Fusion Finders

开发文档

Fusion Finders 团队
2023 年 7 月 13 日

目录

一、绪论

1.Fusion Finders 系统开发背景

视频、照片是生活中常见的图像数据。随着智能手机、相机和高速移动网络的发展，人们使用照片、视频来记录生活的成本越来越低，因此产生了大量的图像数据。同时，在安保、野生动物保护等领域，也产生了大量需要分析、检索的图像数据。

与此同时，产生了大量对视频片段检索并对物体进行定位的需求，如对校园监控视频的片段检索、对公共环境的视频片段检索等。而现今多数视频检索任务需要人工手动完成，因此构建一个自动化分析选取视频片段的软件具有较大的应用价值。

2.项目主要工作

Fusion Finders 系统是一次多模态的创新探索，该系统融合了语

音、文本、视频和图片四个模态，可以通过语音或文本对视频和图片进行提问，找到所需片段或指定目标。

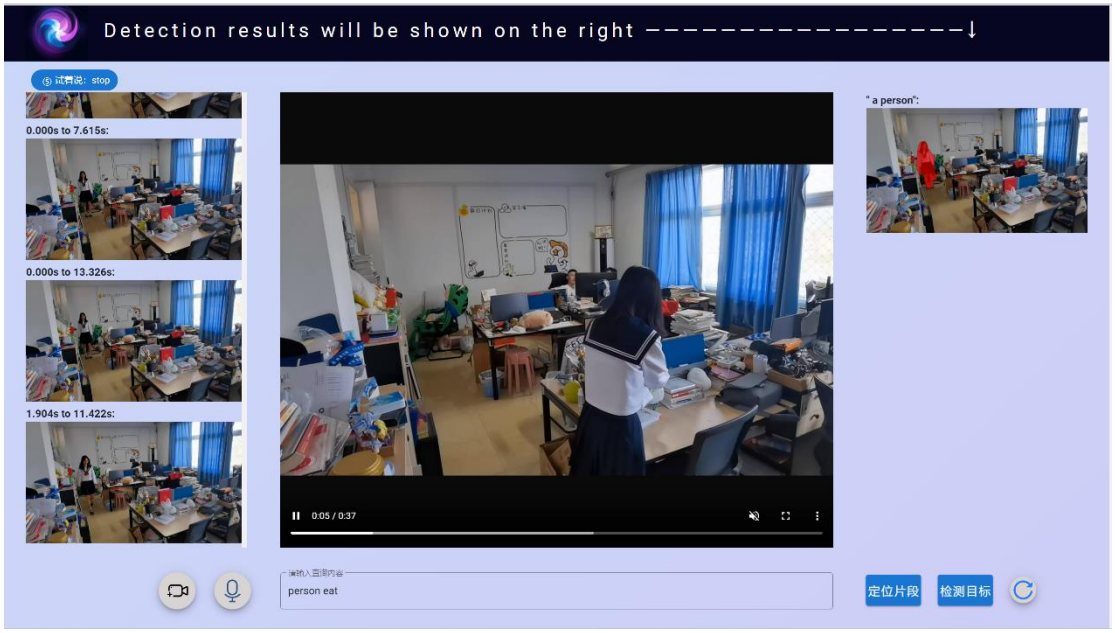
针对于软件需要完成的功能，对本项目的工作内容分析如下：

（1）语音识别

为了让用户更方便地控制软件，我们需要为软件添加语音识别模块。用户点击搜索框旁的“语音”按钮，即可开始语音输入，再次点击即可结束语音。结束后，软件会自动将语音转为文字并执行相关操作。

（2）视频片段定位

用户在软件中上传需要分析的视频，并通过语音或文字输入筛选条件，定位目标视频片段。点击“定位片段”按钮后，软件将调取相关模型寻找满足要求的片段，并最终将最符合描述的 5 个片段返回给用户。



