

第 5 章



图像数据标注

在数据标注领域中,涉及最多的就是图像标注。图像标注也在人工智能与各行各业实际应用相结合的演变过程中扮演着重要的角色,是最简单、最常用的数据标注类型,非常直观且容易上手,因此大多数数据标注员都是从图像标注开始接触标注行业的。图像标注目的是为了使计算机更好地识别图像,主要手段是把标注好的数据给计算机,告诉它这是什么以及附带的相关信息。后面再依靠不断的强化学习,计算机算法就能够把未标注的图像描述出图像内容,从而对图像进行处理。

5.1 图像数据特征和分析

5.1.1 图像标注的定义

图像标注的本质是一个将标签添加到图像上,将视觉变成文本的过程。计算机根据图像里面的信息自动生成相对应的描述文字,是自然语言与计算机视觉领域的结合。其目标范围既可以是在整个图像上仅使用一个标签,也可以是在某个图像内的各组像素中配上多个标签,这就好像我们小时候在做看图说话题目一样。我们也希望算法能够根据图像而描绘其内容含义的自然语句和自然语言。这项工作同时涵盖图像理解和语言生成,这些工作对于人类小朋友来说较为容易,但对于计算机视觉领域来讲,是一个不小的挑战。随着深度学习概念的到来,图像标注技术的发展,实现了两种不同形式的图像信息到文本信息之间进行转换。为人工智能的发展提供了有力而坚实的后盾,实现了 AI 里程碑式的发展。

5.1.2 图像处理的应用领域

通过对路况图片中的汽车和行人进行筛选、分类、标框等,可以提高安防摄像头以及

无人驾驶的识别能力；通过对医疗影像中的骨骼进行描点，特别是对病理切片进行标注分析，能够帮助 AI 提前预测各种疾病；对街景中红绿灯、车辆、高架桥等道路标注的画框标注，可用于帮助自动驾驶和测量识别道路物体；对人脸图像做描点处理，能帮助人工智能告知是不同个体等，这些领域均为图像标注的常见应用。主要用于为计算机视觉的相关算法提供数据集，下面取其中三个最常见的领域具体介绍。

1. 人脸识别

人脸识别是利用人的脸部特征信息来进行身份识别的一种生物识别技术。采用摄像机或摄像头获取到含有人脸的图像或视频流，并在图像中检测和跟踪人脸进而将检测到的人脸进行信息识别的一系列相关技术，通常也叫作人像识别或面部识别。人脸识别由于其采集成本低、识别效率高，在市场中占据重要位置。

传统的人脸识别技术主要是通过可见光图像来进行的，这也是人们熟悉的识别方式，已有 30 多年的研究历史。但这种方式有着难以克服的缺陷，特别是在环境光照发生变化时，识别效果会急剧下降，无法满足实际系统的需要。能解决光照问题的方案有三维图像人脸识别和热成像人脸识别。可这两种技术还有很多需要探索的地方远不成熟，识别效果不尽如人意。迅速发展起来的另一种解决方案是采用主动近红外图像的多光源人脸识别技术。它能够克服光线变化的影响，已经取得了卓越的识别性能，在精度、稳定性和速度方面的整体系统性能超过三维图像人脸识别。这项技术在近两三年迅速发展，使人脸识别技术逐渐走向实用化。

人体的各个生理特征都是与生俱来的，它的唯一性和不易被复制的优势特性为身份鉴别提供了必要的前提。与其他类型的生物特征识别比较，人脸识别具有非强制性、非接触性和并发性、操作简单、结果直观、隐蔽性好等特点。其实人脸识别技术是通过将人脸信息进行多项分析处理得到的结果，其过程中有 6 个关键的步骤。

1) 人脸检测

人脸检测是为了寻找图片中人脸的位置。当有人脸出现在图片中时，会将人脸框选出来，并给出一个坐标信息，或者将人脸切割出来，检测并定位出图片中的人脸特征关键点。

可以使用方向梯度直方图(HOG)来检测人脸位置，具体方法是先将图片灰度化，接着计算图像中像素的梯度。通过将图像转变成 HOG 形式，就可以获得人脸位置。

2) 人脸比对

人脸比对是把不同角度的人脸图像对齐成同一种标准的形状，如图 5-1 所示提取不同图片或视频中人脸的特征，比对人脸特征的相似度，从而判断是否同一个人，并给出相似度数值。

因为在不同时间获取的图片，不可能完全一致。在身份验证、小区门禁系统、闸机智能验证、金融业认证、App 人脸登录、企业刷脸考勤等场景中都需要用到。

3) 人脸属性分析

定位并标记出图片中人脸的位置，并且可对人脸中的耳、口、鼻、眼、眉等多个特征点进行定位，还可识别出多种人脸属性，如年龄、性别、表情等。

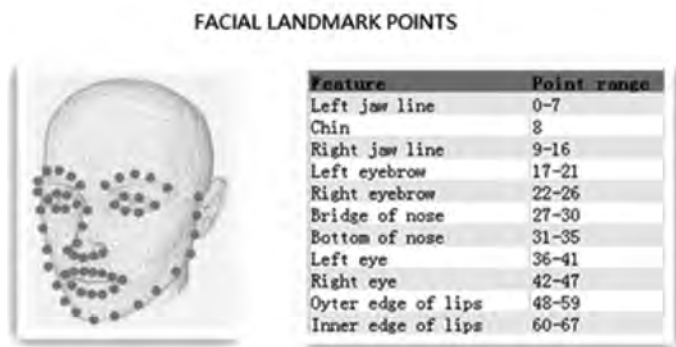


图 5-1 一种人脸特征点位定义(见彩插)

4) 人脸活体检测

RGB 静默、IR 双目、3D 结构光等多种活体检测算法,可拒绝照片、视频、人脸模具等假人脸作弊攻击,保障人脸识别系统的安全稳定运行。

5) 人脸搜索视频比对

从视频中精确搜索给定的图片,可广泛用于公安、工厂、小区的安防布控场景,从海量视频中快速定位嫌疑人或特定人群。

6) 人脸识别

检测视频和图片中的人形、人头、五官、四肢等属性,应用于人脸识别抓拍机、摄像机等产品中,采集点如图 5-2 所示。

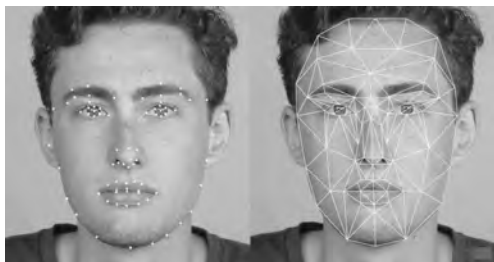


图 5-2 人脸识别头像标注(见彩插)

作为一项飞速发展的技术,人脸识别凭借着出色的运用优势,完全吻合物联网时代人机交互的所需条件。人脸识别的市场十分广阔,许多新兴应用场景处于开发当中。目前主要应用领域包括金融、安防、交通、教育以及智能家居等。现阶段市场主流的人脸识别应用场景具体有如下几个。

(1) 金融行业。

人脸识别技术在金融领域主要作为身份核验的工具,已广泛运用在互联网金融、商业银行等领域。各类商业银行目前均已配备人脸识别机器用于客户开户、存取款等业务。随着互联网金融渗透率不断提高,手机、网银支付逐渐成为日常生活支付的主流方式,其主要身份识别方法即人脸识别技术。

(2) 安检领域。

目前,智能安防发展得如火如荼,为了进一步提升安防应用的适用性,提高数据处理的速度与效率,推动安防从被动防御向主动预警发展,对数据标注的需求与日俱增,这也是由于安检安防领域是人脸识别系统去支撑。目前在全国主要城市的火车站、飞机场等人员进出通道均已搭配人脸识别技术,用来对乘客个人身份信息的验证。除此之外,交通管理部门也将运用人脸识别技术对于行人的违法违规行进行监督管理,提高城市管理效率,主要应用场景如图 5-3 所示。



图 5-3 安检入口处的人脸识别闸机通道

(3) 智能家居。

智能家居领域作为新兴市场,发展十分迅速,具体应用场景有智能人脸识别锁、扫地机器人、智能冰箱、智能咖啡机等。人脸识别作为未来 AIOT 领域的主要交互方式之一,渗透率逐渐提高。通过配备人脸识别技术,智能家居可以变得更加便利与安全,以满足万物智联的需求。

2. 医疗影像

随着数据的暴增和计算机硬件技术的发展,AI 技术逐渐渗透到各行各业的应用之中。对于医疗行业来说,人工智能 AI 是一种很好的工具,能够缓解医生资源紧缺的问题,提高医生工作效率,可以承担部分影像筛查、分类、预警等前期工作,为紧张的医疗资源带来有效缓冲,让医师解放出时间处理疑难病症。

医院可以利用 AI 进行范围内居民健康管理,通过人工智能 AI 模拟医生诊疗过程并给出诊疗建议,实现人机协智的互联网智慧医疗模式,开展大数据人工智能跨区域、跨学科、多中心的临床科研,加速医疗服务队伍专业化能力的培养。

医疗 AI 最终训练出来的模型好坏取决于医疗数据的来源和质量,模型不是难点,数据才是真正的门槛。目前业界流传着一段话:有多少人工,就有多少智能。因此,人工智能医疗影像标注是实现 AI 医疗的必要环节,影像数据的获取能力与标注能力已经成

为 AI 医学影像公司的核心竞争力之一。精准的医疗影像标注,最终将深耕医疗行业,为医疗行业做出贡献。

在医疗领域,主要标注方法有人体标框、3D 画框、骨骼点标记、病历转录等应用,机器学习能够快速完成医学编码和注释,以及在远程医疗、医疗机器人、医疗影像、药物挖掘等场景的应用,助力于提供更高效率的诊断与治疗,制订更为健全的医疗保险计划。

现如今医疗影像生成技术还不够成熟,且为专项领域,进入门槛比较高,因此做影像标注的大多数是专业医生。影像标注与目标检测拉框标注所采用的方法比较类似,不过标注起来需要结合较为严谨的专业医学知识,其对标注准确性的要求非常高。一旦标注出错,会造成十分严重的医疗事故,这便是为何要求只能是医学领域的专业人士来做,例如在职医生和医学研究生。

如果出现了一些较为复杂的医疗影像,如图 5-4 所示的宫颈癌病理切片甚至在三甲医院实习的医生都不能完成。目前获取医疗影像数据主要有两个途径,其一是靠国家政府的公共数据,其二是与医疗机构协同,建立深层合作关系,获得高质量可靠的数据来源。

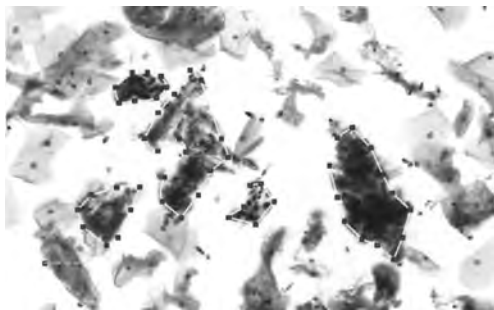


图 5-4 宫颈癌病理切片标注(见彩插)

3. 自动驾驶

自动驾驶系统是一个汇集众多高新技术的综合系统,系统采用先进的通信、计算机、网络和控制技术,对车辆实现实时、连续控制。

无人驾驶汽车要想取得长足的发展,有赖于多方面技术的突破和创新。自动驾驶系统相关的关键技术,包括环境感知、逻辑推理和决策、运动控制、处理器性能等。随着机器视觉、模式识别软件和光达系统的进步,车载计算机可以通过将机器视觉、感应器数据和空间数据相结合来控制汽车的行驶。因此技术的进步为各家汽车厂商自动驾驶的发展奠定了基石。另一方面,普及还存在一些关键技术问题需要解决,包括车辆间的通信协议规范、有人驾驶车辆和无人驾驶车辆共享车道的问题、通用的软件开发平台建立、多种传感器之间信息融合以及视觉算法对环境的适应性问题等。

要完好地实现上述功能,必定需要依赖人工智能的发展,即满足四个基本要素:计算力、海量数据、算法与决策以及传感器的数据采集。而实现完全的自动驾驶则高度依赖于这四个基本要素,并且缺一不可。因此,自动驾驶也被誉为人工智能的终极场景。在汽车自动驾驶的过程中,想要让汽车本身的算法做到处理更多、更复杂的场景,背后就需要有

海量的真实道路数据做支撑,这就需要依靠数据标注。

自动驾驶包含的标注颇为繁多,几乎涵盖了所有车辆和街景的图像信息,其需要标注的内容有:

1) 路面标线类

白色实线、单黄实线、黄色虚线、黄色虚实线、白色虚线、黄色禁止停车线、导流线、禁停网格线、人行道、道路箭头、减速带等。

2) 道路设施类

道路可行驶区域、栏杆、隔离栏、石墩、杆状物(竖直部分)、路边建筑、绿化植被、路灯、红绿灯等。

3) 车辆车型类

非机动车、轿车(小型车辆)、货车(大型车辆)、其他机动车、行人等。

5.1.3 数据质量和标注标准

1. 数据质量标准

人们日常所说的人工智能其实就是一种机器学习,即从数据中自动训练获得规律,并利用规律对未知数据进行处理的过程。如何让机器学习从数据中更准确有效地获得规律,这就是数据标注要做的事情。虽然机器学习领域在算法上取得了重大突破,由浅层学习转变为深度学习,但缺乏高质量的标准数据集已经成为深度学习发展的瓶颈。其影响数据质量的因素主要来源于四个方面:信息因素、技术因素、流程因素和管理因素。

1) 信息因素

产生这部分数据质量问题的原因主要有:源数据描述及理解错误、数据度量的各种性质(如数据源规格不统一)得不到保证和变化频度不恰当等。

2) 技术因素

技术因素主要由于具体数据处理时各环节的异常造成数据质量问题。数据质量问题的产生环节主要包括数据创建、数据获取、数据传输、数据装载、数据使用、数据维护等方面的内容。

3) 流程因素

流程因素是指由于系统作业流程和人工操作流程设置不当造成的数据质量问题,主要来源于系统数据的创建过程、传递过程、装载过程、使用过程、维护过程和稽核过程等各环节。

4) 管理因素

管理因素是指由于人员素质及管理机制方面的原因造成数据质量问题。

如今深度学习算法的训练效果在很大程度上需要依赖高质量的数据集,如果训练中使用的标注数据集存在大量误差,将会导致机器学习训练不充分,无法获得规律,这样在训练效果验证时会出现目标偏离,无法识别的情况。一般的数据在进行机器学习前都要进行加工处理,使数据集的整体质量得到提升,以此来提高算法的训练效果。当数据集的整体标注质量只有80%的时候,机器学习的效果可能只有30%~40%。

随着数据标注质量逐步提高,机器学习的效果也会突飞猛进。但数据标注质量达到90%以上的时候,机器学习效果的提升就没有之前那么明显了,少量的标注数据就无法产生好的效果,需要靠大量的数据进行迭代,才能拥有明显的优化效果。

2. 数据标注标准

对比人眼所见的图像而言,计算机所见的图像只是一堆枯燥的数字。图像标注就是根据需求将这一堆数字划分区域,让计算机在划分出来的区域中找寻数字规律。如图5-5所示场景实际图像转换为如图5-6所示的数字化呈现。



