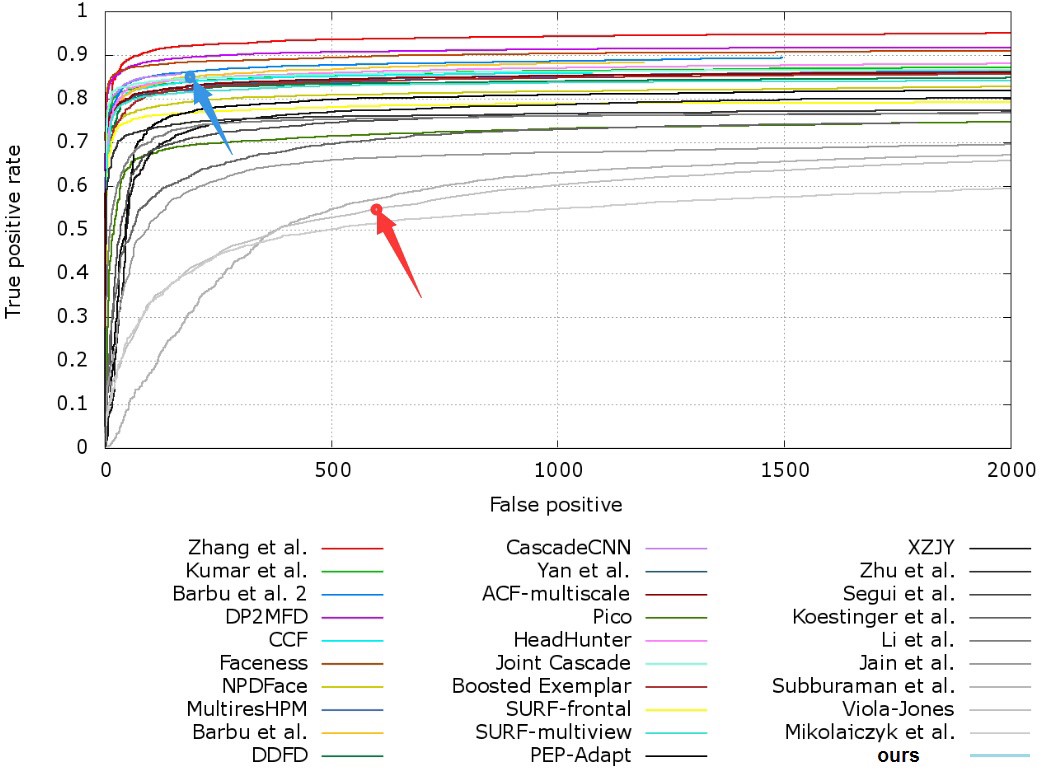
为了更加直观的表示算法性能，一般将各个算法曲线放在同一个图中曲线越靠近左上方，说明算法的性能越好。

（2）算法执行效率

大部分应用领域需要实时地检测人脸，如人脸识别、人脸跟踪，视频监控等。在检测率和误检率达到满意的前提下，检测速度越快越好。算法执行效率在理论层面有算法复杂度的分析。在实际测试过程当中，对给定配置的机型，在相同条件下，对*N*张指定大小的图片进行测试，得到总用时为*Ts*，则可得到算法检测帧率为 NT *f ps*。

### **5.1.3** 实验结果和分析

实验采用了离散的方案在FDDB上进行人脸检测算法的评估。为此，我们和FDDB上最新的已经发表的结果[84–88]进行比较，与学术界己发表的算法的比较结果如图5.2所示，与商业系统的比较结果如图5.3所示。

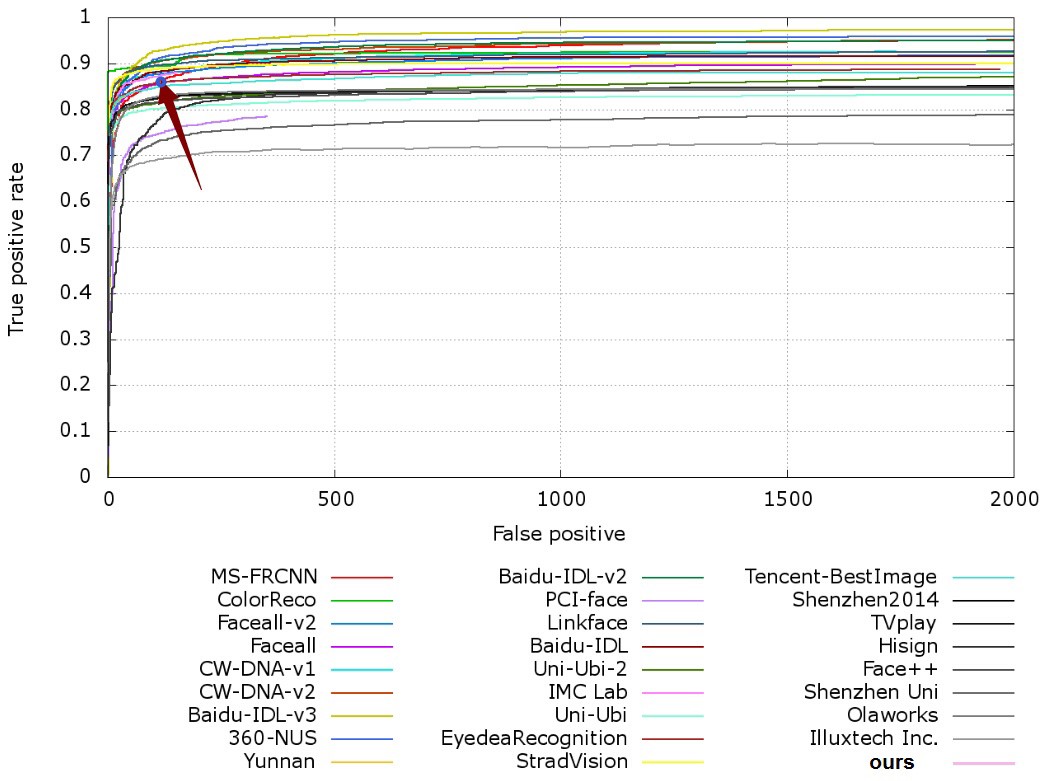


##### 图 5.2 学术界己发表的算法的比较结果

Fig 5.2 Comparison with published methods

图5.2中与其它方法的比较结果数据由FDDB1.提供，其中红色箭头（下）对应的方法为基于Haar Adaboost，而蓝色箭头（上）表示我们的基于DPM的算法，ROC曲线越往左上角说明性能越好，可以看出，我们的算法虽然没有最新算法好，但是和基于Adaboost传统方法比较有非常大的提升。而目前开源的算法当中，基于Haar Adaboost的算法还是大多数。而我们的算法是在开源方法当中性能最好的算法。人脸检测算法从上一结果可以看出，基于DPM的算法虽然没有达到最好的效果，但是却仅次

1. <http://vis-www.cs.umass.edu/fddb/results.html>



##### 图 5.3 与商业系统的比较结果

Fig 5.3 Comparison with unpublished methods

于基于Deep Learning 的方法。而我们基于DPM的人脸检测算法利用了比深度学习少的数据和训练时间。上述最好的算法并没有对算法的检测效率进行比较。我们知道基于Haar Adaboost 算法实时性很高，为此我们跟OpenCV自带的检测器进行了比较了算法的执行效率，其效果如表 5.2所示。