（3）指代物体定位

当用户需要定位视频片段中的物体或目标生物时，可以点击“目标定位”按钮，软件系统将自动根据用户先前输入文本中的指代物体进行检测，并标记到已选好的片段。

3.项目文档组织结构

第一章为绪论，主要讲述项目的背景、工作及结构；第二章为需求分析，从用户角度分析系统的需求度。第三章主要从架构方面描述系统设计，第四章从各结构的细化层面上描述系统设计，第五章讲述系统的实现方式、测试情况及后续的维护工作。

二、Fusion Finders 系统需求分析

1.系统需求介绍

1.1 项目特点

本项目融合了语音、文本、图像及视频四种模态，结合语音识别、文本翻译、视频片段定位及指代目标检测等多种技术实现了一个集成度高、易操作、可视效果好的系统。与此同时，系统针对不同的场景设计了特殊的入口，在安全保证、人文关怀、生态保护以及交通安全等方面构建了有效的应用场景。

1.2 整体需求

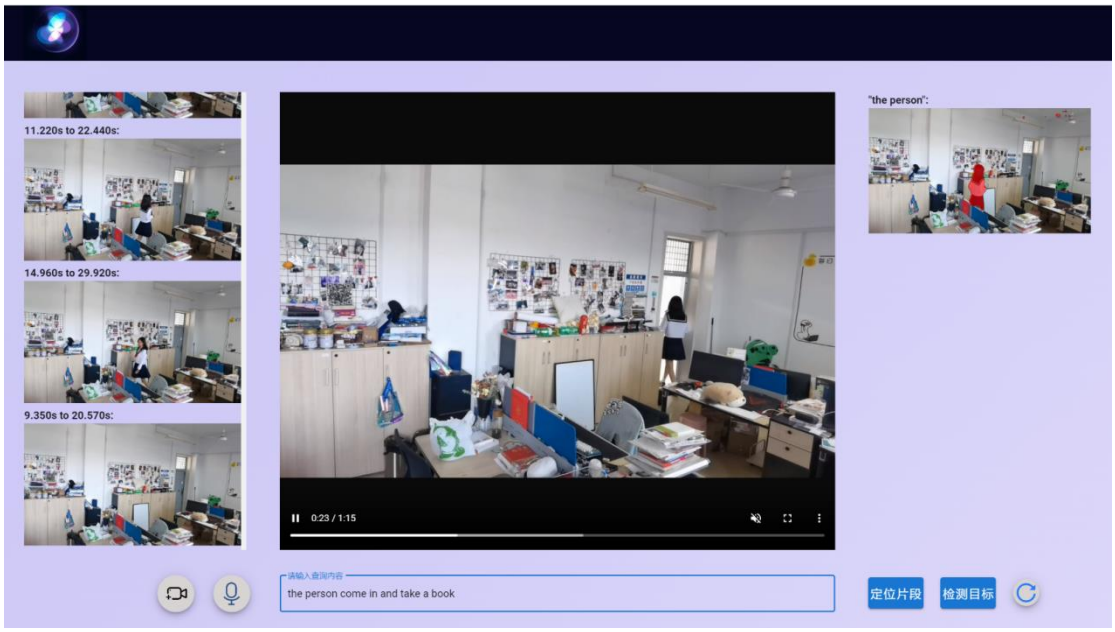
对项目的整体而言，应满足的需求有如下几方面：

1. 语音识别模块。为了更方便、更人性化地对系统进行控制，系统应该添加语音识别模块，通过语音对系统进行输入和控制。与此同时，应该尽量保证用户的输入数据语音和用户的控制系统语音不重合，做到智能、便捷。
2. 视频片段定位模块。用户上传一条视频，随后通过语音或文本提出视频检索条件，系统能够根据用户的描述自动检索视频中的信息，并将检索出来的片段在前端进行展示。
3. 指代目标检测模块。当用户在浏览视频时，用户可以通过语音或文本输入一个想要检测的目标，系统应该能够将用户所指代的目标检测出来，并以可视化方式显示在前端。
4. 翻译模块。用户可以输入母语，系统将自动识别用户的输入并自动翻译为有利于提高检索精确度的语言。