图 5-5 场景实际图像

9	15	6	12	50	90	16	21	15	38	57	7	8	70	20	32	21	34	55	45	67	56	76	2	51	32	8	34	67	34	82	95	7
11	16	7	13	51	91	17	22	16	40	33	3	9	71	21	33	23	25	89	44	68	37	77	3	54	43	9	35	68	55	83	96	64
47	33	44	82	125	48	32	47	71	84	34	40	102	52	64	24	66	120	75	99	33	108	34	85	74	40	66	99	88	114	127	39	90
3	14	19	10	16	34	84	20	23	19	43	56	6	12	74	24	26	26	38	95	42	71	60	99	6	52	11	27	70	31	82	98	6
20	44	57	7	13	23	25	27	27	39	91	48	72	61	31	7	38	47	12	39	72	38	87	100	12	36	12	28	71	38	96	92	11
15	20	11	17	35	95	71	79	10	23	14	20	82	102	28	33	27	31	64	10	16	28	28	40	30	82	96	31	73	84	86	25	28
59	99	25	30	24	48	61	11	17	79	29	41	31	43	97	32	76	65	53	11	65	31	17	43	76	63	91	104	16	80	88	101	13
17	22	13	19	37	97	23	28	27	46	39	9	15	77	27	39	29	41	93	30	14	83	83	9	60	49	15	41	74	61	89	102	14
108	34	39	23	37	70	20	26	88	30	30	40	32	106	61	55	66	46	81	11	55	26	52	95	21	72	61	27	33	86	73	101	114
19	24	13	21	21	26	17	23	61	101	27	23	26	89	63	13	19	81	31	43	32	45	99	24	23	28	19	25	63	103	29	24	24
31	36	30	54	67	17	23	55	35	82	37	49	100	98	59	31	91	17	96	37	23	49	82	12	63	52	15	44	77	84	92	105	12
51	31	22	28	66	106	32	37	31	55	69	15	24	86	26	48	38	50	104	39	16	67	87	27	78	67	33	99	92	19	107	120	32
22	27	18	24	34	46	36	48	102	57	81	70	70	82	22	44	34	46	109	53	19	65	88	14	63	84	20	46	79	66	94	107	19
30	25	26	22	17	41	34	8	14	32	22	24	21	83	23	45	39	47	101	96	30	69	26	13	66	55	21	47	89	67	95	109	20
24	29	20	26	64	104	30	35	29	33	60	16	22	84	24	46	26	48	102	37	31	76	80	16	67	56	22	48	81	65	96	109	21
25	30	21	27	65	105	16	17	8	14	32	92	16	22	17	41	34	8	21	47	30	67	24	38	70	99	25	31	84	69	97	110	22
27	32	23	29	67	107	33	38	32	56	69	19	25	87	27	49	39	31	105	60	84	72	90	19	83	72	100	113	35	74	99	112	24
80	45	69	59	78	4	55	44	10	36	63	36	34	97	21	26	27	23	71	111	27	42	36	80	73	23	22	91	41	53	43	55	89
40	52	42	24	108	63	37	76	36	77	73	62	28	24	87	74	102	115	27	115	41	46	40	84	77	27	32	85	45	37	47	29	113
49	68	92	51	72	119	29	44	28	62	73	29	21	93	43	55	45	57	72	112	38	42	27	61	74	24	30	92	42	54	54	56	661
42	64	88	77	97	23	74	63	29	55	85	75	103	116	28	120	46	31	45	69	52	32	25	100	50	82	52	64	118	72	87	22	24
87	118	44	49	43	67	30	30	26	96	45	80	50	62	116	71	30	84	104	30	31	70	36	65	95	82	110	123	35	76	104	117	29
72	38	29	25	34	25	31	69	105	35	40	34	58	71	21	27	89	3	111	66	90	29	99	25	76	65	31	57	90	77	105	118	30
120	39	30	26	74	114	40	45	39	63	76	26	32	84	44	56	46	38	112	67	91	80	100	26	77	86	72	35	91	78	106	119	21
65	40	31	27	13	27	42	81	121	87	32	46	70	82	99	30	28	45	37	72	112	38	22	63	67	18	29	62	95	82	112	127	33
36	41	65	119	74	98	87	107	53	84	73	39	65	86	55	113	176	38	78	118	44	49	43	67	60	50	36	86	90	44	56	110	63
37	83	123	49	34	48	72	85	35	41	103	45	57	47	59	113	68	92	113	70	94	83	103	29	80	89	35	61	94	31	109	122	34
38	42	34	40	78	80	120	46	51	45	89	82	122	49	53	47	71	84	34	40	102	44	56	46	14	29	32	72	64	118	73	10	72
29	79	119	45	50	44	68	81	21	37	99	41	33	43	55	109	64	88	117	72	96	85	105	31	82	71	37	63	96	83	111	124	36
40	45	56	42	80	81	121	47	32	46	70	83	32	39	101	43	53	45	37	111	66	90	108	22	83	72	33	64	97	54	112	125	37
41	46	82	32	38	100	47	34	59	88	24	29	38	101	51	53	31	32	42	34	108	65	82	98	48	52	47	34	108	86	100	33	79
42	29	42	33	39	77	117	43	48	47	66	79	26	35	97	39	31	41	53	107	65	58	113	67	91	74	94	20	71	60	35	66	33

图 5-6 图像在计算机中的呈现

不过对于深度学习训练图像识别其实是根据像素点进行的,因此图像标注的质量标准也是根据像素点位判定,即标注区域的像素点越接近于实际物的边缘像素点,所标注的质量就越高,随之的标注难度就越大。由于原始图片质量等原因,标注物的边缘可能存在一定数量与实际边缘像素点灰度相似的像素点,这部分像素点会对图像标注产生干扰。面对各类图像标注类型则需要使用不同的检验方式,下面对一些常用的图像标注方式进行说明。

1) 2D 矩形框标注

对于标框标注,标注员应该先对需标记物最边缘像素点进行标绘,然后核查标框的四周能否和标记物最边缘的像素点误差在1个像素以内。如图5-7所示,标框标注的上下左右边框均与图中座椅最边缘像素点误差在1个像素以内,所以这是一张合格的标框标

注图片。

2) 分割标注

与标框标注相比,分割标注质量检验的难度在于标注时需要对标注物的每一个边缘像素点进行检验。如图 5-8 所示,分割标注像素点与被标记物座椅边缘像素点的误差在 1 个像素以内,才能是一张合格的分割标注图片。在其质量检验中需要特别注意检验转折拐角,因为在图像中转折拐角的边缘像素点噪声最大,最容易产生标注误差。



图 5-7 合格标框标注数据



图 5-8 合格分割标注数据(见彩插)

5.1.4 常见的图像文件和标注结果格式

1. 常用图像标注结果格式——XML

XML 又可称为可扩展标记语言,标准通用标记语言的子集,是一种具有结构性的标记语言,用于标记电子文件。XML 的简易性,使得其能够在任何应用程序中读/写数据,这使 XML 很快成为数据交换的唯一公共语言,虽然目前一些应用软件也支持其他格式的数据交换,但不久之后它们都将支持 XML,那就意味着程序可以更容易地与 Windows、macOS、Linux 以及其他平台下产生的信息结合,然后可以很容易加载 XML 数据到程序中并分析它,并以 XML 格式输出结果。

1) XML 的特点

在计算机系统中,标记代表计算机所能理解的信息符号。通过此种标记,计算机之间可以连通处理各类信息,如文章等。XML 可以用来标记数据、定义数据类型,是一种允许用户对自己的标记语言进行定义的源语言。它非常适合万维网传输,采用统一的方法来描述和交换独立于应用程序以及供应商的结构化数据,是 Internet 环境中具有跨平台、依赖于内容特征的技术,也是当今处理分布式结构信息的有效工具。早在 1998 年,W3C 就发布了 XML 1.0 规范,使用它来简化 Internet 的文档信息传输。既然此类格式文件如此广泛,是拥有其何种特点才得到各界开发者的喜爱?

(1) XML 能够从 HTML 中分离数据。

即能够在 HTML 文件之外将数据存储于 XML 文档中,这样可以使开发者专注于用 HTML 做好数据的显示和布局,并使得数据内容变化时不会导致 HTML 文件也需要改

动,从而方便维护页面。XML 也可以把数据以“数据岛”的形式存储在 HTML 页面中,开发者依然能够把精力集中到使用 HTML 格式化和显示数据上。

(2) XML 可成为交换数据的中转站。

对于开发者来说,最耗时间的工作就是在遍布网络的系统之间交换数据,然而计算机系统和数据库系统所存储的数据却有多种形式。基于 XML 可以在不兼容的系统之间交换数据,因此把数据转换为 XML 格式存储将大大减少交换数据时的复杂性,还可以使这些数据能被不同的程序读取。

(3) XML 可应用于 B2B 中。

比如在网络中交换金融信息,目前 XML 已正成为遍布网络的商业系统间交换信息所使用的主要语言,许多与 B2B 有关的完全基于 XML 的应用程序正在开发中。

(4) 利用 XML 能够实现数据共享。

XML 格式数据是以纯文本格式存储,这使得 XML 更易读、更便于记录、更便于调试,使得不同系统、不同程序之间的数据共享变得更加简单。

(5) XML 能够让数据充分利用。

XML 是与软件、硬件和应用程序无关的,数据可以被更多的用户、设备所利用,而不仅限于基于 HTML 标准的浏览器。其他客户端和应用程序可以把 XML 文档作为数据来源处理,就像操作数据库一样,XML 格式的数据能被各式各样的“阅读器”处理。

(6) XML 可用于开发新的语言。

由 XML 发展来的有 WAP 和 WML。WML(Wireless Markup Language,无线标记语言)是用于标记运行于移动设备上的 Internet 程序工具,它就采用了 XML 的标准。

综上所述,XML 使用一个简单而又灵活的标准格式,为基于 Web 的应用提供了一个描述数据和交换数据的有效手段。

2) XML 的规则

XML 文件格式是纯文本格式,在许多方面类似于 HTML。XML 由相应的元素组成,每个 XML 元素包括一个开始标记<,一个结束标记>以及两个标记之间的内容。例如,可以将 XML 元素标记为价格、订单编号或名称。标记是对文档存储格式和逻辑结构的描述。在形式上,标记中可能包括注释、引用、字符数据段、起始标记、结束标记、空元素、文档类型声明(DTD)和序言。

具体规则如下。

(1) 必须有声明语句。

XML 声明是 XML 文档的第一句,其格式如下。

```
<?xml version="1.0" encoding="utf-8"?>
```

(2) 注意大小写。

在 XML 的展示文档中,大小写是有区别的。例如,“A”和“a”是不同的标记。注意在写数据时,前后标记的大小写要保持一致。编写此类文档时最好养成一种习惯,要么全部大写,要么全部小写,或者大写第一个字母,这样可以减少因为大小写不匹配而产生的文档错误。

(3) XML 文档有且只有一个根元素。

良好格式的 XML 文档必须有一个根元素,就是紧接着第一行声明后面所建立的一个元素,其他元素都是这个根元素的子元素,根元素包含文档中其他所有的元素。根元素的起始标记要放在所有其他元素的起始标记之前;根元素的结束标记要放在所有其他元素的结束标记之后。

(4) 属性值使用引号。

在 HTML 代码里面,属性值可以加引号,也可以不加。但是 XML 的规则里面,所有属性值必须加引号(引号不做限制,可以是单引号,也可以是双引号),否则将被视为错误。

(5) 所有的标记必须有相应的结束标记。

在 HTML 中,标记可以不成对出现,而在 XML 中,所有标记必须成对出现,有一个开始标记,就必须有一个结束标记,否则将被视为错误。

(6) 所有的空标记也必须被关闭。

空标记是指标记对之间没有内容的标记,如“”等标记。在 XML 中,规定所有的标记必须有结束标记。

3) XML 的相关标准

虽然 XML 标准本身简单,但与 XML 相关的标准却种类繁多,W3C 制定的相关标准就有二十多个,采用 XML 制定的重要的电子商务标准就有十多个。这一方面说明 XML 确实是一种非常实用的结构化通用标记语言,并且已经得到广泛应用;另一方面,这又为了解这些标准带来一定的困难,除了标准种类繁多外,标准之间通常还互相引用,特别是应用标准,它们的制定不仅使用的是 XML 标准本身,还常常用到了其他很多标准。

XML 标准的体系与 SGML 标准的体系非常相似,XML 相关标准也可分为元语言标准、基础标准、应用标准三个层次。

(1) 元语言标准。

描述的是用来描述标准的元语言。在 XML 标准体系中就是 XML 标准,是整个体系的核心,其他 XML 相关标准都是用它制定的或为其服务的。

(2) 基础标准。

这一层次的标准是为 XML 的进一步实用化制定的标准,规定了采用 XML 制定标准时的一些公用特征、方法或规则。例如,XML Schema 描述了更加严格地定义 XML 文档的方法,以便可以更自动化处理 XML 文档。

1XMLNamespace 用于保证 XML DTD 中名字的一致性,以便不同的 DTD 中的名字在需要时可以合并到一个文档中。

2XSL 是描述 XML 文档样式与转换的一种语言。

3XLink 用来描述 XML 文档中的超链接。

4XPath 描述了定位到 XML 文档结构内部的方法。

5DOM 定义了与平台和语言无关的接口,以便程序和脚本动态访问和修改文档内容、结构及样式等。

(3) 应用标准。

XML 已开始被广泛接受,大量的应用标准,特别是针对因特网的应用标准,纷纷采

用 XML 进行制定。有人甚至认为,XML 标准是因特网时代的 ASCII 标准。在这因特网时代,几乎所有的行业领域都与因特网有关。而它们与因特网发生关系,都必然要有其行业标准,而这些标准往往采用 XML 来制定。如图 5-9 所示为某 XML 文件内容。

```
<folder>xingye-part4-20200702</folder>
<filename>0b398999f8bd49a39d382eb265900b2b_300.jpg</filename>
<path>F:\xingye-part4-20200702\0b398999f8bd49a39d382eb265900b2b_300.jpg</path>
<source>
  <database>Unknown</database>
</source>
<size>
  <width>1280</width>
  <height>720</height>
  <depth>3</depth>
</size>
<segmented>0</segmented>
<object>
  <name>document</name>
  <pose>Unspecified</pose>
  <truncated>0</truncated>
  <difficult>0</difficult>
  <bndbox>
    <xmin>642</xmin>
    <ymin>113</ymin>
    <xmax>1187</xmax>
    <ymax>698</ymax>
  </bndbox>
</object>
</annotation>
```

图 5-9 某图像标注 XML 文件内容

2. 常见图像文件格式

1) BMP 格式

BMP 是 Bitmap(位图)的缩写,是一种与硬件设备无关的图像文件格式,使用非常广,能够被 Windows 操作系统中多种应用程序所支持,是 Windows 的标准图像文件格式。随着 Windows 操作系统的流行以及丰富的应用程序诞生,BMP 图像格式理所当然地被广泛应用。这种格式的特点是采用位映射存储格式,除了图像深度可选以外,不进行任何压缩,几乎没有损耗,图像包含的信息较丰富,但因此导致了它与生俱来的缺点——磁盘占用空间过大。所以,目前 BMP 在单机上比较流行。