我们的人脸检测器在运行时的时间和内存开销都十分的低。我们和OpenCV中的人脸检测器以及文献[89]中的人脸检测器进行比较，我们比较不同的算法在相同情况下的时间和内存开销。对于所有的方法，我们都在分辨率为720p的图片中检测最小为80 × 80的人脸。在一个普通的PC上、使用单线程，仅仅花了 31.5ms就能够处理完

表 5.2 与OpenCV人脸检测比较

Table 5.2 Comparison with OpenCV face detector

|  |  |  |
| --- | --- | --- |
| 指标 | **OpenCV** | **Ours** |
| 速度 | <15fps | 30fps |
| 检测率 | 低 | 高 |
| 误检率 | 高 | 低 |
| 指令集优化 | 20%提升 | 40%提升 |

成。这个速度比文献[89]中的人脸检测器快了1000多倍。OpenCV的人脸检测器需要 62.8ms，而且我们的算法的识别精度要远远高于OpenCV的人脸检测器。同时，在运行时的内存幵销方面，我们的人脸检测器仅仅需要20MB的内存。而在其他的方法中，例如文献[16]中的人脸检测器需要150MB的内存，而文献[90]中的检测器需要866MB的内存。和这些方法相比，我们的人脸检测器更加的实用，特别是在移动设备和嵌入式设备上。



##### 图 5.4 OpenCV与本文方法在零误检下的比较

Fig 5.4 OpenCV detector vs. our method in zero false positive



##### 图 5.5 OpenCV与本文方法在全召回下的比较

Fig 5.5 OpenCV detector vs. our method in full recall

图5.4和图5.5分别展示了基于Opencv的方法和我们的方法在零误检时候和全召回情况下的结果。可以清晰的看到，在图5.4中我们的方法和OpenCV当中的人脸检测器在没有任何误检的情况下，我们的方法并未遗漏掉任何人脸。在图5.5中我们的方法和OpenCV自带的方法在全部检测到人脸的情况下，OpenCV误检了三个人脸，而我们的方法没有任何误检，再一次证明了我们方法的鲁棒性。这也为后续的人脸特征点检测和识别提供了有效的检测方法。

## **5.2** 人脸特征点检测结果与分析

### **5.2.1** 实验数据库与设置

早期的人脸配准数据集主耍在实验场景下拍摄，而最近的人脸配准数据集往往从实际场景采集。相对实验场景的人脸配准数据库，自然场景的人脸配准数据库在姿态变化、表情、遮挡、光照强度方面面临着更大的挑战。在实验当中，我们采用真实场景下的Helen[91] 和LFPW[92]这两个数据库进行对比实验。

**Helen**[91]：Helen人脸数据库是2012年ECCV的论文当中提出的。它包括两个目录，分别为训练样本集和测试样本集。它包含了2330张高分辨率真实场景下的人脸图片，且每个人脸都被精细的194个人脸特征点标记。其中大部分的图片都来自国外社交网站Flickr。其中有2000张作为训练集，330张作为测试集。在Helen数据库当中，每个人脸的大小是550个像素，即使最小的人脸也有150像素，所以它被用于非常精细的人脸特征点定位，是一个基准数据库。其样例图片如图5.6 所示。



图 5.6 Helen数据库样例

Fig 5.6 Example of Helen dataset

**LFPW**[92]：该数据库在2011年被Belhumeur等人创建，全名为：Labeled Face Parts in the Wild。数据库是从网络上获取的真实场景下的图片，其中包含了形变、光照、表情、遮挡等恶劣条件的人脸照片。其目的是为了测试人脸配准（人脸特征点检测）在真实无限制条件下的性能。由于该数据库只提供了图片的网络地址，但是一些网络地址不是一直有效，因此，我们只从网络上获取到了包含1100张训练集当中的810张，以及包含300张测试集当中的240张。该数据库的每张人脸都被三个MTurk2.工人所标

2. <https://www.mturk.com/mturk/welcome>

##### 表 5.3 数据库数据分布

Table 5.3 Distribution of the datasets

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 表情 | | | | | 姿态 | | 遮挡 |
|  | 中性 | 微笑 |惊讶 | 恶心 | 尖叫 | 斜视 | 0°-15° | 15°-30° |  |
| **LFPW** | 48.66% | 39.73% |8.05% | 0.44% | 1.78% | 1.34% | 94.69% | 5.31% | 18.31% |
| **Helen** | 43.03% | 49.09% |2.12% | 2.43% | 0.00% | 3.33% | 94.54% | 5.46% | 13.03% |

记，其中的人脸照片都被标记了29个特征点。为了获取到足够多的数据，我们按照文章[92]的方式，将训练集增加到2000张并用得到的240张做测试集。其样例图片如图5.7 所示。



图 5.7 LFPW数据库样例

Fig 5.7 Examples of LFPW dataset

为了分析这两个数据库的数据分布，我们对这两个数据库在表情、姿态、遮挡这三个方面进行统计，其具体情况如表5.3所示。

### **5.2.2** 性能评价准则

通常人脸特征点的评价准则比较依赖于人脸特征点的个数。常见的m17人脸关键点定位误差评价标准采用了十七个关键点(眉毛四个，眼睛六个，鼻子三个，嘴巴四个)，而m7评价标准采用了眼角、嘴角及鼻尖的七个关键点进行特征点定位。关键点定位主要有两个评价指标：与手工标定比较得化定位误差与定位效率。归一化均方误差(Normalized Mean Error，NME)是常常用来衡量定位效果的主要指标，即：

M1 |*p*i,j − gi,j|2

∑

M

j

=

1

*E*i =(5.1)

##### |*l*i − *r*i|2

其中，*i*表示图像标号，*j*为关键点标号，*M*是关键点的数目，*l*i是左瞳孔中心的位

置，*r*i是心瞳孔中心的位置。|*l*i − *r*i|2表示瞳孔距离IOD(Inter-Ocular Distance，IOD)，采用欧氏距离用以归一化配准误差。*p*i,j是第*i*张图片的第*j*个特征点的预测值。gi,j是第*i*张图片的第*j*个特征点的实际值。一般在计算之前，都将人脸图片根据瞳孔中心的位置归一化到固定距离，其原理是一致的。

如图5.8所示，因为配准误差是各向同性的，红绿蓝三个圆分别表示眼角关键点定位归一化均方误差为0.05*IOD*、0.1*IOD*以及0.2*IOD*。通常也会有对误差设限进行配准评价，通常设在 0.05*IOD* ∼ 0.10*IOD*。在我们的方法中，我们采取了文献[93]中所采用的方法，我们认为如果定位误差超过10%那么就是定位失败。



图 5.8 眼角定位误差示例（红：0.05IOD、绿：0.1IOD、蓝：0.2IOD）

Fig 5.8 Illustration of localization error(Red: 0.05IOD,Green: 0.1IOD, Blue: 0.2IOD)

### **5.2.3** 实验结果和分析

为了证明我们算法的精度和速度，我们分别在Helen和LFPW上两个数据上进行了指定标准的测试，并且对两个数据库算法执行效率进行了实验。实验的参数设置分别与方法ESR[41]和方法RCPR[42]保持一致。其实验结果由表5.4和5.5分别所示。

##### 表 5.4 LFPW数据库测试结果

Table 5.4 Results on LFPW

|  |  |  |
| --- | --- | --- |
| Method Error | Failures | FPS |
| ESR 3.8 | 4% | 3 |
| RCPR 3.5 | 2% | 12 |
| Human 3.28 | 0% | - |
| **Ours 3.44** | **2%** | **20** |