2.详细场景需求分析

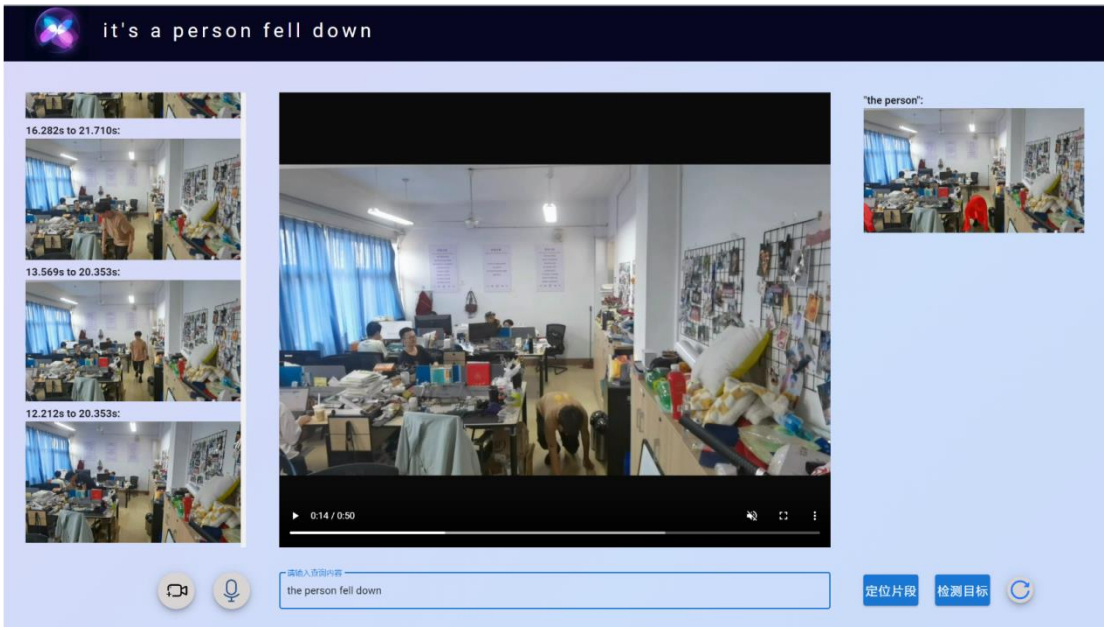
2.1 安全保证--监控检索

当发生诸如用户丢失物品等需要调取监控进行查看的情况时，传统的方法是通过人力来检索，而此方法耗时耗力。因此，此软件需要实现针对一段监控数据，用户输入场景后，可以自动在较短的时间内完成片段的检索，查看对应的片段。