2) JPEG 格式

JPEG 是最常见的一种图像格式,应用十分广泛,特别是在网络和光盘读物上,都能找到它的身影。它由联合照片专家组(Joint Photographic Experts Group, JPEG)开发并命名为 ISO 10918-1, JPEG 仅仅是一种俗称而已。JPEG 文件的扩展名为 .jpg 或 .jpeg, 其压缩技术十分先进,它用有损压缩方式去除冗余的图像和彩色数据,获取极高压缩率的同时能展现十分丰富生动的图像,换句话说,就是可以用最少的磁盘空间得到较好的图像质量。

同时, JPEG 还是一种很灵活的格式,具有调节图像质量的功能,可以调节不同的压缩比例对文件进行压缩,如最高可以把 2.74MB 的 BMP 位图文件压缩至 40.6KB。当然也完全可以在图像质量和文件尺寸之间找到平衡点。

目前各类浏览器均支持 JPEG 这种图像格式,因为 JPEG 格式的文件尺寸较小,下载

速度快,使得 Web 页有可能以较短的下载时间提供大量美观的图像,JPEG 因此也就顺理成章地成为网络上最受欢迎的图像格式。

3) PNG 格式

PNG(Portable Network Graphics)是目前一种新兴的网络图像格式。1994 年年底,由于 Unysis 公司发表 GIF 拥有专利的压缩方法,要求开发 GIF 软件的作者须缴纳一定费用,由此促使免费的 PNG 图像格式的诞生。其主要特点如下。

(1) PNG 一开始便结合了 GIF 及 JPG 两家之长,存储形式丰富,兼有 GIF 和 JPG 的色彩模式,打算一举代替这两种格式;

(2) 能把图像文件压缩到极限以利于网络传输,但又能保留所有与图像品质有关的信息,因为 PNG 是采用无损压缩方式来减少文件的大小,这点与牺牲图像品质以换取高压缩率的 JPG 有所不同。

(3) 显示速度很快,只需下载 1/64 的图像信息就可以显示出低分辨率的预览图像。

(4) PNG 无法进行动画应用效果,如果在这方面能有所涉及,简直就可以完全替代 GIF 和 JPEG 了。

Macromedia 公司的 Fireworks 软件的默认格式就是 PNG。现在,越来越多的软件开始支持这一格式,而且在网络上也越来越流行。

5.1.5 图像标注各类型介绍

图像标注方式有矩形框标注、3D 标注、多边形拉框、打点、语义分割等。当然,具体标注的方法取决于实际项目所使用到的图像标注类型。计算机视觉的飞速发展,离不开大量图像标注数据的支撑,随着各类图像检测、识别算法的商业化落地,市场对图像标注的精准度愈发严格,为了针对不同的应用场景,也衍生出了不同的图像标注方法。

1. 2D 边界拉框

2D 边界框又叫矩形框,标注用途广泛且简单明了,是计算机视觉中最常用最广泛的图像标注类型之一。该标注负责在图像中的某些对象周围绘制框,边框应尽可能地靠近对象的每个边缘。边界框圈出目标,并协助计算机视觉网络找出感兴趣的目标。边界框几乎可以应用于任何目标,而且能大幅提升目标检测系统的准确度,其实质的应用场景如下。

1) 自动驾驶

2D 边界框标注多数会出现在目标检测模型运用里,显然自动驾驶系统就是其典型应用场景,定位捕获交通图像内的车辆、行人等实体就必不可少。

2) 智能建筑

另外还能用在建筑工地上为目标归类创建分类目标的模型,分析工地安全,让机器识别出不同环境中的目标。

3) 智能超市

如图 5-10 所示 2D 边界框可以用于标注产品的图像,识别杂货店的食物及其他物品,自动监测结账流程。



图 5-10 2D 边界框在自动驾驶中的标注场景(见彩插)

2. 3D 标注

3D 标注是一种非常强大的图像标注,与边界框非常相似,是在立体图像中识别对象并在其周围绘制边框。与仅描绘长和宽的 2D 边界框不同,3D 长方体则标注了对象的长、宽和近似深度,让目标以三维效果呈现,使计算机视觉系统在三维空间中学会区分体积和位置等特征。

如图 5-11 所示使用 3D 长方体标注,可以绘制一个框将感兴趣的对象封装起来,并将锚点放置在对象的每个边缘。如果对象的一个边缘不可见或被图像中的另一个对象所遮挡,那就需要根据该对象的大小、高度以及图像的角度,来估算其边缘的位置。

该标注类型常用来开发能够运动的自动系统,从而预测目标在其周围环境中的状况。例如,自动驾驶车辆和移动机器人的计算机视觉系统。



图 5-11 3D 标注呈现效果(见彩插)

3. 多边形

很多时候,图像中的目标由于光照、角度或者遮挡等原因,其形状、大小或方向无法很好地与 2D 边界框或 3D 长方体适配上。同时,开发人员希望对图像中的目标进行更加精确的标注。在这些情况下,便需要选择另一种图像标注技术——多边形进行标注。其也是边界框背后理论的扩展,复杂的多边形比单纯的框更能精确地检测出目标的位置和边界。

在使用多边形时,通常会在需要标注的对象外边缘,标记许多个点来绘制成线。这个

过程与“连点成线,勾勒轮廓”相似。和边界框相比,多边形标注能够捕捉不规则形状的目标,切掉目标边缘无用像素,避免给分类标准造成困扰,提升模型的精确度。其主要应用场景如下。

1) 自动驾驶

因为多边形能勾勒物品轮廓,消除边界框中的无用像素,多边形分割在自动驾驶中非常有用,能够突出显示标志和路标等不规则物体。

2) X 光片

多边形可用于在医用 X 射线中标记器官,以便将它们输入深度学习模型,以训练 X 射线中的畸形或缺陷。

3) 卫星图片

它还能用来精确标注众多不规则目标,如卫星和无人机所检测的对象。如图 5-12 所示需精确探测水生物,多边形分割也比边界框更好。

4. 线和样条线

线标注包含直线和曲线,主要用于描绘图像的各个部分。当需要标注且划分界限的部分太小或者太薄,边界框等方法无法描绘时,便可使用线标注。其主要应用场景如下。

1) 道路场景

尽管线和样条线可以被用于多种用途,但最好用在如图 5-13 所示外观重要特征为线形的目标中,它们在此主要被用于训练驾驶系统,以识别车道及其边界。



图 5-12 多边形标注应用场景



图 5-13 道路场景中的线性标注(见彩插)

2) 仓库机器人

线和样条线也可以被用于训练仓库里的机器人,让它们能够整齐地将箱子挨个摆放,或是将物品准确地放置到传送带上。

3) 制造行业

线和样条可用于标注工厂的图像线跟随机器人工作。这可以帮助自动化生产过程,人力劳动可以最小化。

5. 语义分割

和上述主要着手于描绘对象的外部边缘(或边界)标注不同,语义分割则要更加精确和具体一些。其流程是根据物体的属性,把复杂不规则的图像分割成不同区域,整个图像中的每个像素与标签属性相关联的过程。

例如,图片中一部分可能是“天空”,而另一部分可能是“草地”,语义分割的关键是,各区域由语义信息所定义。在一些需要用到语义分割的项目中,一般会给标注工具提供一系列预定义的标签,以便在定义过程中能够从中选择需要标记的内容。其应用场景如下。

1) 医学成像

针对提供过来的患者照片,标注工具将从解剖学角度对不同的身体部位,打上正确的部位名称标签。因此,语义分割可以被用于处理诸如“在 CT 扫描图像中标记脑部病变”之类难度较大的特殊任务。

2) 自动驾驶汽车

图像标注可以利用语义分割对图像中的每个像素进行标记,使车辆能够感知到道路上的障碍物,分辨出道路、草地和人行道的各个区域。

3) 农田分析

使用如图 5-14 所示的语义分割标注,检测杂草和特定的作物类型,检测森林和雨林的毁坏和生态系统破坏,促进生态保护。



图 5-14 语义分割标注结果(见彩插)

6. 特征点位标注

因为特征点标注在图片上创建点,所以有时也被称为点标注。仅仅几个小点就能为

图片中细小纷繁的目标归类。但特征点标注常常使用许多点来描绘目标的轮廓或框架。特征点大小多样,大些的点有时会用来在区域中区分出重要/标志区域。特征点/点标注用点表示目标,所以最主要的用法是检测并量化小型目标。例如,城市鸟瞰图需要用特征点检测来找到车辆、房屋、树木、水池等感兴趣的目標。也就是说,特征点标注也有其他用法。将重点特征点结合起来便能创建目标轮廓,就像是连点拼图的游戏。这些点形成的轮廓能用来识别面部特征,或者分析人的动作或姿势。其主要应用场景如下。

1) 面部识别

追踪多个特征点能轻松识别出面部表情和其他面部特征。

2) 生物学领域的几何形态测量

如图 5-15 所示人脸骨骼点、场景目标物体以及统计模型。

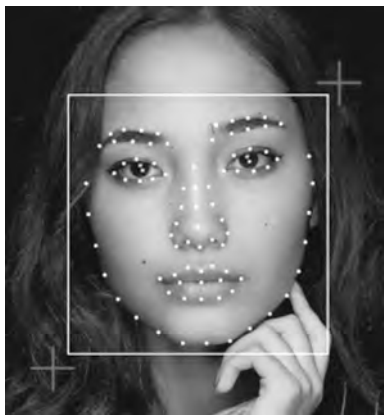


图 5-15 脸部特征点位标注

7. 其他图像标注类型

1) 云标注

点云是三维数据的一种重要表达方式,通过激光雷达等传感器,能够采集到各类障碍物及其位置坐标,而标注员则需要将这些密集的点云分类,并标注上不同属性。

2) 目标跟踪

目标跟踪是指在动态的图像中,进行抽帧标注,在每一帧图片中将目标物体标注出来,描述它们的运动轨迹,这类标注常应用于训练自动驾驶模型以及视频识别模型。

3) 属性判别

属性判别是指通过人工或机器配合的方式,识别出图像中的目标物体,并将其标注上对应属性。

总之,只要选择对应且正确的标注方法,计算机视觉便能实现所有的目标。在了解图像标注的众多种类和用例后,最好进行实验,付诸实践,从而掌握实际应用中的最佳方法。

5.2 图像数据采集和整理

5.2.1 数据采集的渠道和方法

数据采集,又称数据获取,是利用一种装置从系统外部采集数据并输入到系统内部一个接口的过程。在计算机广泛应用的今天,数据采集的重要性是十分显著的,它是计算机与外部物理世界连接的桥梁纽带。面对各种类型信号采集,其难易程度差别会很大,实际采集时,杂质和噪声也可能带来一些麻烦。

数据采集时,有一些基本原理要注意,往往原始材料与第一手资源针对性强且更为准确。但相应的采集比较耗时耗力,而对于互联网信息以及文献资料、研究报告等丰富的现成资料,采集相对比较快捷,其中还有更多的实际问题要解决。数据采集技术广泛应用在

各个领域,如摄像头、麦克风,都是常见的数据采集工具。

对于来自不同领域的源数据,需要遵循特定的数据采集流程。具体而言,数据采集在明确数据来源之后,可以根据特定行业与应用定位,确定采集的数据范围与数量,并通过核实的数据采集方法,展开后续的数据采集工作。

在互联网行业快速发展的今天,数据采集已经被广泛应用于互联网及分布式领域,数据采集领域已经发生了重要的变化。首先,分布式控制应用场合中的智能数据采集系统在国内外已经取得了长足的发展;其次,总线兼容型数据采集插件的数量不断增大,与个人计算机兼容的数据采集系统的数量也在增加。

国内外各种数据采集机先后问世,将数据采集带入了一个全新的时代。目前一些大数据公司和很多数据研究机构,获取图像数据的方法主要包括如下四种。

1. 系统日志采集

日志文件是用于记录数据源执行的各种操作行为,包括股票记账、流量管理、Web 服务器记录等用户访问行文的记录,很多互联网企业都有自己的海量数据采集工具,多用于系统日志采集。如 Hadoop 的 Chukwa, Cloudera 的 Flume, Facebook 的 Scribe 等,这些工具均采用分布式架构,能满足大数据的日志数据采集和传输需求。

2. 互联网数据采集

通过网络爬虫或者网站公开 API 等方式从网站上获取数据信息,该方法能够按照一定的规则,自动地把数据从网页中抽取出来进行处理,使其统一存储到本地数据文件。它可以支持图片、音频、视频等文件或附件的采集,同时附件与正文可以自动关联。处理网站中包含的内容之外,还可以使用 DPI 或 DFI 等带宽管理技术实现对网络流量的采集。

3. App 移动端或终端数据采集

App 是获取用户移动端数据的一种有效方法,App 中的 SDK 插件可以将用户使用 App 的信息汇总给指定服务器,即便用户在没有访问时,也能获知用户端的相关信息,包括安装应用的数量和类型等。单个 App 用户规模有限,数据量有限;但数十万 App 用户,获取的用户终端数据和部分行为数据也会达到数亿的量级。

日常生活中我们常见道路或者商场内部的监控摄像头,还有各种语音控制设备都是良好的数据采集方式,这些终端通过日常的工作形态,将其运作过程中的数据保存下来,后期再经过合理的清洗和分类,就能获得最真实原始的计算机视觉训练数据。

4. 向数据服务机构进行购买

有很多专门做行业研究的组织、公司和数据服务机构,能够在某一特定领域获得大量的数据,通常具备规范的数据共享和交易渠道,这些组织、公司和机构会通过有偿或无偿的方式将数据共享给数据需求者,人们可以在平台上快速、明确地获取自己所需要的数据。而对于企业生产经营数据或学科研究数据等保密性要求较高的数据,也可以通过与企业或研究机构合作,使用特定系统接口等相关方式采集数据。

以上均为常见的数据采集方法,当然针对特定项目场景算法模型的需要,在通过上述批量获取数据之外,对于一些人像、车辆、街景等数据需求者可以自行采集数据,到相似现场进行拍摄收集图片,对语音可进行人工朗读、转录成音频文件来收集,或者直接从书籍、文章中提取特定文本内容等。也可以委托数据标注平台采集数据,现在很多数据标注平台都提供数据采集的业务,它们将数据采集任务以众包的形式发布在网站上,付费给网友来协助采集。

5.2.2 图像数据采集案例

前文介绍了数据采集的方法和流程之后,下面来具体描述一些常见数据采集项目的过程与要求。

1. 人脸数据采集

关于人脸数据的获取,可以从第三方数据机构进行购买,或者可自行准备一些采集设备进行收集。在采集之前,应该需要提前了解应用场景,明确采集数据的规格,对包括年龄、人种、性别、表情、拍摄环境、姿态分布等予以准确限定,明确图片尺寸、文件大小与格式、图片数量等要求,同时在获得被采集人许可之后,还需对被采集人拍摄不同光线、不同角度、不同表情的数据,并在收集后对数据做脱敏处理。

如下为一个简单的人脸数据采集项目的示例卡片。

年龄分布——18~30岁
性别分布——男女比例:6:4
人种分布——黑种人:50;白种人:40;黄种人:10
表情类型——正常,挑眉,向左看,向右看,向上看,向下看,闭左眼,闭右眼,微张嘴,张大嘴,嘟嘴,微笑,大笑,惊讶,悲伤,厌恶
采集方式——自行使用采集设备
拍摄环境——光线亮的地方,光线暗的地方,光线正常的地方
图片尺寸——1200×1600
文件格式——JPG
数据量级——20 000
适用领域——人脸识别,人脸检测

