ICS 35.020
L70

团　　体　　标　　准

T/ISC 0006—2020

基于 AI 的多媒体内容识别基本要求

Multimedia content recognition requirements based on AI technology

2020－09－24 发布　　2020－12－01 实施

中国互联网协会发布

目　　次

前　言

本文件按照 GB/T 1.1-20200《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由中国互联网协会标准工作委员会提出并归口。

本文件起草单位：深圳市腾讯计算机系统有限公司、恒安嘉新（北京）科技股份公司、深圳市网安计算机安全检测技术有限公司。

本文件主要起草人：盛安宇、庞永杰、鞠奇、黄申、肖万鹏、孙子荀、王永霞、代威、黄超、洪跃腾。

基于AI的多媒体内容识别基本要求

1 范围

本文件给出了基于AI的多媒体内容识别的技术框架和基本要求，包括不限于多媒体视觉技术、多媒体内容文字识别技术、多媒体语义理解技术等，在语音、视频、图片、NLP等多媒体内容上的检索识别能力，以及多媒体内容识别中的数据安全能力要求。

本文件适用于互联网企业、政府、科研单位开展基于AI的多媒体内容识别的设计、开发、应用等方面的指导和参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35273—2020 个人信息安全规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

恶意信息 malicious information

指国家相关法律法规中规定的，以及根据具体业务场景不同而对业务有负面影响的信息。

4 缩略语

以下缩略语适用于本文件：

AI 人工智能 (Artificial Intelligence)

ASR 自动语音识别（Automatic Speech Recognition）

NLP 神经语言程序学(Neuro-linguistic programming)