##### 表 5.5 Helen数据库测试结果

Table 5.5 Results on Helen

|  |  |  |
| --- | --- | --- |
| Method Error | Failures | FPS |
| ESR 7.1 | 13% | 2 |
| RCPR 6.5 | 8% | 6 |
| Human 3.3 | 0% | - |
| **Ours 6.7** | **9%** | **13** |

可以看出，我们的方法不仅保证了针对多姿态形变、表情、遮挡情况下的低错误率，而且由于我们提出的自适应初始化的策略，使得算法执行效率很高。在Helen数据库上，我们在实验的过程当中，根据不同的迭代次数，可以得到对应的平均误差，为此我们比较了与其它方法的收敛性能，如图5.9所示。



##### 图 5.9 快速收敛示例

Fig 5.9 Illustration of fast convergence

由图5.9可以看出，采用了自适应的初始化策略之后，算法在保证定位精度的前提下有更快的收敛速度，这也再次体现了本算法的鲁棒性。Helen数据库图像分辨率很高，并且标注的关键点非常密集，其特征定位也更加精细。但是随着定位点的增多，回归的次数也会增加，从而降低了算法的执行效率，所以在Helen数据库上的帧率远低于在LFPW上的帧率。这也指导我们在实际的工程应用当中，特征点数和速度之间要有一个折中，通常我们会选择十几个特征点作为后续的处理。最后给出算法在不同数据库上的样例检测结果，如图5.10所示。



图 5.10 人脸特征点定位结果示例

Fig 5.10 Example result of the two dataset

## **5.3** 人脸识别结果与分析

### **5.3.1** 实验数据库与设置

**LFW**[61]：LFW（Labeled Faces in the Wild）数据库是由美国马萨诸塞大学（ University of Massachusetts）阿姆斯特分校（ Amherst ）计算机视觉实验室整理完成，对应文章为一篇技术报告，该实验室还共同维护了上文讲到的FDDB人脸检测基准数据库。其主要目的用于研究非受限情形下的人脸识别问题，且现已成为学术界和工业界评价人脸识别性能的基准数据集（benchmark）。LFW包含了13233张自然状态下人脸图片，其大多为名人，包含了5749个人，其中4069个人只有一张图片，1680个人多于一张图片。LFW 的测试任务是判断一对人脸照片是来自同一个人还是不同的人。测试集包含 6000 对人脸图像。对于人脸识别问题，包含人脸确认（Face Verification）和人脸辨识（Face Identification）两个任务，在该数据库上，主要是人脸确认的任务。在我们的实验过程当中，利用该数据库测试了基于KISSME的人脸确认任务，其目的在于验证其度量方式的有效性。LFW数据库样例图片如下图5.11所示。可以看出，该数据库包含了多姿态、表情、光照和部分遮挡情况下的人脸图片，其难点与监控环境下的人脸识别面对的挑战一致。



##### 图 5.11 LFW数据库样例图片

Fig 5.11 Example of LFW dataset

**Pubfig**[94]：PubFig（Public Figures Face Database）数据库是一个真实条件下巨大的人脸数据库，它包含了200个人的58797图片，其中大部分图片都是从网上收集得到。该数据库包含训练集和测试集两部分：其中训练集包含60个人。该部分用来训练算法本身，为了防止过拟合等问题，该数据集与测试集并没有任何重叠和交集。测试集包含了140个人，主要用于测试算法性能。

由于版权问题，该数据库并只开放了这些图片的网络链接，由于图片的链接会慢慢消失，工作人员会定期更新不能用的链接，让其能够近似比较。由于该数据库数量巨大，虽然不能精确比较算法性能，但是在近似情况下，整体的性能与原始的数据集图片并无太大差别。

由于该数据集从真实环境下获得，所以仍然保持着真实条件下姿态、光照、表情等影响因素，因此十分有挑战性。另一方面，在提出该数据库的文章[94]中，提出了高级的人脸属性描述，其中包括了73个人脸属性，例如性别、种族、头发颜色等等。在后续的特征提取实验中，本文也采用了该特征提取的方法，以便统一的比较。该数据库示例图片如图5.12所示。

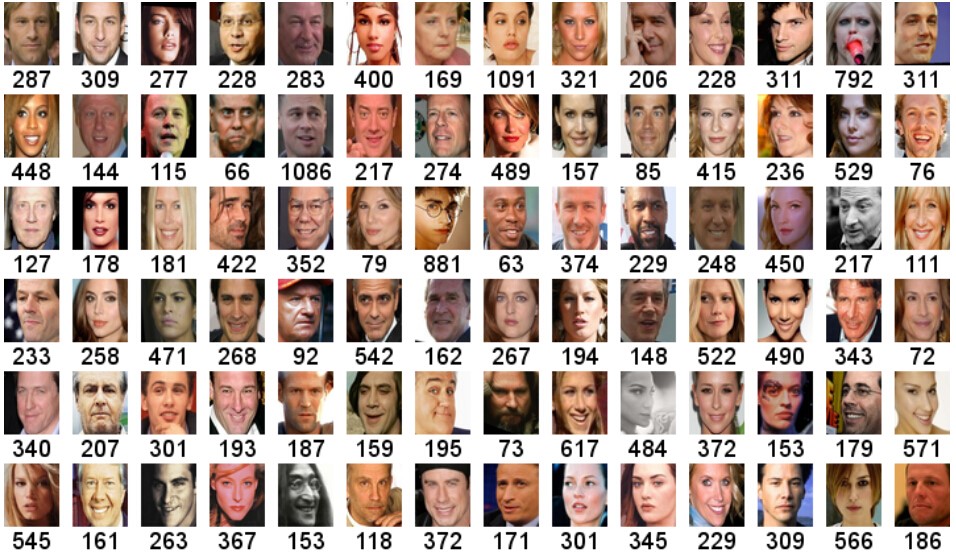


图 5.12 PubFig数据库样例图片(图片下面数字代表每个人的图片个数)

Fig 5.12 Example of PubFig dataset (Below each image is the total number of images for that person)

### **5.3.2** 性能评价准则

上文提到，人脸识别分为人脸确认（Face Verification）和人脸辨识（Face Identification）两种任务。人脸确认判断两张人脸图像是否属于同一个人，本质上是属于一个二分类问题，所以我们采用ROC曲线来评价人脸确认任务的性能。关于ROC曲线的说明上文已经提到，此部分再不做说明。

人脸辨识是将一张人脸图像分成*N*个身份类别之一，其本质上是属于一个多分类问题，且难度随着*N*变大而增加。2014年，阿尼尔·简(Anil Jain)的研究小组提出了在LFW上人脸辨识任务的测试协议[95]。这是一对多的匹配，比一对一的人脸确认难度高很多。该测试协议包含基于封闭集和开放集两种不同的人脸辨识任务。在封闭集上报告的是 Rank-1辨识率，开放集上报告的是在 1% 的虚警率下 Rank-1 的检测和辨识率

(DIR1% FAR)。

由于本文方法采用了多任务度量的机制，在每个任务上需要同一个人的不止一张的人脸数据作为训练，所以采用了PubFig数据库作为人类辨识的测试数据集。为了在同样数据库上进行比较，我们跟文章 [96]采用了相同的评价机制，即通过召回率曲线（Precision Recall Curve, PRC）的方式，我们对分类器的得分进行了排序，通过不同的阈值得到不同的预测结果。其中，横轴召回表示测试样本被标记（分类）所占的百分比，标记意味着测试样本的分类器得分大于当前设定的阈值（我们认为只有当分类器结果大于指定阈值的时候才指定相应标签），那么纵轴对应的准确率表示被正确分类的样本比例。根据表5.1混淆矩阵的定义，我们可以得到对Recall和Precision的定义，即：

|  |  |
| --- | --- |
| *Recall* = *TP*/*P*, *P* = *TP* + *FN* | (5.2) |
| *Precision* = *TP*/(*TP* + *FP*) | (5.3) |