2. 车辆数据采集

在对车辆数据的采集中,最常用的方式是通过交通监控视频进行图片抓取,数据中最好包含车牌、车型、车辆颜色、品牌、年份、位置、拍摄时间等车辆信息。定好统一的图片尺寸、文件格式、图片数量规定,必要时做脱敏处理(即数据漂白),实时保护隐私和敏感数据。

如下为一个车辆数据采集项目示例卡片。

车型类型——小轿车、SUV、面包车、客车、货车、其他
车辆颜色——白、灰、红、黄、绿、其他
拍摄时间——光线亮的时候,光线暗的时候,光线正常的时候
车牌颜色——蓝、白、黄、黑、其他
图片尺寸—— 1024×768
文件格式——JPG
图片数量——75 000 张
适用领域——自动驾驶、车牌识别

3. 街景数据采集

与车辆数据采集相同,街景数据采集也能来自监控视频,可将视频进行截图与收集,还可借助车载摄像头、水下相机等进行街景拍摄。例如,谷歌在进行街景拍摄时,通过采集、定位到街景传感器吊舱、街景眼球、街景塔、街景三轮车、街景雪地车、街景水下相机等多种渠道进行全方位图像采集。采集的街景数据主要包括城市道路、十字路口、隧道、高架桥、信号灯、指示标志、行人与车辆等场景。同时,对于采集的数据需要做统一的图片尺寸、文件格式、图片数量规定与脱敏处理。

如下是一个街景数据采集项目示例卡片。

数据环境——城市道路
路况类型——十字路口、高架桥、隧道
数据量级——10 000 张
采集设备——车载摄像头
图片尺寸—— 1920×1200
文件格式——PNG
图片数量——15 000 张
应用领域——自动驾驶

4. 人体姿态采集

数据对象:办公室及相应工作场景周边的人体姿态。

如下是一个人体姿态采集项目示例卡片。

图片要求——不能直接在网络上获取,必须是真实拍摄的照片,不可以做修改
年龄分布——18~60 岁
性别分布——男女比例:1:1
拍摄光线要求:正常光线、较暗光线和较亮光线
照片清晰度:清晰度大于 1080P,无闪光灯及其他灯光问题

文件格式: JPG

照片数量: 30 000 张

适用领域: 人体识别

1) 场景 1: 办公室内

姿态要求:

(1) 手平放在桌子上, 露出胳膊之间的间隙, 左前方放置绿植遮挡部分手臂, 遮挡部分不超过左手臂的三分之一。

(2) 左手平放在桌子上, 右手五指张开, 直立在桌子上, 不要遮挡身体躯干, 且要求露出胳膊之间的空隙。

(3) 左手平放在桌子上, 右手五指张开, 直立在桌子上, 不要遮挡身体躯干, 且要求露出胳膊之间的空隙, 左前方放置绿植遮挡部分手臂, 遮挡部分不超过左手臂的二分之一。

2) 场景 2: 办公室过道

指房间与其他房间或墙面形成的狭窄区域, 非写字楼的楼道及其他场景的楼道。姿态要求:

(1) 单人全身出境, 不要遮挡面部。

(2) 站立于过道中间, 不要依靠两侧墙体或其他物体。

3) 场景 3: 其他工作场景

如业务大厅工作人员、教室等。

姿态要求: 靠墙环抱手臂, 双臂环抱胸前, 臂膀紧靠在墙上, 墙体完整且不为纯白色。室内站立状态下正面背靠绿植或侧面靠绿植。所有照片不可以佩戴口罩或用其他物品遮挡面部。桌面除绿植外, 不要有其他物体, 单人上半身出境。

5. 停车位和交通标志采集

采集对象: 车位范围及交通标志

采集场景: 停车场和标识道路

采集地域: 武汉、上海、广州、深圳

采集内容: 采集各个停车场停车位图片, 以及在停车过程中采集交通标志图片

采集方法: 驾驶员负责开车, 采集员负责在后排拍摄

图片格式: JPG

图片数量: 300 000 张

停车场个数: 500 个

适用领域: 自动停车

5.2.3 数据预处理和清洗

数据预处理是一种数据挖掘技术, 其目的是把原始数据转换为可以理解的格式或者符合数据挖掘的格式。现实世界中获取的数据是大部分不完整、不一致的脏数据, 无法直

接进行数据挖掘或者数据挖掘结果差强人意。为了提高数据挖掘的质量,产生了数据预处理技术。数据预处理有多种方法,包括数据清洗、数据集成、数据归约、数据变换。需要注意的是,这些方法不一定会同时使用,或者有时几种方法会结合在一起使用。在数据挖掘之前使用数据预处理技术先对数据进行一定的处理,将极大地提高数据挖掘的质量,降低实际数据挖掘所需的时间。

当数据质量核查完毕之后,针对有问题的数据要进行数据清洗和转换,并且还包含对合格数据的转换。如图 5-16 所示数据清洗从名字上也看得出就是把“脏”的“洗掉”,指发现并纠正数据文件中可识别的错误的最后一道程序,包括检查数据一致性,处理无效值和缺失值等。因为获取数据后,并不是每一条数据都能够直接使用,来自各处数据源的数据内容并不“完美”。有些数据存在缺失不完整,有的数据存在噪声不一致,有的数据异常存在错误和重复,按照一定的规则把异常数据“洗掉”,这就是数据清洗。

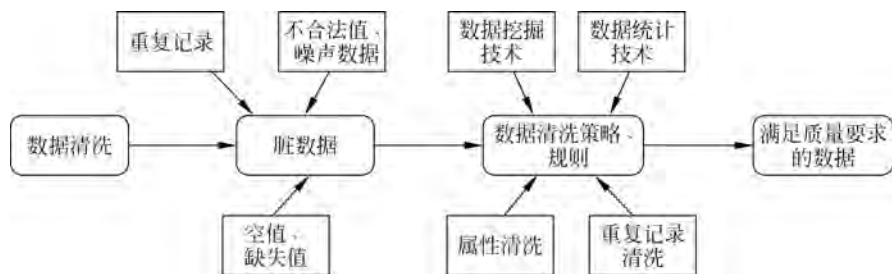


图 5-16 数据清洗流程

1. 数据清洗流程

在具体的数据清洗流程中,能够按指定错误类型、判断错误实例、修正发现错误、正确数据回收的具体流程来进行开展。

1) 指定错误类型

在这个环节中,能够通过手动检查或者数据样本的分析方式,检测分析数据中包含的异常,然后在此基础上定义清洗转换规则和工作流程。根据数据源的数量以及缺失、不一致或者冗余情况,决定数据转换和清洗步骤。

2) 判断错误实例

在核查过程中,如果只利用人工方式,一般耗时耗力,准确率也难以保障。因此在这个过程中,能够首先通过统计、聚类或者关联规则的方法,自动检查数据的属性错误。提高核查的效率,且扩大范围,使之发现更多的错误。

3) 修正发现错误

对于纠正错误,便按照最初定义的数据清洗要求和工作流程有序进行。在纠正过程中,为了方便处理,应该对数据源进行分类处理,并在同一类型数据中将属性值统一格式,做标准化处理。

此外,在处理之前还应该对源数据进行备份,以免造成数据修改错误或者意外丢失,需要进行撤销操作等情况。

4) 正确数据回收

通过以上三大环节,基本可以得到正确干净的数据,这时需要把原来的“脏”数据给替换掉,实现干净数据回流,以提高数据质量,同时也避免了将数据重复清洗的工作。

2. 数据清洗的主要作用

1) 修改错误

出错的数据是数据源环境中一直都存在的一类问题。数据错误的形式包括以下。

(1) 数值错误。

数据直接是错误的,例如,超过固定域集、超过极值、拼写错误、属性错误、源错误等。

(2) 类型错误。

数据的存储类型与实际情况不符,如日期类型的以数值型存储,时间戳存为字符串等。

(3) 编码错误。

存储的数据编码错误,例如,将 UTF-8 写成 UTF-80。

(4) 格式错误。

存储的数据格式问题,如半角全角字符、中英文字符等。

(5) 异常错误。

如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期越界、数据前后有不可见字符等。

(6) 依赖冲突。

某些数据字段间存储依赖关系,例如,城市与车牌字符应该满足对应关系,但可能存在二者不匹配的问题。

(7) 多值错误。

大多数情况下,每个字段存储的是单个值,但也存在一个字段存储多个值的情况,其中有些可能是不符合实际业务规则的。

这类错误产生的原因是业务系统不够健全,尤其是在数据产生之初的校验和入库规则不规范,导致在接收输入后没有进行判断或无法检测而直接写入后台数据库造成的。

2) 剔除重复数据

因为各类情况,数据集中可能存在重复记录或重复字段(列),对于这些重复项目(行和列)需要做去重处理。对于重复数据的处理,基本方式是“排序和合并”,先把数据库中的记录按一定规则排序,然后通过比较邻近记录是否相似来检测记录是否重复。这里面其实包含两个操作,一是排序,二是计算相似度。

(1) 常见的排序算法。

插入排序、冒泡排序、选择排序、快速排序、堆排序、归并排序、基数排序、希尔排序。

(2) 常见的判断相似度的算法。

基本的字段匹配算法、标准化欧氏距离、汉明距离、夹角余弦、杰卡德距离、马氏距离、曼哈顿距离、闵可夫斯基距离、欧氏距离、切比雪夫距离、相关系数、信息熵。

对于重复的数据项,尽量需要经过业务确认并进行整理提取出规则。在清洗转换阶

段,对于重复数据项尽量不要轻易做出删除决策,尤其不能将重要的或有业务意义的数据过滤掉,校验和重复确认的工作必不可少。

3) 统一规格

大多数项目的数据都来自不同系统,且分散在各个业务线,不同业务线对于数据的要求、理解和规格不同,会导致对于同一数据对象描述规格完全不同,因此在清洗过程中需要统一数据规格并将具有一致性的内容抽象出来。数据字段的规则大致可以从以下几个方面进行统一。

(1) 名称。

对于同一个数据对象的名称首先应该是一致的。例如,对于访问深度这个字段,可能的名称包括访问深度、人均页面浏览量、每访问 PV 数。

(2) 类型。

同一个数据对象的数据类型必须统一,且表示方法一致。例如,普通日期的类型和时间戳的类型需要区分。

(3) 单位。

对于数值型字段,单位需要统一。例如,万、十万、百万等单位度量。

(4) 格式。

在同一类型下,不同的表示格式也会产生差异。例如,日期中的长日期、短日期、英文、中文、年月日制式和缩写等格式均不一样。

(5) 长度。

同一字段长度必须一致。

(6) 小数位数。

小数位数对于数值型字段尤为重要,尤其当数据量累积较大时会因为位数的不同而产生巨大偏差。

(7) 记数方法。

对于数值型等的千分位、科学记数法等的记数方法的统一。

(8) 缩写规则。

对于常用字段的缩写,例如,单位、姓名、日期、月份等的统一。例如,将周一表示为 Monday 还是 Mon 还是 M。

(9) 值域。

对于离散型和连续型的变量都应该根据业务规则进行统一的值域约束。

(10) 约束。

是否允许控制、唯一性、外键约束、主键等的统一。

统一数据规格的过程中,重要的一点是确认不同业务线带来数据的规格一致性,这需要业务部门的参与、讨论和确认,以明确不同体系数据的统一标准。

4) 修正逻辑

如果一个数据集的来源存在多个数据集,很可能存在数据异常或冲突的问题。例如,不同的数据源对于订单数量的数据统计冲突问题,结果出现矛盾的记录。通常,这是由于不同系统对于同一个数据对象的统计逻辑不同而造成的,逻辑的不一致会直接导致结果

的差异性；除了统计逻辑和口径的差异，也有因为源数据系统基于性能的考虑，放弃了外键约束，从而导致数据不一致的结果；另外，也存在极小的数据丢失的可能性，通常由于并发量和负载过高、服务器延迟甚至宕机等原因导致数据采集的差异。

对于这类数据矛盾，首先需要明确各个源系统的逻辑、条件、口径，然后定义一套符合各个系统采集逻辑的规则，并对异常源系统的采集逻辑进行修正。

某些情况下，也可能存在业务规则的错误导致的数据采集的错误，此时需要从源头纠正错误的采集逻辑，然后再进行数据清洗和转换。

5) 转换构造

数据变换是数据清理过程的重要步骤，是对数据的一个标准处理，几乎所有的数据处理过程都会涉及该步骤。数据转换常见的内容包括：数据类型转换、数据语义转换、数据值域转换、数据粒度转换、表/数据拆分、行列转换、数据离散化、数据离散化、提炼新字段、属性构造等。

(1) 数据类型转换。

当数据来自不同的数据源时，不同类型的数据源数据类型不兼容可能导致系统报错。这时需要将不同数据源的数据类型进行统一转换为一种兼容的数据类型。

(2) 数据语义转换。

传统数据仓库中基于第三范式可能存在维度表、事实表等，此时在事实表中会有很多字段需要结合维度表才能进行语义上的解析。例如，假如字段 M 的业务含义是浏览器类型，其取值分别是 1/2/3/4/5，这 5 个数字如果不加转换则很难理解为业务语言，更无法在后期被解读和应用。

(3) 数据粒度转换。

业务系统一般存储的是明细数据，有些系统甚至存储的是基于时间戳的数据，而数据仓库中的数据是用来分析的，不需要非常明细的数据。一般情况下，会将业务系统数据按照数据仓库中不同的粒度需求进行聚合。

(4) 表/数据拆分。

某些字段可能存储多种数据信息，例如，时间戳中包含年、月、日、小时、分、秒等信息，有些规则中需要将其中部分或者全部时间属性进行拆分，以此来满足多粒度下的数据聚合需求。同样地，一个表内的多个字段，也可能存在表字段拆分的情况。

(5) 行列转换。

某些情况下，表内的行列数据会需要进行转换（又称为转置），例如，协同过滤的计算之前，user 和 term 之间的关系即互为行列并且可相互转换，可用来满足基于项目和基于用户的相似度推荐计算。

(6) 数据离散化。

将连续取值的属性离散化成若干区间，来帮助消减一个连续属性的取值个数。例如，对于收入这个字段，为了便于做统计，根据业务经验可能分为几个不同的区间：0~3000、3001~5000、5001~10 000、10 001~30 000、大于 30 000，或者在此基础上分别用 1、2、3、4、5 来表示。

(7) 数据标准化。

不同字段间由于字段本身的业务含义不同,有些时间需要消除变量之间不同数量级造成的数值之间的悬殊差异。例如,将销售额进行离散化处理,以消除不同销售额之间由于量级关系导致的无法进行多列的复合计算。数据标准化过程还可以用来解决个别数值较高的属性对聚类结果的影响。

(8) 提炼新字段。

很多情况下,需要基于业务规则提取新的字段,这些字段也称为复合字段。这些字段通常都是基于单一字段产生,但需要进行复合运算甚至复杂算法模型才能得到新的指标。

(9) 属性构造。

在有些建模过程中,也会需要根据已有的属性集构造新的属性。例如,几乎所有的机器学习都会将样本分为训练集、测试集、验证集三类,那么数据集的分类(或者叫分区)就属于需要新构建的属性,用户作机器学习不同阶段的样本使用。