OCR 光学字符识别（Optical Character Recognition）

5 概述

本标准给出了多媒体内容识别的基本技术架构，包括:图片识别、文本识别、视频识别，以及与数据标注系统的接口等。多媒体内容识别基本技术架构见图1。

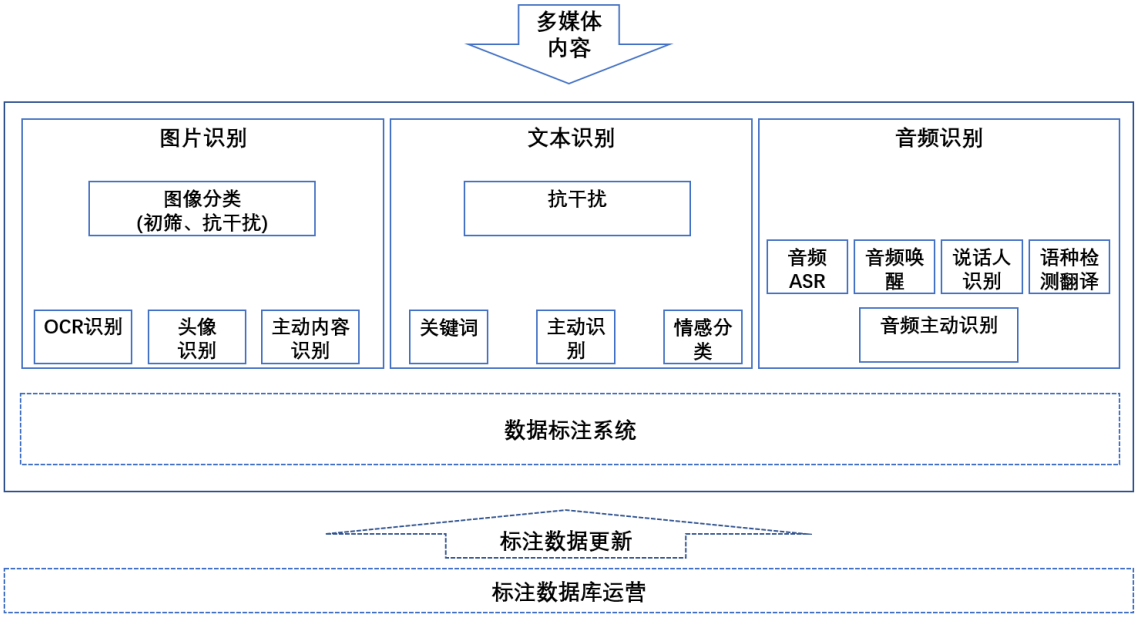

图1　多媒体内容识别技术架构

图片识别方面，通过图片分类对干扰图片进行过滤、对图片进行初步分类，再将内容送入相应的识别模块，如OCR识别、头像识别、主动内容识别等。

文本识别方面，通过文本抗干扰进行过滤，过滤、替换干扰内容，再将内容送入相应的识别模块中，如关键词、文本主动识别、情感分类等。

音频识别方面，通过音频ASR、音频唤醒、说话人识别、语种检测翻译等模型，识别特征明显的音频样本，再将不易识别样本通过核心主动识别模型进行识别。

视频识别方面，通过截帧、抽取音频流、抽取文字流的方式，通过对视频中的图片、文本、音频内容的识别得到对视频进行识别的综合结果。

6　基于 AI 的多媒体内容识别技术要求

6.1　图片识别

6.1.1　初筛图片分类

初筛图片分类，是将图片按照不同的类别进行初步分类的算法模型。分类方法包括，但不限于：

——按照内容排版分类：图文混合类、纯文本类图片、拼图类图片等；

——按照内容分类：色情类图片、暴力类图片、低俗类图片等；

——按照类型分类：漫画类图片、场景类图片、人群类图片、物品类图片等。

技术要求如下：

a)　应支持多种格式的图片的读取；

b)　应具有图片旋转、缩放的鲁棒性，不会因为图片的旋转、缩放导致算法结果有较大差异；

c)　应支持基于多维度的图片初步分类模型；

d)　应支持图片分类类别的扩展；

e)　应对干扰因素具有一定的鲁棒性，例如 干扰线、噪点、局部马赛克等；

f） 应提供训练图片模型的接口，进行图片标注和训练；
g） 应提供低延迟、高并发的模型算法；
h） 应能发现并识别诸如信息隐写、图床等信息藏匿技术藏匿的信息。

6.1.2 OCR 识别

OCR识别是对图片进行文字识别，为恶意信息识别提供最准确的文字内容。技术要求如下：
a） 适应多种业务场景的图片文字分布；应具备各种方向旋转体的图片OCR识别能力；
b） 多图文本行混合，具备分级批识别能力；
c） 应支持对各种文字变体，如 宋体、幼圆、黑体 等计算机字体或手写字体的识别能力；
d） 应支持对不同大小字体的识别能力；
e） 应对干扰因素具有一定的鲁棒性，例如复杂背景、干扰线、噪点等；
f） 应提供训练、标注图片模型的接口，进行图片标注和训练。

6.1.3 头像识别

头像识别是为业务场景的图片提供敏感人物识别，技术要求如下：
a） 具备敏感人物的人脸识别能力，并可以识别PS或者带装饰品（口罩或墨镜等）人脸、侧脸、人脸以及模糊人脸等各种类型；
b） 可以共享识别网络底层参数，并使用独立分支的参数进行特定人物的分类识别；
c） 应对干扰因素具有一定的鲁棒性，例如 漫画人脸、模糊图片等；
d） 应提供敏感人物模型的运营接口，支持敏感人物的增加、删除、更新、查询能力，进行敏感人物模型的训练、更新、优化。

6.1.4 主动内容识别

主动识别是指根据具有恶意目的业务的特性，将业务分类细化，在不同业务场景下，对不同的分类做不同的处理。技术要求如下：
a） 应具备识别不同类型恶意内容的能力,包括不限于：色情类图片、暴力类图片、血腥类图片、违法违禁商品类图片；
b） 具备基于多维度模型的扩展能力，可以整合多个模型提供统一的识别结果；
c） 应具备相似样本的识别能力，能够基于已经发现的样本图片，识别相似或相近的内容的能力；
d） 对干扰因素具有一定的鲁棒性，例如 拼图图片、局部马赛克图片等；
e） 应提供训练图片模型的接口，进行图片标注和训练。

6.2 文本识别

6.2.1 抗干扰模型

文本预处理模型是针对文本中的对抗文字进行预处理、过滤的基础模型，技术要求如下：
a） 应具备识别不同类型的干扰文本的能力,包括不限于：拆分组合字替换、同音字识别替换、同义词替换、象形字替换、干扰符替换、表情符替换、特殊字符替换等能力；
b） 应具备对不同字符编码的处理能力，可将不同的字符编码转换成统一的字符编码；
c） 应具备模型扩展能力，可以针对不同类别的字符进行替换、过滤；
d） 应提供文字样本扩展的接口，进行文字样本的增加、修改、删除、查询能力；
e） 应支持发现不在已知范围内的新词识别，例如：互联网流行语言、小众群体黑话等。