在这里，Recall代表测试样本中被指定类别标签的百分比，而Precision表示在所有指定类别标签的这些样本中被正确分类的比例。

### **5.3.3** 实验结果和分析

为了证明基于KISSME度量学习的有效性，我们在LFW上数据库上进行了测试，并重现了文献[78]的方法。实验过程中采用了数据库当中的十折交叉验证方式，平均每个集合有包含300个相似对和300个不相似对，最终的结果是交叉验证的平均结果。在实验当中，特征表达采用了文献[97]所采用的方法。

首先，我们根据人脸特征点检测之后的6个特征点（嘴角2个，眼睛3个，鼻子1个）在三个尺度上提取SIFT[23]特征。相应的特征维数是3456维。为了便于计算，我们利用PCA将特征维度降至100维的子空间。为了保证公平性，也验证了未降维的2456维特征，可以发现PCA降维对其影响较小。只有通过线性的SVM方法去分类受部分影响。在图5.13当中，我们与常用的度量学习方法进行了比较，如LDML[97]，ITML[98]，

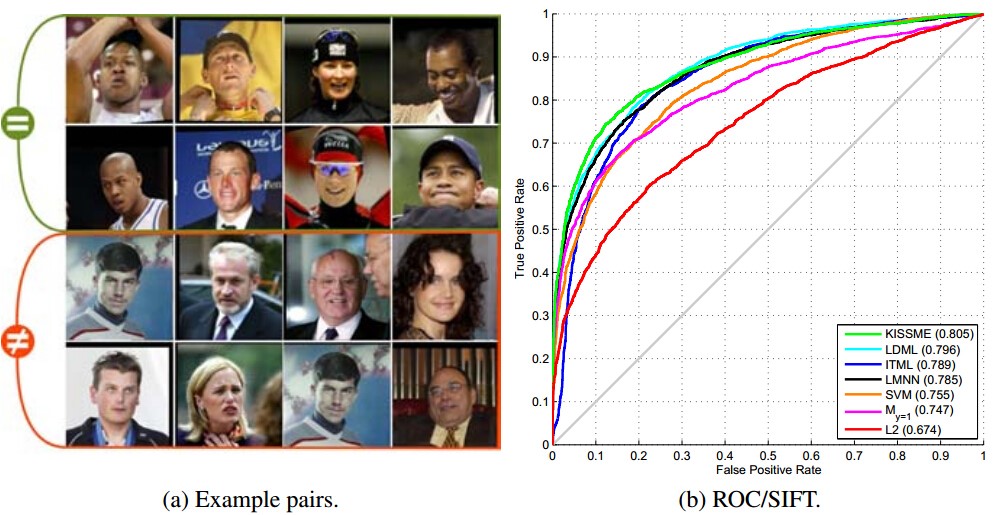


图 5.13 LFW数据库KISSME人脸确认结果. (a) 相似对和不相似对示例 (b) 不同方法对应的ROC曲线

Fig 5.13 Face verification results on the LFW dataset. (a) Examples of similar and dissimilar pairs (b)

ROC curves for different method.

LMNN[70]，SVM[99]以及马氏距离、欧式距离作为基准进行比较。可以发现，本方法比其它度量学习方法性能稍微提升。为了进一步测试，我们在PubFig上数据库上进行了测试，其测试数据也基于十折交叉验证，每个集合包含了1000个相似对和1000个不相似对，每个集合都来自14个不同的人。并且测试集当中没有出现训练集当中的人。其中采取了该数据库所在文章提供的“high-level”的属性特征。同样，我们与LDML[97]，ITML[98]，LMNN[70]，

SVM[99]以及两个基准方法进行比较，其效果图5.14所示。可以看出，基于KISSME的性能整体上仍然是最好的。

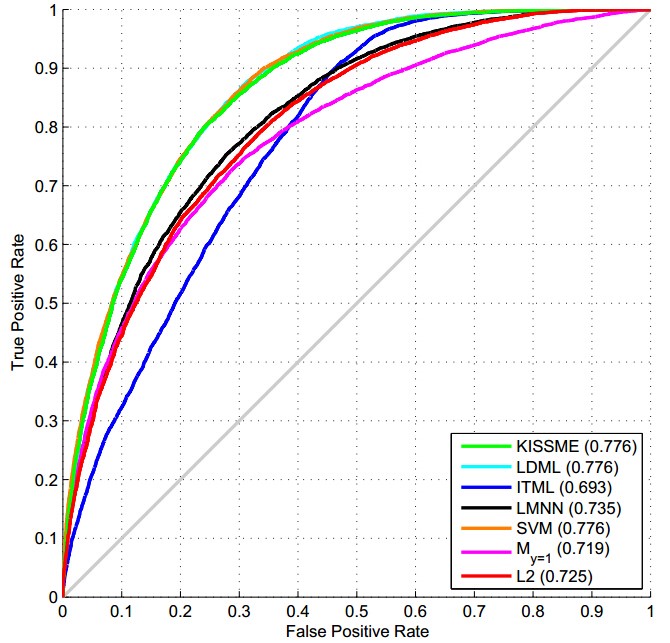


图 5.14 PubFig数据库KISSME人脸确认结果

Fig 5.14 Face verification results using KISSME on the PubFig dataset

KISSME度量学习方法最大的特定是训练时间短，执行效率高。上述实验证明了其识别的准确率，可以看出相比其它方法虽然都有一定程度的提升，虽然与一些方法的提升效果不是很明显，但是算法执行效率却相差巨大。为了验证其执行效率，我们比较了不同度量学习方法的训练时间，其测试机型为3.06GHZ Xeon 24核服务器。如表5.6所示。由表5.6可以看出，KISSME方法的训练时间非常快，这得益于该方法不需

##### 表 5.6 平均训练时间

Table 5.6 Average training times

**Dataset**

**Method**

**KISSME**

**SVM**

**ITML**

**LDML**

**LMNN**

**LFW 0.05s** 12.78s 24.81s 307.23s 1198.69s **PubFig 0.07s** 0.84s 20.82 2868.91s 783.66s

要额外的优化，而传统的度量学习方法将问题转化为可解的优化问题，但计算过程迭代次数较多，计算缓慢。相反，KISSME方法具有简单、快速、有效的特点避免了复杂度高的迭代过程，同时性能优于ITML、 LMNN等方法。该结果与文章[78]所示的结果基本一致。这也为我们将该方法拓展到两级的度量学习任务用于人脸辨识任务提供了基础。

为了证明我们提出的两级度量学习任务的有效性，我们在PubFig数据库上进行了人脸辨识任务的验证。首先，在第一级的任务中，我们分期提取了LBP[22]、SIFT[23]和高级的属性特征[94]。按照评价准则及测试标准，我们首先跟单任务的KISSME方法以及多任务的KISSME方法进行了比较。如图5.15 (a)所示，我们的方法（TMTL)在全召回率的情况下得到了71.2%的准确率。并且证明了多任务的执行效果比单任务的效果更好。