6) 数据压缩

数据压缩通常是表示在保持原有数据集的完整性和准确性,以及不丢失有用信息的前提下,按照一定的算法和方式对数据进行重新组织的一种技术方法。

对大规模的数据进行复杂的数据分析与数据计算通常需要耗费大量时间,所以在之前需要进行数据的约减和压缩,减小数据规模,而且还可能面临交互式的数据挖掘,根据数据挖掘前后对比对数据进行信息反馈。这样在精简数据集上进行数据挖掘显然效率更高,并且挖掘出来的结果与使用原有数据集所获得结果基本相同。

数据压缩的意义不止体现在数据计算过程中,还有利于减少存储空间,提高其传输、存储和处理效率,减少数据的冗余和存储的空间,其包括无损压缩和有损压缩两种类型。数据压缩常用于磁盘文件、视频、音频、图像等,这对于底层大数据平台具有非常重要的意义。数据压缩有多种方式可供选择:

(1) 数据聚合。

将数据聚合后使用,例如,如果汇总全部数据,那么基于更粗粒度的数据更加便利。

(2) 维度约减。

通过相关分析手动消除多余属性,使得参与计算的维度(字段)减少;也可以使用主成分分析、因子分析等进行维度聚合,得到的同样是更少的参与计算的数据维度。

(3) 数据块消减。

利用聚类或参数模型替代原有数据,这种方式常见于多个模型综合进行机器学习和数据挖掘。

3. 常见数据清洗的方法

根据上面所述,如果需要获取到高质量的数据,则在拿到数据之后,需要将不规整的数据转换为优质数据,因而完成准确、简洁的数据清洗工作便成为数据预处理中的重要环节。面对不同的业务需求,数据清洗包含以下几种方法。

1) 数据缺失处理

当数据到达我们手上时,难免会丢失部分信息,无法做到全部完成。例如,数据库中

的图像在传输过程中经过了压缩处理、表格中的列值不会全部强制性不为空类型、设备异常对数据改变没有日志记载等。面对这些问题,解决的方法大致为以下三种。

(1) 消除存在缺失值的日志记录。

就是将含有缺失属性值的对象直接删除,从而得到一个完备的数据信息。理论上讲,删除具有缺失值记录的方法主要有直接删除法和权重法。其中,直接删除法是对缺失值进行处理的最原始、最简单的方法。在缺失属性的数据占整个数据集比例较小时,这种方法比较适用,一般是在进行分类任务中缺失类别标号属性时常采用。如果数据集中大量的数据都存在缺失值问题,则此种方法失效。或者整体的数据量级过小,这样删除宝贵的数据资源,便会很大地影响挖掘结果的准确性。

(2) 补全缺失值。

采用一定的值对缺失的属性进行填充补齐,进而使数据信息完备化。填补缺失值的核心观念是通过最可能出现的值来补全缺失值,这相对于把不完全样本全部删除所造成的信息丢失要少得多。在数据挖掘中,面对的一般都是大型的数据库,数据会附带几十个甚至几百个属性,因为其中一个属性的错误或者丢失,而抛弃其他大量的属性值,是对信息的极大浪费,因此便有了以可能值对缺失值进行填补的思想与方法,常用的有以下几种方法。

① 人工填写:需要用户对数据相关信息属性非常了解,面对大数据集时,此种方法效率太低。

② 指定值填充:将所有空值使用一个指定值进行填充,使用此方法有可能导致严重的数据偏离。

③ 均值填充:如果缺失的是定量数据,则用该字段存在值的平均值来插补缺失的值,即数值型数据属性;如果缺失的是定性数据,则依照统计学中的众数原理,用该属性中出现频率最高的值来插补缺失的值。

(3) 最大可能性的预估填补。

在缺少情况为随机缺失的条件下,假如模型对于完整的样本是正确的,那么通过观测数据的边际分布可以对未知参数进行极大似然估计。这种方法也称为忽略缺失值的极大似然估计。对于极大似然的参数估计,实际中常采用的计算方法是期望值最大化。使用前提是增大样本,并且有效样本的数量足以保证极大似然估计值是渐近无偏的并且服从正态分布。但是这种方法可能会陷入局部极值,收敛速度也不是很快,并且计算很复杂。

2) 重复数据处理

如果两个值中的所有字段和属性都相同,对于重复的值一定要剔除。但数据集的量级不够大,则可根据业务场景,只选取其中几个字段来进行去重操作,避免数据集过小。

3) 噪声数据处理

噪声数据中会存在错误和异常,即偏离期望值的数据,会影响数据分析的结果。造成这种情况有多方面的原因,如数据收集工具的问题,数据输入以及传输错误,技术问题等。如果想优化算法模型的效果,通过使用标注数据来迭代获取最优的线性算法,然后训练数据集中含有大量的噪声数据,那么就会大大影响算法的收敛速率,并且还会对训练生成模型准确性造成负面影响,即越迭代效果越差。上述问题过程是很多项目算法模型后期都

会遇到和存在的问题,就是因为没有重视噪声数据的清洗所造成的。其基本的噪声数据处理有三个方法,分别是分箱法、回归法、聚类法。

(1) 分箱法。

这是一个经常使用到的方法。所谓的分箱法,就是将需要处理的数据根据一定的规则放进箱子里,然后测试每一个箱子里的数据,并根据数据中的各个箱子的实际情况采取方法处理数据。

(2) 回归法。

回归法就是利用函数的数据进行绘制图像,然后对图像进行光滑处理。回归法有两种,一种是单线性回归,一种是多线性回归。单线性回归就是找出两个属性的最佳直线,能够从一个属性预测另一个属性。多线性回归就是找到很多个属性,从而将数据拟合到一个多维面,这样就能够消除噪声。

(3) 聚类法。

聚类法的工作流程是比较简单的,但是操作起来却是复杂的。所谓聚类法,就是将抽象的对象进行集合分组,成为不同的集合,找到在集合以外的孤点,这些孤点就是噪声。这样就能够直接发现噪点,然后进行清除即可。

4. 常见图像数据清洗方法

1) 图像降噪

现实中的数据图像在数字化和传输过程中会经常出现成像设备与外部环境噪声干扰等影响因素,这种现象称为含噪图像或噪声图像。减少数据图像中噪声的过程称为图像降噪。

在演示的过程中可以看到当二值化之后的图片会显示很多小黑点,这些都是不需要的信息,会对后面进行图片的轮廓切割识别造成极大的影响,降噪是一个非常重要的阶段,降噪处理的好坏直接影响了图片识别的准确率。

最简单的叫作数据结构中学到的 DFS 或者 BFS(深度和广度搜索)。我们对 $w \times h$ 的位图先搜索所有连通的区域(值为 1 的,看起来是黑色的,连接起来的区域)。所有连通区域算一个平均的像素值,如果某些连通区域的像素值远远低于这个平均值,我们就认为是噪点。然后用 0 代替它。

2) 倾斜矫正

拍照或者选取的图片不可能完全是水平的,倾斜会影响后面切出来的图片,所以要对图片进行旋转。倾斜矫正最常用的方法是霍夫变换,其原理是将图片进行膨胀处理,将断续的文字连成一条直线,便于直线检测。计算出直线的角度后就可以利用旋转算法,将倾斜图片矫正到水平位置。



微课视频

5.2.4 图像处理工具

收集到的数据往往不能直接使用,通常需要经过图像处理程序的加工处理才能使用。能够进行图像处理的软件和工具有很多,常见的图像处理工具有 ProcessOn、Microsoft Office Visio、Photoshop 等。

1. ProcessOn

ProcessOn 是一个面向垂直专业领域的作图工具和社交网络,成立于 2011 年 6 月,并于 2012 年启动。它能够基于云服务的免费流程梳理、创作协作工具,与同事和客户实时进行协作协同设计,实时创建和编辑文件,并可以实现更改的及时合并与同步,这意味着跨部门的流程梳理、优化和确认可以即刻完成。支持绘制思维导图、流程图、UML、网络拓扑图、组织结构图、原型图、时间轴等。

ProcessOn 专注于为作图人员提供价值,利用互联网和社交技术颠覆了人们梳理流程的方法习惯,继而使商业用户获得比传统模式更高的效率和回报,改善人们对流程图的创作过程。

将全球的专家顾问、咨询机构、BPM 厂商、IT 解决方案厂商和广泛的企业用户紧密地连接在一起。

1) 应用场景

(1) 计划制定。

ProcessOn 的思维导图能够用来制定计划,包括工作计划、学习计划、旅游计划等,计划可以按照时间或项目划分,罗列出计划的内容、完成时间、执行人等信息后,可以标注出每个事项的完成度,将繁杂的日程整理清晰。

(2) 笔记。

ProcessOn 的思维导图可以记录学习或工作笔记。把大篇幅的学习内容进行拆分,缩短文字数量,总结知识点的从属关系进行概括,便于理解和记忆。

(3) 梳理流程。

如图 5-17 所示 ProcessOn 的流程图可以梳理系统流程、工艺流程、管理流程等,用图的形式来展现某一过程。

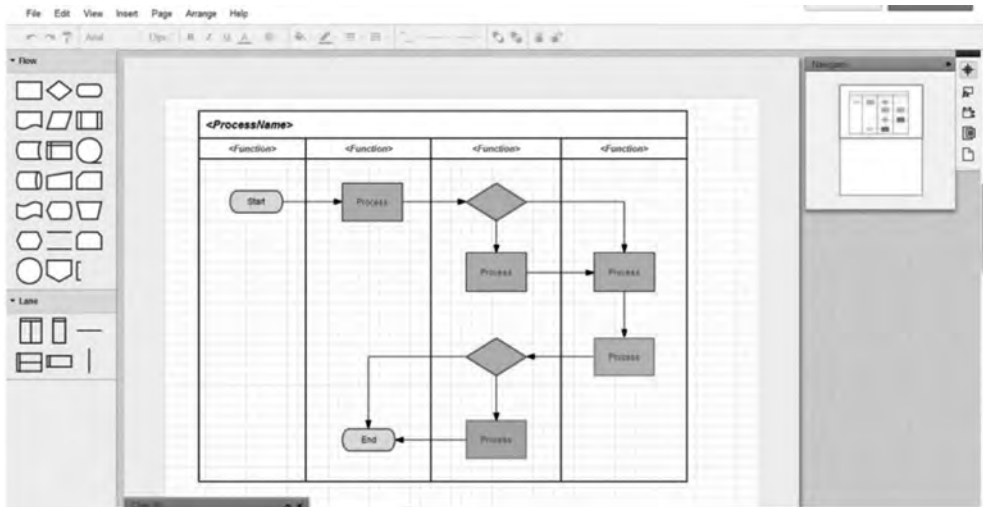


图 5-17 ProcessOn 界面

2) 常用快捷键

(1) 通用类快捷键如表 5-1 所示。

表 5-1 通用类快捷键

键 名	功 能
Alt	按住 Alt 键,通过鼠标可以对页面进行拖动
Ctrl	按住 Ctrl 键,单击一个图形,将其添加到选择图形中,或者从中移除
Ctrl+(+)/(一)	放大/缩小
Ctrl+A	全部选中
Esc	取消选中,并取消当先操作
T	插入文本
I	插入图片
L	插入连线

(2) 选中图形时快捷键如表 5-2 所示。

表 5-2 选中图形时快捷键

键 名	功 能
方向键	将选中图形向左、向上、向下、向右移动
Ctrl+方向键(← ↑ ↓ →)	每次微移一个像素
Ctrl+Z	撤销
Ctrl+Y	恢复
Ctrl+X	剪切
Ctrl+C	复制
Ctrl+V	粘贴
Ctrl+D	复用
Ctrl+Shift+B	格式刷
Delete, BackSpace	删除
Ctrl+]	将选中的图形置于顶层
Ctrl+[	将选中的图形置于底层
Ctrl+Shift+]	将选中的图形上移一层
Ctrl+Shift+[	将选中的图形下移一层
Ctrl+L	锁定选中的图形
Ctrl+Shift+L	将选中的图形解锁
Ctrl+G	组合选中的图形
Ctrl+Shift+G	将选中的图形取消组合

(3) 编辑文本时快捷键如表 5-3 所示。

表 5-3 编辑文本快捷键

键 名	功 能
空格	编辑文本
Ctrl+B	粗体

续表

键 名	功 能
Ctrl+I	斜体
Ctrl+U	下划线
Ctrl+Enter	保存文本编辑

2. Microsoft Office Visio

Office Visio 是 Office 软件系列中负责绘制流程图和示意图的软件,是一款便于 IT 和商务人员就复杂信息、系统和流程进行可视化处理、分析和交流的软件。使用具有专业外观的 Office Visio 图表,可以促进对系统和流程的了解,深入了解复杂信息并利用这些知识做出更好的业务决策。

OfficeVisio 很多时候可以用来处理一些专业的图像,使其标注数据能够更加精确。或者为一些特定的项目场景制定专门对口的数据环境,使其模型训练更加得专一性。例如,可将畸变过大的图像数据,进行长宽比重塑,使其消除畸变,满足模型识别规格;除去字符底色和背景,并将其镶嵌到其他场景环境中,拟造出更加多元化的数据,提升算法模型的兼容性。如表 5-4 所示为此图像编辑工具的常用快捷键列表

表 5-4 Office Visio 快捷键

键 名	功 能
Ctrl+Shift+U	取消对所选组合中形状的组合 (“形状”菜单,“组合”子菜单,“取消组合”)
Ctrl+Shift+F	将所选形状置于顶层 (“形状”菜单,“顺序”子菜单,“置于顶层”)
Ctrl+Shift+B	将所选形状置于底层 (“形状”菜单,“顺序”子菜单,“置于底层”)
Ctrl+L	将所选形状向左旋转 (“形状”菜单,“旋转或翻转”子菜单,“向左旋转”)
Ctrl+R	将所选形状向右旋转 (“形状”菜单,“旋转或翻转”子菜单,“向右旋转”)
Ctrl+H	水平翻转所选形状 (“形状”菜单,“旋转或翻转”子菜单,“水平翻转”)
Ctrl+J	垂直翻转所选形状 (“形状”菜单,“旋转或翻转”子菜单,“垂直翻转”)
F8	为所选形状打开“对齐形状”对话框 (“形状”菜单,“对齐形状”)
Ctrl+Shift+P	切换“格式刷”工具的状态 (“开”或“关”)
Ctrl+1	选择“指针工具”
Ctrl+3	选择“连接线工具”
Ctrl+Shift+1	选择“连接点工具”

续表

键 名	功 能
Ctrl+2	选择“文本工具”
Ctrl+Shift+4	选择“文本块工具”
Ctrl+Shift+3	选择“图章工具”

3. Photoshop

Photoshop 是目前公认的最好的功能非常强大的平面图像编辑工具及平面美术设计软件,功能相当多,性能稳定,使用方便。在实际生活和工作中,可以将数码照相机拍摄下来的照片进行编辑和修饰,也可以将现有的图形和照片用扫描仪扫入计算机进行加工处理,还可以把摄像机摄入的内容转移到计算机上,然后用它实现对影像的润色。