6.2.2 关键词识别

关键词模型是基于关键词进行文本内容识别的基础模型，技术要求如下：

a) 应具备多关键词的匹配、反匹配、多重组合等能力；
b) 应具备按照不同类别、不同业务进行关键词分库查询匹配的能力；
c) 应具备多级词库匹配的能力，例如：提供通用词库、业务词库等词库的组合查询。同时也应具备多级词库的管理能力；
d) 应具备关键词的增加、删除、查找、更新 能力，进行关键词的运营。

6.2.3 文本主动识别

文本主动识别模型是基于文本语义，对文本内容进行识别的模型，技术要求如下：

a) 应具备基于文本语义进行文本内容识别的能力，包括不限于：色情类文本、违法类文本、暴力恐怖类文本、涉政类文本、谣言类文本、辱骂类文本、广告引流类文本的识别能力；
b) 应具基于多维度模型的扩展能力，可以整合多个模型提供统一的识别结果；
c) 应具备相似样本的识别能力，能够基于已经发现的段落样本，识别相似或相近的内容的能力；
d) 应提供文本样本扩展的接口，进行文字样本的增加、修改、删除、查询能力。

6.2.4 情感分类

情感分类是基于文本语义，对文本描述的主观情绪进行分析识别的模型，技术要求如下：

a) 应具备基于文本语义进行基本情绪识别的能力，包括不限于：正面类情绪、负面类情绪、中性类情绪的识别；
b) 应具备基于多维度模型的扩展能力，可以整合多个模型提供统一的识别结果；
c) 应提供语料库扩展的接口，进行语料样本的增加、修改、删除、查询能力。

6.3 音频识别

6.3.1 说话人识别

说话人识别，是生物特征识别技术的一种，指利用人的发声特征进行身份鉴定的技术，因此也被成为“声纹识别”，用于识别特定人群的语音信息。技术要求如下：

a) 应具备基于声纹信息，识别具体说话人的能力；
b) 应具基于多维度模型的扩展能力，可以整合多个模型提供统一的识别结果；
c) 应具备基于已有添加的说话人语料库样本识别同一说话人所说语音的识别能力；
d) 应提供语音语料样本库的扩展接口，进行语音语料库的增加、修改、删除、查询能力；
e) 应对干扰因素具有一定的鲁棒性，例如 干扰背景音乐、远场说话、耳语说话等。

6.3.2 语音关键词唤醒

语音关键词唤醒通常指设备在休眠状态下，使用预定义关键词唤醒设备从而进行操作的过程。在音频识别领域中，将语音唤醒技术应用在互联网音视频关键词的检测中。技术要求如下：

a) 应具备语音关键词的匹配能力；
b) 应具备按照不同类别、不同业务进行关键词分库匹配的能力；
c) 应具备多级词库的组合匹配的能力，包括多级词库的管理能力；
d) 应具备相应关键词的增加、删除、查找、更新能力，并开展关键词运营；
e) 应对干扰因素具有一定的鲁棒性，例如干扰背景音乐等。

6.3.3 语种检测/翻译

语种检测/翻译是指检测不同语种的内容，将语音内容翻译为固定的某种语言以便后续进行识别的能力。

a) 应具备多个语种的检测、翻译能力，包括不限于：英语、维语、藏语等预言；
b) 应对干扰因素具有一定的鲁棒性，例如干扰背景音乐等；
c) 应提供语料训练的接口，进行语料样本的标注、训练，以及对模型的优化。

6.3.4 ASR 唤醒

语音ASR是指将输入的语音信息，输出转换成可阅读的文字的算法模型。技术要求如下：

a) 应具备对特定语种的文字转换能力。包括不限于汉语；
b) 应对干扰因素具有一定的鲁棒性，例如干扰背景音乐、耳语、远场说话等；
c) 应具备对识别内容进行语义修正的能力，即通过上下文关联性，对识别的文本内容进行修正，选择正确的同音字、同音词的能力；
d) 应提供语料训练的接口，进行语料样本的标注、训练，以及对模型的优化。