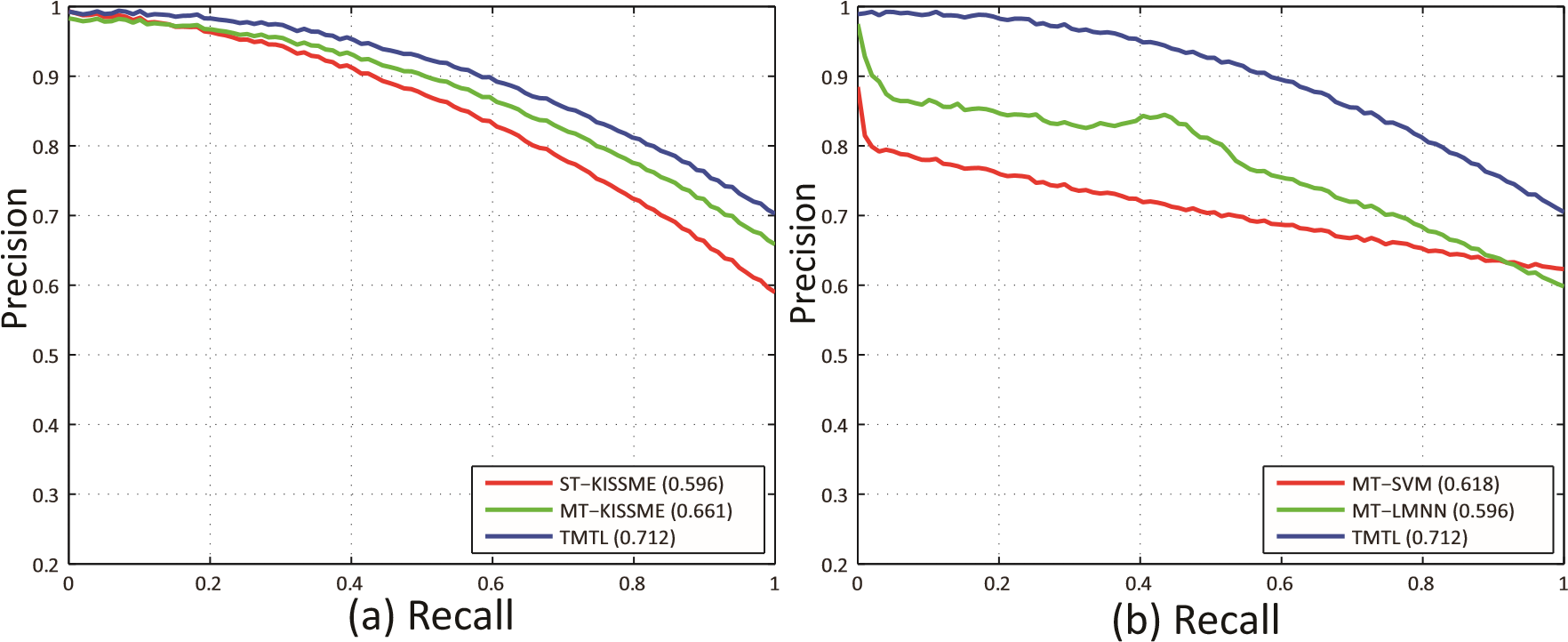


图 5.15 TMTL与(a)单任务KISSME和MT-KISSME， (b)其它多任务方法比较

Fig 5.15 Comparison with (a) single-task KISSME and MT-KISSME, (b) other MTL methods

紧接着，我们与其它多任务的度量学习方法MT-LMNN[71]和多分类SVM[99]进行了比较。由图5.15 (b)可以看出，我们的方法比多分类SVM（one-vs-all）和MT-LMNN分别高9.4%和10.6%。这也证明了我们的两级度量方法从相似对以及特定任务的学习中获取了更多的信息，从而提高了识别率。

## **5.4** 本章小结

本章主要对人脸检测、人脸特征点定位、人脸识别这三个任务进行了实验验证。其中人脸检测验证了基于DPM方法的人脸检测效果，实验得出该方法可以实时的鲁棒的检测出多视角的人脸。在人脸特征点定位当中，验证了两级级联回归的特征点定位方法，并验证了我们提出的自适应初始化方法的有效性，为实时的特征点定位提供了保证，实验得出，该方法可以实时准确的定位出人脸特征点的位置，为后续人脸识别任务提供了有效的归一化操作。在人脸识别任务中有两部分内容，针对人脸确认任务，本章验证了基于KISSME方法在人脸确认任务当中的准确性和高效性，并分别在两个真实数据库下进行了实验。针对人脸辨识任务，验证了本文提出的两级多任务度量学习（TMTL）的有效性，实验证明，本方法在人脸辨识任务中取得了比单任务学习及其它多任务度量方法更好的性能。

第六章 结论与展望

# 第六章 结论与展望

（1）研究工作总结

本文主要的研究工作是围绕人脸问题展开，针对人脸技术在智能监控环境下应用问题，分别对人脸检测、人脸特征点检测（人脸配准）、人脸识别（人脸辨识和人脸确认）这三个任务进行了研究和应用。人脸检测和识别是计算机视觉中一个广泛研究的问题，其主要困难在于人脸随光照、姿态、表情、角度等因素的变化而变化。而在监控系统当中，由于摄像机本身架设角度的问题，使得人脸姿态形变问题更加严重。因此，本文主要通过“三步走”的研究方向，围绕这一问题做出了以下贡献：

第一，针对智能监控场景下多姿态人脸检测的问题，本课题采用了一种基于可变型部件模型（DPM）的人脸检测方法。本文将目标检测领域的可变型部件模型应用到多姿态的人脸检测中，有效的解决了人脸因姿态变化的检测问题。实验表明，该方法不但能够实时准确的检测出多姿态的人脸，而且可以有效的防止误检，这大大提高了在实际系统的应用性，因为在一个实际系统当中，误检率是制约系统性能的重要指标之一。此外，由于采用了FHOG特征，该特征对光照、表情等有一定的鲁棒性。

第二，针对人脸检测后的归一化问题，需要对多姿态的人脸进行特征点检测。针对特征点问题当中的形变、遮挡及实时性的要求，本课题提出了一种基于两级级联回归自适应初始化的人脸特征点定位方法。该方法能够对多姿态的人脸实现精准的定位，而且保证在算法快速收敛，实现定位的实时性。首先采用了形状索引特征，并通过相似变换消除全局的旋转和尺度旋转，从而消除了人脸角度和表情带来的影响，而且提取速度极快。之后，采用了两级级联回归模型，回归指的是逐步校准各个基准点的位置，随着阶段的增加，人脸的基准点位置也越来越精确。而通过自适应初始化的策略，一方面可以对人脸的不同姿态进行由粗略到精细的定位，另一方面使得回归模型能够及时的判断初始化失败。从而提高了人脸特征点定位精度和速度。

第三，针对监控环境下人脸识别的问题。本课题提出了一种基于两级的多任务度量学习的人脸识别方法。传统的度量学习方法在应用于人脸识别时会产生过拟合、对图像噪声缺乏稳健性和训练速度慢等问题，且只针对从单一图像中一个特征学习马氏距离的度量，无法表达人脸图像的有效信息。基于此，本课题通过提取多种特征，并针对人脸辨识的问题，将一个大的学习问题分解为多个子任务，挖掘更加有区分性的信息，并在全局优化的条件下学习新的最优度量矩阵，从而提高人脸识别的准确率，并结合监控视频的特点，从而提高监控视频下人脸识别的鲁棒性。以上三个方面的工作相辅相成，在一个完整的人脸识别流程中，每个步骤都对识别效果有非常大的影响。前一步的工作都是为后一步的工作做好基础，即作为预处理。同时，我们在每一步当中，都对人脸姿态、形变等问题进行了考虑，都针对该问题提出了相应的改善措施。总之，所以这些算法都是为得到速度更快、精度更高的人脸检测和人脸识别算法而服务的。