由于 Photoshop 功能强大,目前正在被越来越多的图像编排领域、广告和形象设计领域以及婚纱影楼等领域广泛使用,是一个非常受欢迎的应用软件。支持几乎所有的图像格式和色彩模式,能够同时进行多图层的处理;它的绘画功能和选择功能让编辑图像变得十分方便;它的图层样式功能和滤镜功能给图像带来无穷无尽的奇特效果。下面就简单介绍一下图像处理方面较为常用的相关功能。

1) Photoshop 的主界面

如图 5-18 所示,单击菜单栏中的“窗口”选项,可打开或关闭工具箱面板及其他浮动面板。

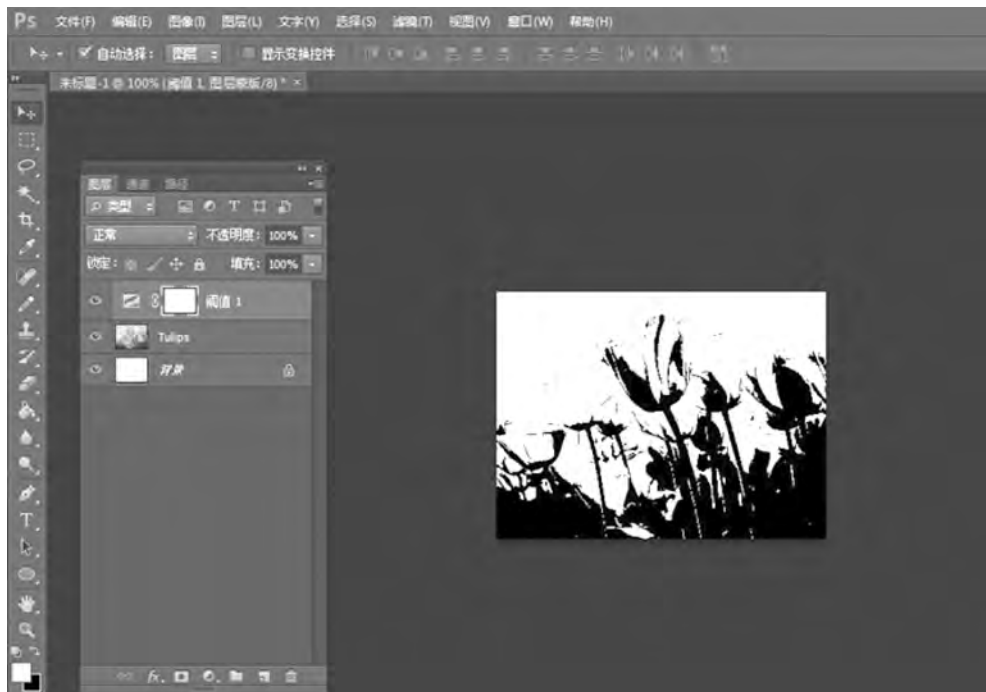


图 5-18 Photoshop 的主界面

2) 改变图像大小

第一步：依次单击“文件”→“打开”按钮，打开需要调整尺寸的原始图像。

第二步：依次单击“图像”→“图像大小”按钮，弹出“图像大小”对话框，在宽度和高度输入框里输入相应的数字，单击“确定”按钮。

3) 抠取图像

很多的图像处理工作，经常需要从现有的图像素材中截取或者抠取一部分使用，这就要用到抠取图像技术。在 Photoshop 中，抠取图像的方法有很多种，这里介绍利用选择工具抠取图像的方法。

(1) 矩形框选工具。

需要抠取的图像具有矩形规则。将鼠标放在矩形选框工具图标上，并按住鼠标，则会出现一个工具列表，这些工具能够建立矩形或椭圆形选区，甚至可以建立只有一个像素的水平或垂直选区。

(2) 套索工具。

可用于选取不规则形状的属性数据。套索工具分为两种，一种是多边形套索工具，另一种是磁性套索工具。多边形套索工具可以通过单击屏幕上的不同点来创建直线多边形区域。磁性套索工具能够自动对齐到图像的边缘，常用于创建精确的复杂选区。

(3) 魔棒工具。

魔棒工具能够依据颜色的相似性来判断选取。它可以选择一个图像中与其他区域颜色不同的区域，此工具的应用广泛，作用很大。

灵活运用好以上三种选择工具，便可从一幅图像中抠取想要的区域使用。

4) 图像的颜色改变

对图像的色泽、背景、曝光率等的调整，也是图像处理中经常会进行的操作。例如，数码相机拍摄或者扫描仪扫描出来的图像，由于设备及环境原因有时会出现颜色失真的现象，图像效果不理想，便可使用 Photoshop 对图像颜色进行调整。

(1) 了解颜色的属性。

颜色有三个属性：色相、饱和度和明度。色相是颜色的名称，用来描述颜色种类。饱和度是指一种色彩的浓烈或鲜艳程度，饱和度越高，颜色中的灰色组分就越低，颜色的色彩浓度就越高。明度是指颜色的明暗程度，它主要取决于该颜色吸收光线的程度。

(2) 颜色调整方法。

依次单击菜单“图像”→“调整”→“色相/饱和度”按钮，打开“色相/饱和度”对话框，拖动颜色三个属性的滑块，即可调整图像的颜色。

5) Photoshop 的图层

图层是 Photoshop 十分重要的功能。如图 5-19 所示，引入图层的概念，便于把一张复杂的图像分解为相对简单的多层结构，每个图层都可以进行独立调整，而图层又通过上下叠加的形式来组成整个图像。操作时可以根据需要添加很多图层，方便地对图像的效果进行灵活调整，可以通过图层浮动面板来管理图层。

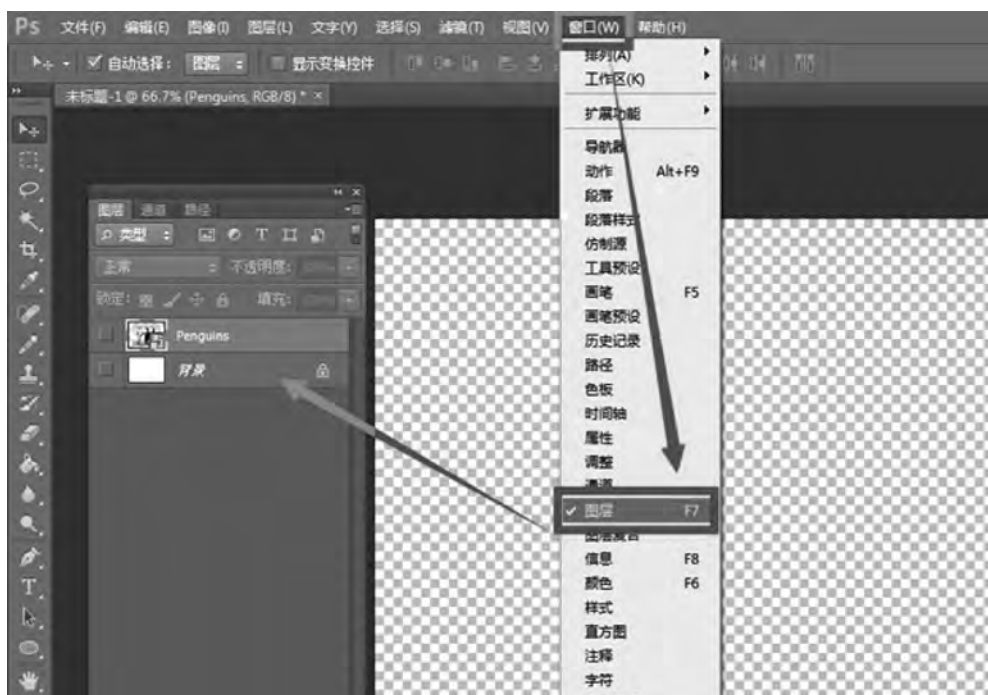


图 5-19 图层管理

5.3 图像数据标注工具和方法

数据标注是人类运用计算机等工具把各类型的数据如文本、语音、图像、视频等通过不同的标注方式为它们定义标签,并提供给机器学习的过程。在人工智能发展过程中,数据标注是必不可少的一个环节。此行业十分注重高效和准确,一个好用、功能强大的数据标注工具可以帮助团队节约成本,提高数据标注效率,为算法模型数据的获取如虎添翼。



微课视频

5.3.1 目标检测工具及使用方法

LabelImg 是一个可视化的图像标定工具。其主要特点是操作简单、使用方便。Faster R-CNN, YOLO, SSD 等目标检测网络所需要的数据集,均需要借此工具标定图像中的目标,且生成的 XML 文件是遵循 PASCAL VOC 格式的。

1. LabelImg 的安装环境

(1) 下载工具 Anaconda。下载地址为 <https://www.anaconda.com/products/individual#download-section>。

(2) 安装 Anaconda,选择默认安装目录将能勾选的都勾选上,选择自动添加环境变量。

(3) 如图 5-20 所示,利用 pip 安装 PyQt5。



图 5-20 安装 PyQt5 界面

(4) 如图 5-21 所示,安装 lxml,文件较小直接运行即可: pip install lxml。

```
C:\Users\DELL>pip install lxml
Requirement already satisfied: lxml in c:\programdata\anaconda3\lib\site-packages (4.4.1)
```

图 5-21 安装 lxml 界面

(5) 如图 5-22 所示,安装 LabelImg: pip install labeling。

```
C:\Users\DELL>pip install labeling
Collecting labeling
  Downloading https://files.pythonhosted.org/packages/91/04/3a48958976f3f82a1d21edcd26c87f8c753275a88578b9a9e4ca3a18a93/labeling-1.8.3-py2.py3-none-any.whl (248kB)
    | 256kB 5.6kB/s
Requirement already satisfied: pyqt5 in c:\programdata\anaconda3\lib\site-packages (from labeling) (5.13.2)
Requirement already satisfied: lxml in c:\programdata\anaconda3\lib\site-packages (from labeling) (4.4.1)
Requirement already satisfied: PyQt5-sip<13,>=4.19.19 in c:\programdata\anaconda3\lib\site-packages (from pyqt5->labeling) (12.7.0)
Installing collected packages: labeling
Successfully installed labeling-1.8.3
```

图 5-22 图层管理

(6) 在 cmd 中输入“labelimg”就会弹出工具主界面。

2. LabelImg 的使用方法

1) 配置预定义标签名称

预定义标签名称的配置,如图 5-23 所示需要到 labelimg.exe 文件所在路径下的 data\predefined_classes.txt 文本文件中进行填写或修改,按行存放,每行代表一个预定义标签名称。

2) 基本功能

在 LabelImg 窗口的左边有一些操作的功能,如图 5-24 所示。其中,Open 是打开单个图像; Open Dir 是打开文件夹; Change Save Dir 是图像保存的路径; Next Image 是切换到下一张图像; Prev Image 是切换到上一张图像; Verify Image 是校验图像; Save 是保存图像; Create RectBox 是创建一个标注框; Duplicate RectBox 是重复标注框; Delete RectBox 是删除标注框; Zoom In 是放大图像; Zoom Out 是缩小图像; Fit Window 是图像适用窗口; Fit Width 是图像适应宽度。当然,使用操作按钮不是很方便,下面介绍一些快捷键,可为无聊的标注工作节省一些时间。



图 5-23 LabelImg 标签预定义配置文本

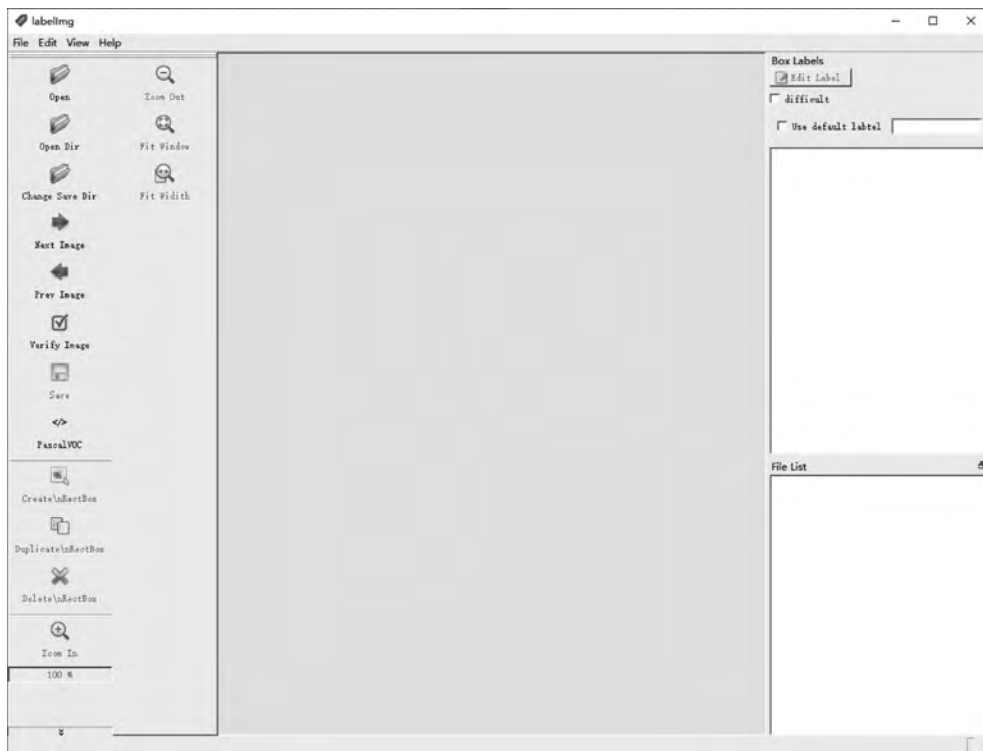


图 5-24 LabelImg 主界面和基本功能窗口

3. LabelImg 中常用的快捷键

LabelImg 常用快捷键如表 5-5 所示。

表 5-5 LabelImg 快捷键

快捷键	功 能
Ctrl+O	打开图像文件
Ctrl+U	打开图像所在的文件夹
Ctrl+R	更改结果保存的位置
W	开始画标注框
Ctrl+S	保存标注结果
D/A	下一张图像/上一张图像
Delete	删除画的标注框
Ctrl+(+)/(-)/(=)	放大/缩小/保持图像原始大小
Ctrl+E	编辑标签
Ctrl+Shift+O	打开的文件夹只显示 XML 格式的文件
Ctrl+L	修改标注框边线的颜色
Ctrl+Shift+S	将图像另存为
Ctrl+J	修改标注框
Ctrl+D	复制框
Ctrl+H	隐藏所有的标注框
Ctrl+A	显示所有的标注框
Space	标记当前的图像
Ctrl+D	复制当前标注框的标签和框
Ctrl+Q	退出软件
Ctrl+F	适合窗口的图像大小

4. 操作方法

(1) 单击 Open,Dir 按钮打开如图 5-25 所示需要标注的图像所在文件夹,打开第一张需要标注的图像。

(2) 单击 Change Save Dir 按钮确定 XML 文件保存路径。

(3) 单击 Create RectBox 按钮或按 W 键对图像中需要标注的部分进行画框,过程中需按住鼠标。

(4) 画完框后,松开鼠标左键,会弹出选择标签类别信息框。如图 5-26 所示,可以提前在预定义标签文档中编辑好所需用到的标签信息。

(5) 如图 5-27 所示,选择所需要的标注类别或者输入新的类别,然后单击 OK 按钮。

(6) 继续标注,直到一张图像的所有目标都标注成功以后,单击 Save 按钮,此时就在标注图像的文件夹下生成一个对应图像名的 XML 格式的文件,里面保存了标注信息。