6.3.5 音频主动识别

音频主动识别是指将输入的语音信息，不经过文字转换就可进行识别分类的模型。技术要求如下：

a) 应具备基于音频信息进行音频主动识别的能力，包括不限于娇喘类音频、色情类音频、政治类音频等。
b) 应具基于多维度模型的扩展能力，可以整合多个模型提供统一的识别结果。
c) 应具备相似样本的识别能力，能够基于已经发现的音频段落样本，识别相似或相近的内容的能力。
d) 应提供语料训练的接口，进行语料样本的标注、训练，以及对模型的优化。
e) 应对干扰因素具有一定的鲁棒性，例如干扰背景音乐、远场说话、录制声音、音频伪装等。

7 标注数据更新接口

7.1 数据增删改查接口

数据标注更新接口，是经过人工审核将标注好的样本和标注样本库进行交互的一系列接口，用于对标注样本库进行操作。技术要求如下：

a) 应具备增加、删除、查找、修改已经标注的音频、视频、图片、文本、文件、链接图文综合媒体等样本及其附属信息的能力。
b) 应具备批量进行增加、删除、修改、查询的功能。
c) 应具备相应的日志记录功能。
d) 应具备相应的人员权限管理功能。
e) 应具备自动去重能力，当发现样本库样本相似性较高的情况下，可以标记或提醒添加人注意相关

样本的添加。

7.2 数据多租户接口

数据的多租户接口是对于标注样本库，支持生成多个样本库，并可对多个样本库进行操作的接口。技术要求如下：

a) 应具备增加、删除、查找、修改某一个音频、视频、图片、文本、文件、链接图文综合媒体等样本的样本库及其附属信息的能力；
b) 样本库中的样本应可独立运营，不受其他样本库的干扰。包括不限于其他样本库中样本的增加、删除、修改，以及其他样本库的增加、删除、修改；
c) 同一个或相近似的样本，应支持存在于一个或多个的样本库中；
d) 样本库应该支持对多租户生成权限不重叠的样本库。

8 基于 AI 的多媒体内容识别的数据安全

8.1 数据采集

数据采集环节指数据获取和创建过程。要求如下：
a) 应告知用户被收集的数据类型、使用目的，并获得用户授权许可；
b) 应经用户同意，采集用户个人信息，不应欺诈、诱骗、强迫个人提供信息。

8.2 数据存储

数据存储环节指将数据持久保存在系统中，数据存储应采取必要的安全措施，确保数据存储系统安全及数据自身安全，要求包括不限于:
a) 应及时安装更新或补丁；
b) 应管理移动介质和移动办公设备；
c) 应管理数据的访问权限，按照数据的重要性、敏感性等因素进行相应的账号管理和权限分配；
d) 应对于重要或敏感数据加密存储、重要数据进行备份。

8.3 数据传输

数据传输环节指数据在组织内外部的流转过程，应确保以下安全措施：
a) 应确保数据传输链路安全采用加密传输技术或协议；
b) 应设置数据传输冗余链路。

8.4 数据加工

数据加工指对采集的原始数据进行计算分析，加工生成新的数据，数据加工环节应确保以下安全措施:
a) 按照数据采集环节对用户的承诺，遵守承诺范围使用用户数据进行挖掘分析；
b) 在进行数据分析和挖掘的过程中，对个人敏感信息采取数据脱敏等措施，保护用户个人权益或用户隐私；
c) 进行数据挖掘分析后形成的数据应采取相应的安全保护措施，如加密等。

8.5 数据转移

数据转移指将原始数据、加工过的数据等不同形式的数据分发给外部实体，数据转移环节应确保如下安全措施：
a) 将数据提供给第三方前，对第三方的数据安全防护能力进行评估，明确告知第三方数据使用权限，并对第三方进行必要的安全监督管理；
b) 将数据提供给第三方前，对将要转移的数据的重要性、敏感性进行评估。

8.6 数据删除

数据删除指删除组织的数据及其副本。数据删除环节应确保以下安全措施：

a) 当用户终止使用服务后，按法律法规要求留存；

b) 按用户要求删除用户数据的。

8.7 个人信息安全

应遵循GB/T 35273—2020《个人信息安全规范》国家标准的要求，并采取必要安全措施。