（2）不足和展望

人脸检测和人脸识别一直都是计算机视觉中的研宄热门方向，30多年来，经过大量的研究，人脸识别的精度已经得到了长足的进步。但是，直到今天，人脸识别依旧是个十分具有挑战性的问题，在监控场景下面临的困难更加艰巨，人脸检测和识别的精度与理想的状态还有很大距离。本文虽然在人脸检测和识别算法上提出了一些新方法，也取得了许多新的进展。但是由于作者时间、精力以及研宄水平有限，本文中的工作也存在一些不足之处:

第一，虽然我们采用的人脸检测对角度、表情等具有一定的不变性，但是人脸角度非常大的时候（大于90◦），人脸检测会变得很困难。另外人脸检测越鲁棒，人脸特征点检测面临的困难越大，人脸特征点检测的精度受影响，那么人脸识别任务因为对齐误差造成识别率下降。如何同时提高这三个部分的识别率，是后续要考虑的问题。

第二，在本文的研究方法当中，人脸检测与人脸配准都依赖人脸面部特征，而这个过程是串行的方式解决这两个问题，而两个任务串行带来特征提取的重复及速度的下降，可以进一步考虑融合这两个任务，使得人脸检测和特征点的检测统一起来，可以有效避免人脸检测的鲁棒性带给特征点检测的困难。而融合的思想是基于精确的检测会给配准提供更好的初始化，精确的配准能够提供更严格的人脸误检测删除标准。

最终通过融合检测配准可以简化人脸分析的预处理步骤，加速人脸分析相关的应用。

第三，在本文的研究过程中，主要考虑的是主流的真实场景下人脸检测和识别问题。由于监控设备越来越高清化，本文的研究重点并没有放在低分辨率的人脸检测和识别当中。而低分辨率的人脸检测和识别对一个真实的智能监控场景也至关重要。在智能监控环境下，由于低分辨率人脸识别困难并且人脸在完全侧面或背面时，人脸检测无法实现，所以针对这个问题，研究者越来越关注行人重识别的研究（ReID）。如果结合这两种技术，不但可以弥补因人脸丢失问题带来的信息缺失，而且可以通过已经确认的人脸身份信息给人体加注标签信息。相信未来的结合人脸和人体技术实现智能监控环境下的布控系统是未来智能监控系统的发展方向。

第四，鉴于目前深度学习人脸识别算法取得的重大进展，未来的人脸识别的发展方向肯定是以深度学习为主导的。手工设计的特征是受局限的，如何通过传统浅层的机器学习算法和深度学习算法结合起来也是值得研究的一个方向。

# 参考文献

1. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001: 511–518.
2. I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision (IJCV)*, 2004, 60(2): 135–164.
3. A. M. Burton, S. Wilson, M. Cowan and V. Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 1999, 10(3): 243–248.
4. C. Pagano, E. Granger, R. Sabourin, G. Marcialis and F. Roli. Adaptive ensembles for face recognition in changing video surveillance environments. *Pattern Recognition (PR)*, 2015, 48(11): 3385–3406.
5. I. Haritaoglu, D. Harwood and L. S. Davis. W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1997: 222–222.
6. Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki and K. Fujimura. The intelligent ASIMO: System overview and integration. In: *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2002: 2478–2483.
7. M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2003: 53–53.
8. K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In: *Eleventh International Symposium on Robotics Research*, 2005: 192–201.
9. W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 2003, 35(4): 399–458.
10. C.-F. Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior and Y. Tian. Ibm smart surveillance system (s3): a open and extensible framework for event based surveillance. In: *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005: 318–323.
11. Z. Wang, Z. Miao and C. Zhang. Extraction of High-Resolution Face Image from Low-Resolution and Variant Illumination Video Sequences. In: *Image and Signal Processing*, 2008: 97–101.
12. S. Z. Li, L. Zhu, Z. Q. Zhang, A. Blake, H. J. Zhang and H. Shum. Statistical Learning of Multi-view Face Detection. In: *European Conference on Computer Vision (ECCV)*, 2002: 67–81.
13. B. Wu, H. Ai, C. Huang and S. Lao. Fast rotation invariant multi-view face detection based on real Adaboost. In: *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2004: 79–84.
14. C. Huang, H. Ai, Y. Li and S. Lao. High-performance rotation invariant multiview face detection.

*IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2007, 29(4): 671–86.

1. J. Li, T. Wang and Y. Zhang. Face detection using SURF cascade. In: *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2011: 2183–2190.
2. H. Li, Z. Lin, J. Brandt, X. Shen and G. Hua. Efficient boosted exemplar-based face detection. In:

*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014: 1843–1850.

1. C. Zhang and Z. Zhang. A Survey of Recent Advances in Face Detection. 2010.
2. H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision (IJCV)*, 2004, 56(3): 151–177.
3. Y. Li, S. Gong, J. Sherrah and H. Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, 2004, 22(5): 413–427.
4. F. Raphaël, B. Olivier and C. Daniel. A Constrained Generative Model Applied to Face Detection.

*Neural Processing Letters*, 1997, 5(2): 11–19.

1. Z. Q. Zhang, M. J. Li, S. Z. Li and H. J. Zhang. Multi-View Face Detection with FloatBoost. In: *Applications of Computer Vision*, 2002: 184–188.
2. T.Ojala,M.PietikäinenandT.Mäenpää.GrayScaleandRotationInvariantTextureClassificationwith Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002, 24(7): 971–987.
3. D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV)*, 2004, 60(2): 91–110.
4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005: 886–893.
5. C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing (TIP)*, 2002, 11(4): 467–76.
6. Z. Cao, Q. Yin, X. Tang and J. Sun. Face recognition with learning-based descriptor. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010: 2707–2714.
7. Y. Taigman, M. Yang, M. Ranzato and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014: 1701–1708.
8. Y. Sun, X. Wang and X. Tang. Deep Learning Face Representation from Predicting 10,000 Classes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014: 1891–1898.
9. F. Schroff, D. Kalenichenko and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015: 815– 823.
10. R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In: *IEEE International Conference on Image Processing (ICIP)*, 2002: 900–903.
11. S. Marcel, Y. Rodriguez and G. Heusch. On the Recent Use of Local Binary Patterns for Face Authentication (accepted for publication but withdrawn because of author charges). *IDIAP*, 2006:

6–34.