(7) 单击 Next Image 按钮或按 D 键,对下一张图像进行标注。

5.3.2 图像分割工具及使用方法

labelme 可以定义为升级版的 LabelImg 工具,能够支持对图像进行多边形、矩形、圆、折线、点、语义分割等形式的标注,适用于目标检测、语义分割、图像分类等多个领域的数据标注任务。作为一款开源工具,labelme 布局简单,图形界面运用的是 pyqt。labelme



微课视频

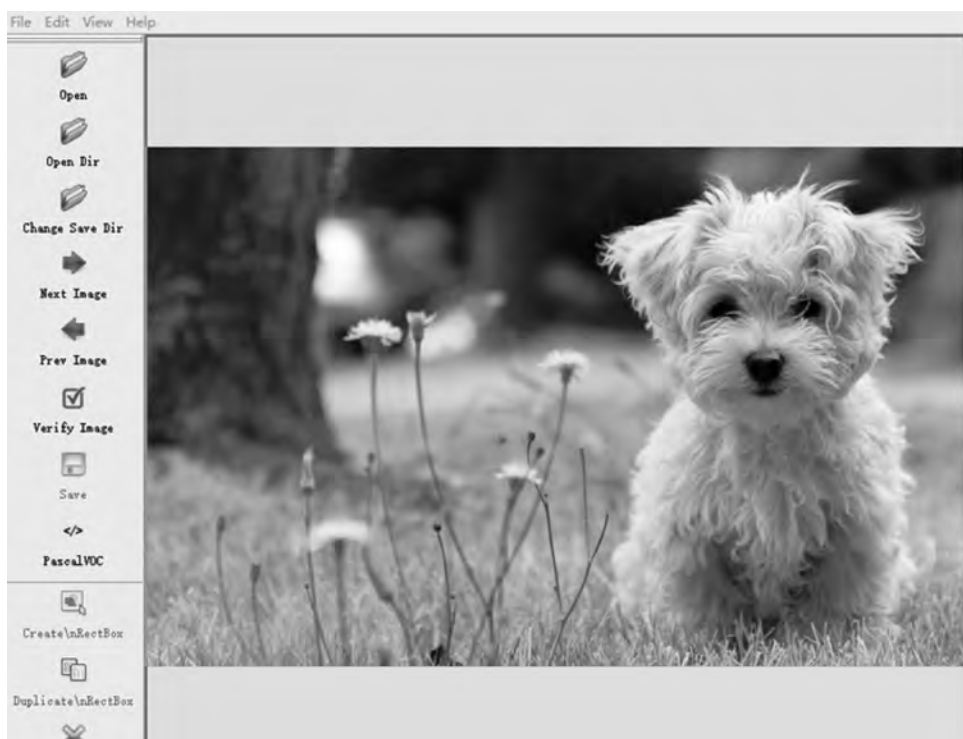


图 5-25 打开标注图像示例

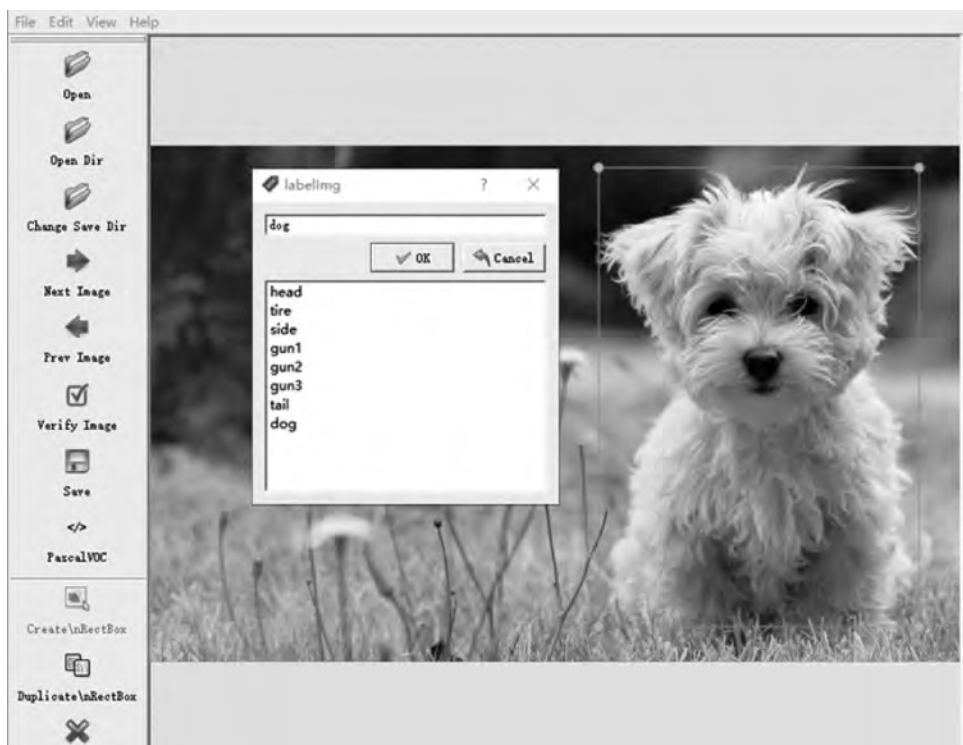


图 5-26 编辑标签示例

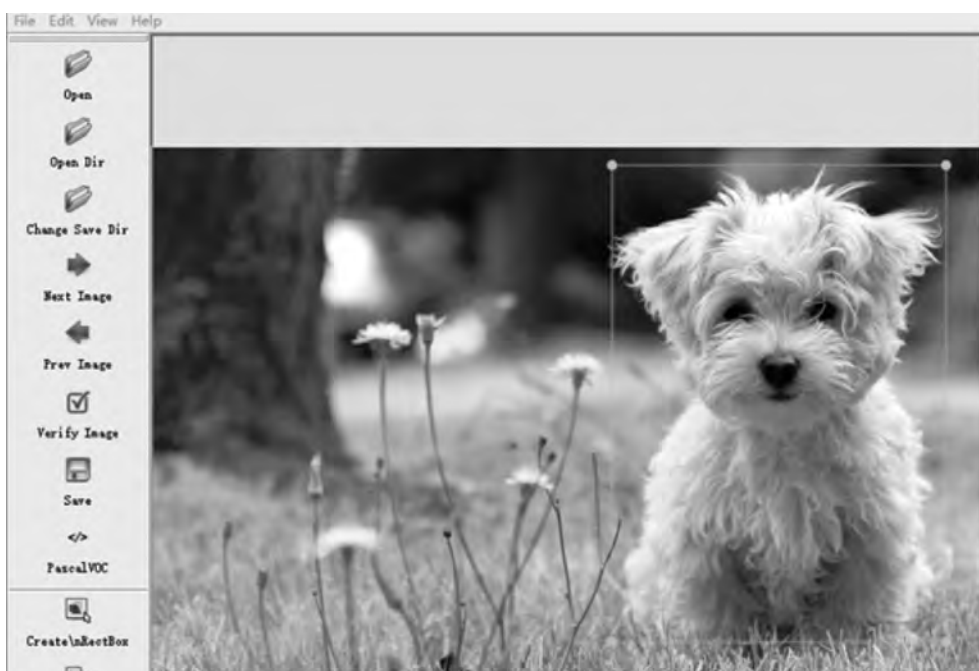


图 5-27 选择标签类型示例

可以生成 VOC 格式和 COCO 格式的数据集且以 JSON 文件格式存储标注信息。

1. labelme 的安装

- (1) 首先安装 Anaconda, 可以安装 Python 2.7 版本或 Python 3.6 版本。
- (2) 如图 5-28 所示, 安装成功后, 打开 Anaconda Prompt, 然后依次输入以下命令。

```
1 # python2
2 conda create --name=labelme python=2.7
3 source activate labelme
4 # conda install -c conda-forge pyside2
5 conda install pyqt
6 pip install labelme
7 # if you'd like to use the latest version. run below:
8 # pip install git+https://github.com/wkentaro/labelme.git
9
10 # python3
11 conda create --name=labelme python=3.6
12 conda activate labelme
13 # conda install -c conda-forge pyside2
14 # conda install pyqt
15 pip install pyqt5 # pyqt5 can be installed via pip on python3
16 pip install labelme
17
18 # Pillow 5 causes dll load error on Windows.
19 # https://github.com/wkentaro/labelme/pull/174
20 conda install pillow=4.0.0
```

图 5-28 Labelme 安装命令

(3) 成功后,在 Anaconda Prompt 中输入“labelme”,如果打开如图 5-28 所示的界面,则表明安装成功。

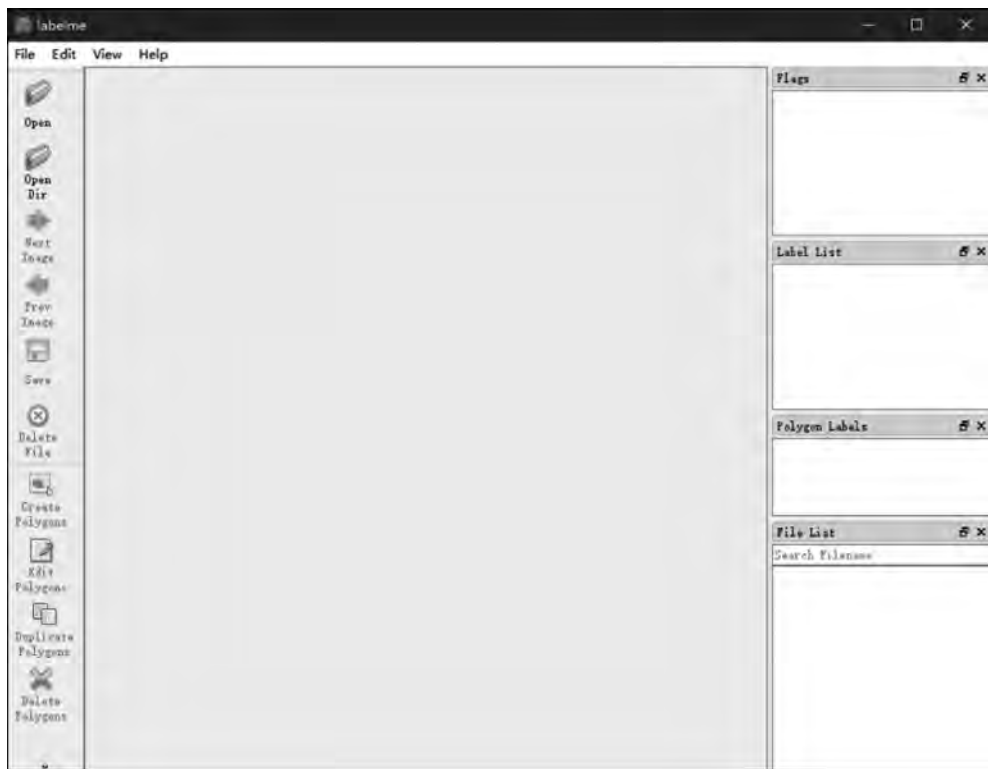


图 5-28 labelme 初始窗口

2. labelme 的基本功能

如图 5-29 所示,在 labelme 窗口的左边有一些操作的功能。其中,Open 是打开单个图像;Open Dir 是选择文件夹地址,打开多张图片;Next Image 是切换到下一张图片;Prev Image 是切换到上一张图片;Save 是保存.json 文件;Delete File 是删除.json 文件;Create Polygons 是绘制多边形边界;Edit Polygons 是编辑多边形边界;Duplicate Polygons 是复制选择的多边形边界;Delete Polygons 是删除选择的多边形边界;Undo 是撤销上一步;Zoom In 是放大图像;Zoom Out 是缩小图像;Fit Window 是图像适用窗口;Fit Width 是图像适应宽度。当然使用操作按钮不是很方便,下面介绍一些快捷键,可为无聊的标注工作节省一些时间。

3. labelme 的快捷键

为了提高标注的效率,labelme 提供了大量的快捷键。在 home 文件夹内能够看到一个隐藏文件“.labelmerc”,打开“.labelmerc”,就能查询到默认的快捷键,也可以根据习惯自定义操作快捷键。系统默认的快捷键如表 5-6 所示。

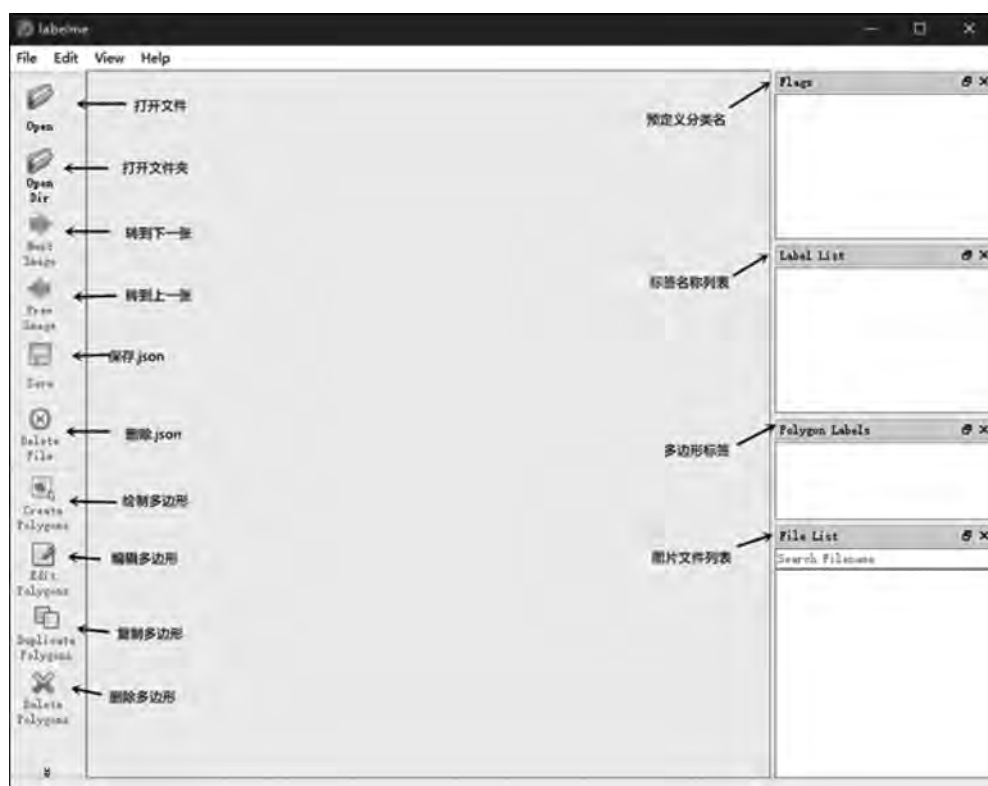


图 5-29 labelme 主界面

表 5-6 labelme 快捷键

快捷键	功 能
Ctrl+O	打开图像文件
Ctrl+U	打开图像所在的文件夹
Ctrl+S	保存标注结果
Ctrl+Shift+S	将图像另存为
Ctrl+W	关闭标注
Ctrl+Q	退出软件
D/A	下一张图像/上一张图像
Delete	删除画的标注框
Ctrl+(+)/(=)	放大/缩小/保持图像原始大小
Ctrl+F	适合窗口的图像大小
Ctrl+N	创建多边形标注框
Ctrl+R	创建矩形标注框
Ctrl+D	复制框
Ctrl+J	修改标注框
Ctrl+L	修改标注框边线的颜色
Ctrl+Shift+L	修改标注框填充颜色

4. 操作方法

(1) 单击 Open Dir 按钮打开需要标注的图像所在文件夹,打开第一张需要标注的图像。

(2) 使用 Create Polygons 绘制多边形边界,使用 BackSpace 键撤销一步。

(3) 首尾点合并输入类别名称。

(4) 选择所需的标注类别或者输入新的类别,然后单击 OK 按钮。如图 5-30 所示使用过的标签名会保留在列表中,可以直接选择。出现多个同类对象时,可以在标签名后加上数字以作区分。之后如果想要修改标签名,可以在右侧 Polygon Labels 栏中选择要修改的标签名,右击,选择 Edit Label 修改标签。



图 5-30 labelme 设置标签名(见彩插)

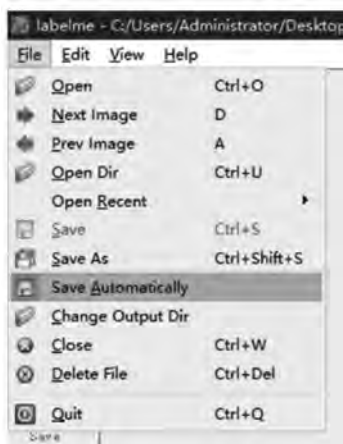


图 5-31 labelme 设置标签名

(5) 继续标注,使用 Ctrl+-、Ctrl++ 和 Ctrl+鼠标滚轮放大或缩小图片,使用 Edit Polygons 移动边界点。选中的边界点会变成正方形,请注意不要将整个掩码移动了。

(6) 直到一张图像的所有目标都标注成功以后,单击 Save 按钮,保存路径要与原图片路径相同,否则 labelme 无法读取保存后的.json 文件。此时就在标注图像的文件夹下生成一个对应图像名的 JSON 格式的文件,里面保存了标注信息。如图 5-31 所示可以在打开 labelme 后,通过 File 下拉菜单,选择 Save Automatically,自动将.json 文件保存在图片路径下。

5.4 图像数据标注案例

数据标注是一项熟能生巧的工作,需要通过大量的实战练习,才能够熟练掌握这门技术,并且更为精准地标注数据。本节分别对人体框图像、车辆识别和手势关键点进行数据标注进行介绍,涵盖了标框标注、分割标注和描点标注以及区域多边形标注。

5.4.1 人体框图像标注——目标检测

对于行人的标注是很多项目中都会出现的场景,需要将人体的各个特征属性都标注清晰,且标注框贴合。由于人体特征较为明显,因此这类型的标注大多数使用 LabelImg 标注工具。以十字路口检测行人通过目标检测算法为例。



微课视频

1. 人体图像数据要求

数据场景——各类十字路口行人通过时的图像

目标类型——各类人群,无人过路(负样本),路口人流量大和人流稀疏时都需要

数据环境——晴天、雨天、白天、夜晚

图片格式——JPG、PNG

数据量级——10 000(正负样本比例 19 : 1)

2. 数据集分类

(1) 参与标注——训练集: 8000 张(均为正样本)。

(2) 不参与标注——测试集: 1000 张(正负样本比例 1 : 1)。

3. 具体标注流程

1) 预定义标签名称说明

(1) 头部: 上至头顶,下沿为肩膀圆弧处,要包含肩膀圆弧处。

(2) 上半身: 上至脖子末端与肩膀交接处,下沿为上衣末端(如遇连体衣则预估到臀部位置),需要把手全部包含进上半身。

(3) 下半身: 上至臀部位置,基本框入臀部,下沿为脚底。

(4) 人体框: 上至头顶,下沿脚底,左至人体最左侧,右至人体最右侧。

2) 标注要求及注意事项

(1) 先进行标注框处理,画出框后再选择标签,并重复以上动作。

(2) 标注采用从上往下的顺序进行,依次为: 头部、上半身、下半身,最后为人体全身,标注完一个人后再标注另一个人。

(3) 距离较远或是模糊的人,只需要标注人体全身框,不需要标注头部、上半身和下半身。

(4) 除了全身框,头部、上半身和下半身看见多少就标注多少,看不见的不标注,框与

框允许重叠。

(5) 在标注人体全身框时,人体不可见部分能够默认为在人体全身框内。

(6) 头部、上半身和下半身框需要在人体全身框内。

(7) 人体全身框必须要贴合里面头部、上半身和下半身的框,不可过大或是压线,不可见区域除外。

(8) 人物穿着的服装和鞋帽需要框选在内,手拎物品和背包不需要框选。

3) 标注流程

(1) 打开 LabelImg 标注工具,单击 Open Dir 按钮,如图 5-32 所示打开需标注的文件夹。



图 5-32 用 LabelImg 标注工具打开行人图像

(2) 对图片中的行人进行画框标注,单击 Create RectBox 按钮,如图 5-33 所示标注矩形框。需要标注 4 个框。



图 5-33 对头部进行标框标注(见彩插)

(3) 依次框选完毕,如图 5-34 所示添加标签,分别为头部、上半身、下半身和人体全身。



图 5-34 对行人上半身进行标框标注

(4) 检查标框的头部、上半身、下半身以及全身框的标签名称是否正确,如图 5-35 所示与目标边缘吻合,全身框是否压线,对不可见部分的处理是否合理。



图 5-35 对行人下半身进行标框标注(见彩插)

(5) 图片中所有目标标注完毕之后,单击 Save 按钮保存,如图 5-36 所示,单击 Next Image 按钮标注数据集的下一张图片,直至数据集标注完毕。

5.4.2 车辆轮廓抠图——图像分割

如今车辆智能驾驶或者智慧停车都已经深入人们的生活,并且越来越便捷。从之前的车辆检测到现在的自动驾驶,人工智能算法有了质的飞跃,这无疑是数据更加繁多更加



微课视频



图 5-36 对行人全身进行标框标注

精细的结果,因此较为简单的矩形框标注已经无法满足高质量高精度的算法模型了。更为精细的多边形分割技术,逐渐被广大算法训练师所接受。对于车辆分割标注方面的数据项目,具体流程如下。

1. 车辆图像数据要求

数据场景: 各类道路上的车辆图片(城市道路、高速公路等)
 数据环境: 晴天、雨天,日间、黑夜
 拍摄角度: 路边监控摄像头、行车记录仪、驾驶室内从外拍摄
 数据质量: 车辆轮廓清晰、车速过快导致车辆轮廓模糊(负样本)
 图像格式——JPG、PNG
 数据量级——5000(正负样本比例: 9 : 1)

2. 数据集分类

- (1) 参与标注——训练集: 4500 张(均为正样本)。
- (2) 不参与标注——测试集: 500 张(正负样本比例 1 : 1)。

3. 具体标注流程

1) 标注要求及注意事项

- (1) 分割点描绘需要尽量贴近目标,间距控制在 1 像素以内。
- (2) 尽量多拐点,保证标注范围内不存在其他特征信息。
- (3) 如图 5-37 所示有一些车辆在其本体上安装或镶嵌了一些设施,在分割标注时,应将其框选在内。

(4) 如果是放置或捆绑在车辆本体上的物品,如图 5-38 所示则不应该在分割时框选在内。



图 5-37 镶嵌设施分割标注示例(见彩插)



图 5-38 捆绑物体分割标注示例(见彩插)

(5) 当需要分割标注被物体遮挡隔断的车辆时,如图 5-39 所示被隔断的部分也应框选出来,并与车整体轮廓相连接。

(6) 当有人员站立或者遮挡住车辆时,应不选取人员图像,如图 5-40 所示直接预判出遮挡的车辆图像,框选出车辆轮廓。



图 5-39 隔断物体分割标注示例(见彩插)



图 5-40 遮挡物体分割标注示例(见彩插)

2) 标注流程

(1) 打开编译好的 labelme 多边形车辆标注工具,如图 5-41 所示双击图片路径和 XML 数据路径,确定数据集及标签文件存储位置。

(2) 如图 5-42 所示在右侧视觉姿态选取栏内选择车辆停放姿态,例图中停放姿态为 LR。

(3) 如图 5-43 所示选择目标车辆颜色和遮挡及困难属性,例图中为白色无遮挡。

(4) 如图 5-44 所示选择车辆类型,例图中为轿车车型。

(5) 选取车辆目标类型,如图 5-45 所示为整车标注。

(6) 车辆属性选取完毕之后,按住鼠标左键,如图 5-46 所示将目标框选在内,需贴边框选。



图 5-41 车辆分割标注工具界面(见彩插)



图 5-42 车辆视觉姿态选择(见彩插)



图 5-43 颜色、遮挡及困难属性选择(见彩插)



图 5-44 车辆类型选取(见彩插)



图 5-45 车辆目标类型选取(见彩插)



图 5-46 车辆分割目标框选(见彩插)

(7) 框选完毕之后,单击“开始”按钮,在目标框内开始勾勒车辆轮廓。

(8) 勾勒完毕车辆目标后,在最后线条闭环处单击鼠标右键,如图 5-47 所示,然后单击“结束”按钮,完成车辆分割标注。

(9) 标注完毕之后,如图 5-48 所示,单击“保存”按钮,进行下一张标注,直至数据集标注完毕。



图 5-47 车辆目标分割标注结果(见彩插)



图 5-48 车辆分割标注完毕结果(见彩插)

5.4.3 手势特征点——点位标注

现如今各类交通指挥终端,或者场景动作识别系统,都会运用到人物手势识别,去分析手势动作,从而达到指挥机器设备运作的功能。因此这类型数据标注为特殊数据标注,具有特定的标注规则。

1. 数据要求

手势类型——手势展示数字样式(一到十)、常用手势语,如“ok”、剪刀手、大拇指等

数据类型——手势无遮挡、手势存在遮挡(负样本)

图像格式——JPG、BMP、PNG

数据量级——5000(正负样本比例 9 : 1)

2. 数据集分类

(1) 参与标注——训练集: 4500 张(正负样本比例: 8 : 1);

(2) 不参与标注——测试集: 500 张(均为正样本)。

3. 具体标注流程

1) 手势标注规则和注意事项

(1) 需要标注的手部关键点有 21 个,具体位置和说明如图 5-49 所示。且每张图都是 21 个点,从 0 到 21。

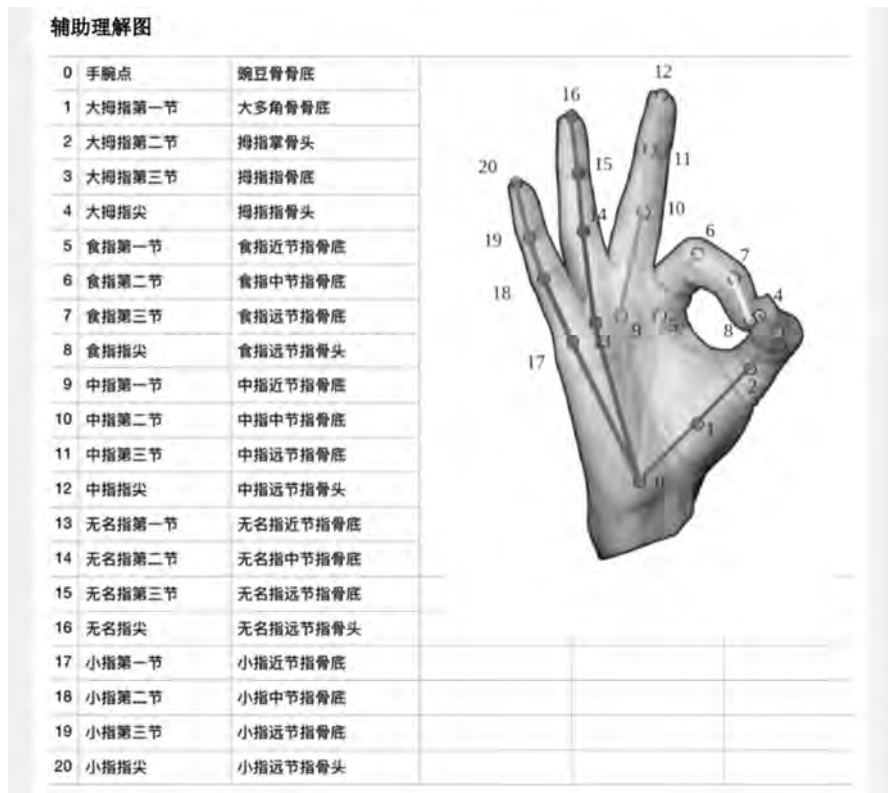


图 5-49 手势点位图(见彩插)

(2) 手部的整体矩形框要贴合手,边缘要紧贴手指尖和手腕,但不能露出来。

(3) 需从大拇指开始标注,按照手指头关节从下往上的顺序进行标注,第 21 个点的位置在手腕末端关节处。

(4) 每个关键点的标记位置应为关节中心处,如果遇到关节被遮挡或图片残缺无法判断出关键点特征的情况,则需要自行合理推测关节所在的位置,然后进行标注。

2) 具体标注流程

(1) 打开 labelme 图像标注工具,单击 Open Dir 按钮,打开文件夹里的数据。

(2) 首先需对手部的整体进行框选。单击主菜单中的 Edit 按钮,并在下拉菜单中选择 Create Rectangle 选项,框选完毕之后定义标签名称为“hand”。

(3) 接下来再进行 21 个指关节标注。单击主菜单中的 Edit 按钮,在下拉菜单中选择 Create Point 选项。找到手部的手腕处,标注为 0 号点,由此开始,依次进行连续标注。每根手指 4 个点,包括手指的 3 个指节和指尖。

(4) 在每一个关键点标注时,都要选择此关键点是可见或不可见,遇到有遮挡的手势时,要预估出关键点位的所在位置,被遮挡的手指如果为握拳形态,一律默认为握的最紧。

(5) 按照顺序标注号 21 个指关节,并标记号相应的序号标签。

(6) 21 个指关节标注完成后,单击左侧的 Save 按钮,保存完成后。如图 5-50 所示,单击 Next Image 按钮,继续进行下一张手势图像标注,直至标注完成数据集所有手势。

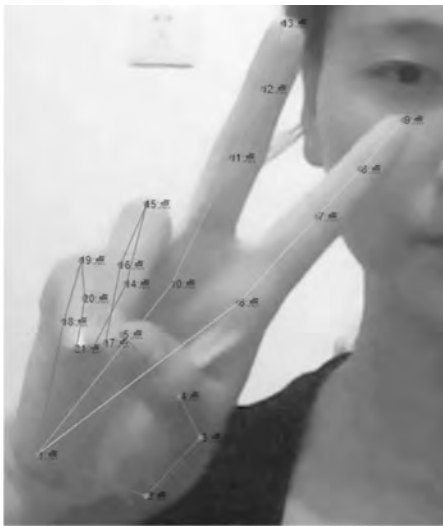


图 5-50 手势 21 节点标注完成结果(见彩插)

思考题

1. 图像标注的定义是什么?
2. 影响图像数据标注质量的因素来源于哪几方面?
3. 说出几种常见的图像文件格式。