1. P. Felzenszwalb, D. Mcallester and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008: 1–8.
2. P. F. Felzenszwalb, R. B. Girshick, M. A. David and R. Deva. Object detection with discriminatively trainedpart-basedmodels.*IEEETransactionsonPatternAnalysisandMachineIntelligence(TPAMI)*, 2010, 32(9): 1627–1645.
3. P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 2004, 8(19): 415–428.
4. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision (IJCV)*, 2005, 61(1): 55–79.
5. T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, 1995, 61(1): 38–59.
6. T. F. Cootes, G. J. Edwards and C. J. Taylor. Active appearance models. 1998: 484–498.
7. T. F. Cootes, G. J. Edwards and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2001, (6): 681–685.
8. H. Azizpour and I. Laptev. Object Detection Using Strongly-Supervised Deformable Part Models. In: *European Conference on Computer Vision (ECCV)*, 2012: 836–849.
9. P. Dollar, P. Welinder and P. Perona. Cascaded pose regression. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010: 1078–1085.
10. X.Cao,Y.Wei,F.WenandJ.Sun.FacealignmentbyExplicitShapeRegression.In:*IEEEConference on Computer Vision and Pattern Recognition (CVPR)*, 2012: 2887–2894.
11. X. P. Burgosartizzu, P. Perona and P. Dollar. Robust Face Landmark Estimation under Occlusion. In: *IEEE International Conference on Computer Vision (ICCV)*, 2013: 1513–1520.
12. J. Gower. Procrustes Analysis. *International Encyclopedia of the Social and Behavioral Sciences*, 2001, 39(4): 12141–12143.
13. S. Sclaroff and J. Isidoro. Active Blobs. In: *IEEE International Conference on Computer Vision (ICCV)*, 1998: 1146–1146.
14. S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision (IJCV)*, 2004, 56(3): 221–255.
15. Y. Sun, X. Wang and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013: 3476–3483.
16. N. Duffy and D. Helmbold. Boosting Methods for Regression. *Machine Learning*, 2002, 47(2): 153– 200.
17. H. Li, G. Hua, Z. Lin, J. Brandt and J. Yang. Probabilistic Elastic Part Model for Unsupervised Face Detector Adaptation. In: *IEEE International Conference on Computer Vision (ICCV)*, 2013: 793–800.
18. G. Zhenhua, Z. Lei and Z. David. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing (TIP)*, 2010, 19(6): 1657–63.
19. X. Xiong and F. D. L. Torre. Supervised Descent Method and Its Applications to Face Alignment.

In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013: 532–539.

1. S. Ren, X. Cao, Y. Wei and J. Sun. Face Alignment at 3000 FPS via Regressing Local Binary Features.

In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014: 1685–1692.

1. M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 1991, 3(1): 71–86.
2. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1997, 19(7): 711–720.
3. X. He, S. Yan, Y. Hu, P. Niyogi and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2005, 27(3): 328–340.
4. D. Zhang and Z.-H. Zhou. (2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, 2005, 69(1): 224–231.
5. M. Li and B. Yuan. 2D-LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, 2005, 26(5): 527–532.
6. K. I. Kim, K. Jung and H. J. Kim. Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters (SPL)*, 2002, 9(2): 40–42.
7. A. F. Abate, M. Nappi, D. Riccio and G. Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 2007, 28(14): 1885–1906.
8. O. M. Parkhi, A. Vedaldi and A. Zisserman. Deep face recognition. *The British Machine Vision Conference (BMVC)*, 2015, 1(3): 6.
9. Y. Sun, Y. Chen, X. Wang and X. Tang. Deep learning face representation by joint identificationverification. In: *Advances in Neural Information Processing Systems (NIPS)*, 2014: 1988–1996.
10. G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2007.
11. T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967, 13(1): 21–27.
12. S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory (TIT)*, 1982, 28(2): 129–137.
13. J. Friedman. Flexible metric nearest neighbor classification. 1994.
14. T. Hastie and R. Tibshirani. Discriminant Adaptive Nearest Neighbor Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1996, 18(6): 607–616.
15. J. Baxter and P. Bartlett. The Canonical Distortion Measure in Feature Space and 1-NN Classification. In: *Advances in Neural Information Processing Systems (NIPS)*, 1998: 245–251.
16. E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell. Distance Metric Learning, With Application To Clustering With Side-Information. *Advances in Neural Information Processing Systems (NIPS)*, 2002, 15: 505–512.
17. A. Bellet, A. Habrard and M. Sebban. A survey on metric learning for feature vectors and structured data. *Computing Research Repository (CoRR)*, 2013.
18. S. Gong, M. Cristani, S. Yan and C. L. Chen. Person Re-Identification. *Visual Analysis of Behaviour*, 2013, 42(7): 301–313.
19. K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 2009, 10: 207–244.
20. S. Parameswaran and K. Q. Weinberger. Large margin multi-task metric learning. In: *Advances in Neural Information Processing Systems (NIPS)*, 2010: 1867–1875.
21. Z. Cui, W. Li, D. Xu, S. Shan and X. Chen. Fusing Robust Face Region Descriptors via Multiple

Metric Learning for Face Recognition in the Wild. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013: 3554–3561.

1. F. Yun, Y. Shuicheng and T. S. Huang. Correlation metric for generalized feature extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008, 30(12): 2229–2235.
2. M. Guillaumin, J. Verbeek and C. Schmid. Multiple Instance Metric Learning from Automatically Labeled Bags of Faces. In: European Conference on Computer Vision (ECCV), 2010: 634–647.
3. M. Guillaumin, J. Verbeek and C. Schmid. Is that you? Metric learning approaches for face identification. *IEEE International Conference on Computer Vision (ICCV)*, 2009, 30(2): 498–505.
4. H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In: *Asian Conference on Computer Vision (ACCV)*, 2011: 709–720.
5. Q. Cao, Y. Ying and P. Li. Similarity Metric Learning for Face Recognition. In: *International Conference on Computer Vision (CVPR)*, 2013: 2408–2415.
6. M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth and H. Bischof. Large scale metric learning from equivalence constraints. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012-06: 2288–2295.
7. K. Huang, Y. Ying and C. Campbell. Generalized sparse metric learning with relative comparisons.

*Knowledge and Information Systems*, 2011, 28(1): 25–45.

1. C. Shen, J. Kim, L. Wang and A. Hengel. Positive semidefinite metric learning with boosting. In: *Advances in Neural Information Processing Systems (NIPS)*, 2009: 1651–1659.
2. J. Friedman. Another approach to polychotomous classification. *Technical report, Department of Statistics, Stanford University*, 1996.
3. V. Jain and E. Learned-Miller. FDDB: A Benchmark for Face Detection in Unconstrained Settings. *Technical report, University of Massachusetts, Amherst*, 2010, (UM-CS-2010-009).
4. M. H. Zweig and . Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 1993, 39(4): 561–77.
5. J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013: 3468–3475.
6. D. Chen, S. Ren, Y. Wei, X. Cao and J. Sun. Joint cascade face detection and alignment. In: *European Conference on Computer Vision (ECCV)*. Springer, 2014: 109–122.
7. B.Yang,J.Yan,Z.LeiandS.Z.Li.Convolutionalchannelfeatures.In:*IEEEInternationalConference on Computer Vision (ICCV)*, 2015: 82–90.
8. R. Ranjan, V. M. Patel and R. Chellappa. A deep pyramid deformable part model for face detection.

In: *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2015: 1–8.

1. V. Kumar, A. Namboodiri and C. V. Jawahar. Visual Phrases for Exemplar Face Detection. In: *IEEE International Conference on Computer Vision (ICCV)*, 2015.
2. X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012: 2879–2886.
3. X. Shen, Z. Lin, J. Brandt and Y. Wu. Detecting and aligning faces by image retrieval. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013: 3460–3467.
4. V. Le, J. Brandt, Z. Lin, L. Bourdev and T. Huang. Interactive facial feature localization. In: *European Conference on Computer Vision (ECCV)*, 2012: 679–692.
5. P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman and N. Kumar. Localizing parts of faces using a consensus of exemplars. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011: 545–552.
6. M. Dantone, J. Gall, G. Fanelli and L. Van Gool. Real-time facial feature detection using conditional regression forests. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012: 2578–2585.
7. N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In: *IEEE International Conference on Computer Vision (ICCV)*, 2009-10: 365–372.
8. L. Best-Rowden, H. Hu, C. Otto, B. F. Klare and A. K. Jain. Unconstrained Face Recognition: IdentifyingaPersonofInterestfromaMediaCollection.*IEEETransactionsonInformationForensics and Security (TIFS)*, 2014, 9(12): 2144–2157.
9. M. Köstinger, P. Roth and H. Bischof. Synergy-based learning of facial identity. 2012: 195–204.
10. M. Guillaumin, J. Verbeek and C. Schmid. Is that you? Metric learning approaches for face identification. In: *IEEE International Conference on Computer Vision (ICCV)*, 2009-09: 498–505.
11. V. D. Jason, K. Brian, J. Prateek, S. Suvrit and I. S. D. Information-theoretic metric learning. In: *International Conference on Machine Learning (ICML)*, 2007-06: 209–216.
12. C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011, 2: 27.

攻读硕士学位期间的科研成果

# 攻读硕士学位期间的科研成果

已发表论文

1. Hong Liu, **Xuewu Zhang** and Pingping Wu. Two-level multi-task metric learning with application to multi-classification. *IEEE International Conference on Image Processing (ICIP)*, pp. 2756-2760, September 27-30, Quebec, Canada, 2015. (EI, ISTP)
2. Hong Liu, **Xuewu Zhang** and Pingping Wu. Regression based landmark estimation and multi-feature fusion for visual speech recognition. *IEEE International Conference on Image Processing (ICIP)*, pp. 808-812, September 27-30, Quebec, Canada, 2015. (EI,

ISTP)

1. Pingping Wu, Hong Liu and **Xuewu Zhang**. Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1249-1253, May 4-9,

Florence, Italy, 2014. (EI, ISTP)

1. Pingping Wu, Hong Liu\*, Xiaofei Li, Ting Fan and **Xuewu Zhang**. A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion. *IEEE Transactions on Multimedia (TMM)*, 18:3, pp. 326-338, 2016. (SCI, JCR一区).
2. 张学武, 吴平平, 刘宏. 基于视频序列的实时人脸性别识别.《华中科技大学学报

（自然科学版）》, 41(S1):116-120, 2013. (EI)

已接收论文

1. Pingping Wu, Hong Liu\*, **Xuewu Zhang** and Yuan Gao. Spontaneous versus Posed Smile Recognition via Region-specific Texture Descriptor and Geometric Facial Dynamics. *Frontiers of Information Technology & Electronic Engineering*, 2016. (SCI)

已投稿论文

1. Pingping Wu, Hong Liu\*, **Xuewu Zhang** and Yuan Gao. How Do You Smile? Towards

A Comprehensive Smile Analysis System. **Submitted** to NeuroComputing, 2016. (二审中)

参与项目

1. 深圳市战略新兴产业项目：网络环境下的智能监控系统公共技术服务平台
2. 深圳文化创意产业发展专项：机器人剧场-机器人演出关键技术及示范平台
3. 科技部创新人才推进计划项目：面向复杂场景人机交互的嵌入式仿生视觉技术与系统

致谢

# 致谢

时光荏苒，转眼间三年的研究生生活即将结束，回首这三年时光，百感交集，历历在目。在这无比充实的研究生生活即将结束之际，在此谨向无数关心、爱护、支持和帮助过我的老师、亲人、朋友们致以最衷心的感谢和诚挚的问候。

北大是我的起点，也是我一生的荣誉。在这三年的研究生生涯中，首先特别感谢的是我的导师刘宏教授。古云：师者，传道、授业、解惑也！刘老师不仅给实验室所有人提供了良好的科研条件，更给每个人更多自由成长的空间。刘老师不仅在学习和生活中给了我极大的帮助和鼓励，更在价值观、思想认识提高方面对我有着深刻的影响。每周的讨论班，他几乎从未缺席，而且都会给每个人细心的指导，大到研究方向、思路，小到ppt的排版、展示，一切与科研相关的东西，都有着近乎苛刻的要求。在每次的投稿当中，刘老师不厌其烦地修改着论文，邮件当中的几十个论文版本便是最好的见证。论文的结构框架、创新点的提取、实验的对比是否完善合理、算法复杂度、参考文献排版是否标准、定冠词和不定冠词的应用是否准确、标点符合是否恰当等等，所有每一个细节都透漏着刘老师严谨的学术态度以及高标准且极富耐心的做事风格。在HRI实验室大家都认可的一句话是：你可能没有见过凌晨四点的洛杉矶，但是你肯定见过凌晨四点刘老师为我们改论文时的灯光明亮的办公室。不光如此，刘老师给予了我们无数的机会去锻炼各方面的能力，不断挖掘自己的潜力，并以一个严格的北大人的标准要求着我们，影响着我们，也让我们明白，北大不是光环、不是标签，而是一份责任、一种标准。刘老师认真、精益求精、执着坚持的做事风格深深的影响着每一个人。从此，高标准、严要求已然成为实验室的新常态，顶天立地的培养模式已经深深的印刻在每一个人的心中。

感谢丁润伟师姐这几年来的关心和照顾。虽然您是实验室的老师，但是我们还是喜欢亲切的叫您师姐。因为您，实验室变得更加温暖。丁师姐每天处理着实验室的琐碎，感谢您的付出，您辛苦了！很荣幸见证了您和姐夫的婚礼，还有幸能够在毕业的时候看到您的宝宝的降临，衷心祝福他（她）茁壮成长。

感谢师兄师姐们给予我生活和学习上极大的指导和帮助，他们是：顾照鹏、赵文勇、萧伟、张佳佳、吴平平、李哲媛、王灿、刘梦源、孙煜科、张结、陈鹏进、高远、张巧夺、肖芳、翟森、陈玲。特别感谢吴平平师姐，感谢师姐一直以来对我的帮助和关心，也很感谢你对我科研的协助与支持。还有感谢同级入学的你们：马里千、康日生、唐浩、陈月朝、庞程、田璐、钟铖铖，大家一起努力，为梦想而奋斗。感谢实验室的各位师弟师妹们：胡靓、王芷、岳梦荻、孙淼、陶树宁、孙小虎，以及研一的黄伟波、黄伟鹏、关乔、金永庆、何侵嵚、马佳瑶、杨冰，还有研零的涂涓辉、刘畅、赵晨阳、董耀凤、石伟、兰海鹏从你们身上我学到了很多优点，也看到了青春活力，未来实验室顶天立地还得你们努力！

深深感谢我的父亲母亲及家人，你们永远是我的坚强后盾和避风港湾，感谢你们对我的默默支持和付出，并给予我信心和力量，使我不断努力，做更好的自己。同样感谢生命旅程中的挫折与艰难，使我更加坚强，不断成长！最后，谨以此文献给所有关心和帮助过我的人！

感谢参加论文评审和答辩的各位老师对本文的批评指正！

1. . <http://animetrics.com/forensicagps/> [↑](#footnote-ref-1)
2. . <http://isscctv.com/products/face/> [↑](#footnote-ref-2)
3. . <http://www.tm-group.com/products/nice-security-portfolio/nice-vision/> [↑](#footnote-ref-3)
4. . <http://www.cognitec.com/facevacs-videoscan.html> [↑](#footnote-ref-4)
5. . <http://video.boschsecurity.com/category/intelligent-video-analysis/12> [↑](#footnote-ref-5)
6. . <http://www.sdscp.com/> [↑](#footnote-ref-6)
7. π

   这里统一使用*B*来表示*B*1和*B*2。在图2.8中，如果计算的是无向直方图(Contrast Insensitive)，则0◦ − 360◦被平均分成了8份(*p* = 8)，0◦ − 22.5◦和180◦ − 202.5◦可以被视为一个通道，以此类推，每个通道占用45◦。如果计算的是有向直方图(Contrast Sensitive)，则0◦ − 360◦被平均分了16份(*p* = 16)，每个通道占用22.5◦。根据以上分类，可以得到位 [↑](#footnote-ref-7)