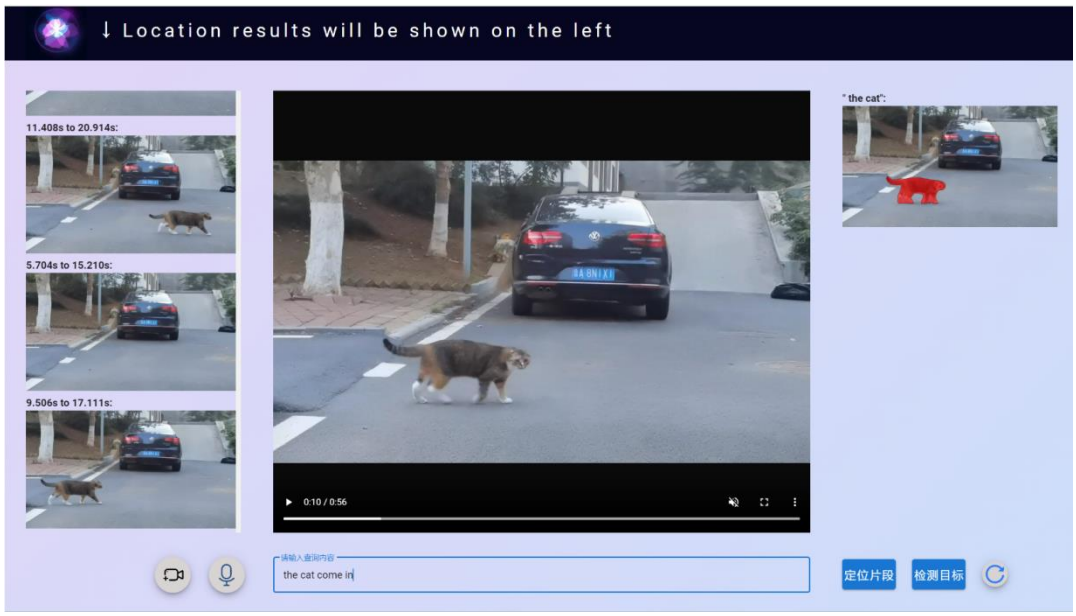
2.2 人文关怀--危险检测

此软件希冀可以为保护特殊人群做出贡献，即通过实时检测特殊人群是否处于险境来提供救援。当特殊人群发生危险时，如独居老人摔倒，此软件可以根据监控视频检测到摔倒动作的发生，并记录摔倒的方位、状态，有助于及时的抢救。与此同时，软件可以检索到危险发生的时间和地点，方便后续问诊。



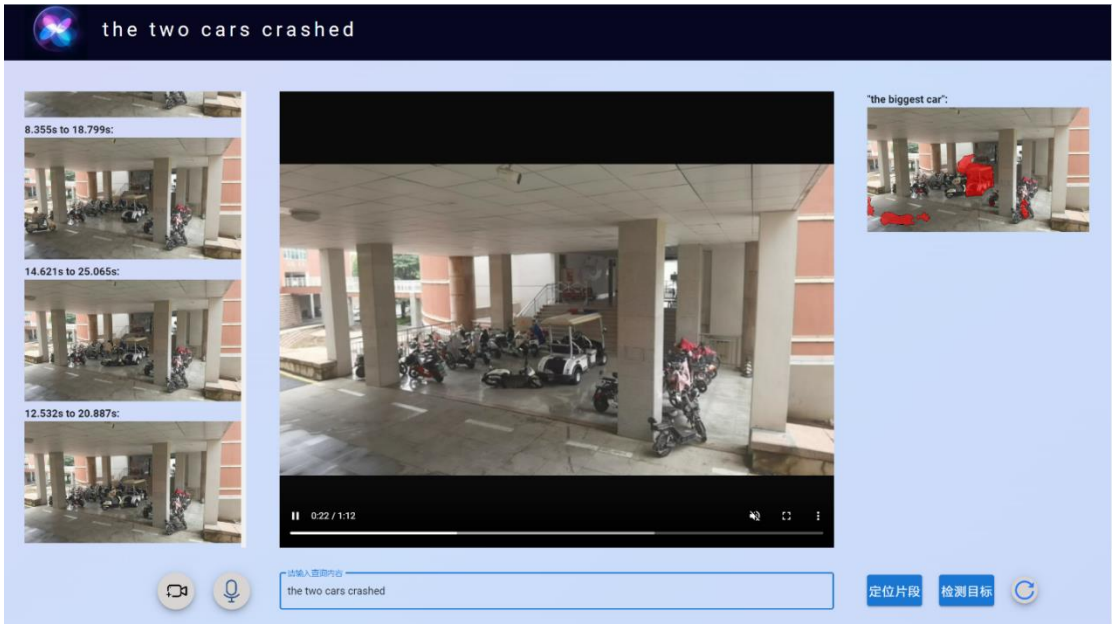
2.3 生态保护--动物检测

对生态研究者来说，他们需要获取野生动物或珍稀物种的视频图片，因此需要长时间将相机置于他们可能的出现区域进行录制，而录制之后对视频的检索同样耗费人力。因此，此系统可以实现对一段视频片段中指定动物的检测，并及时返回结果，节省时间和人力。



2.4 交通安全—事故分析

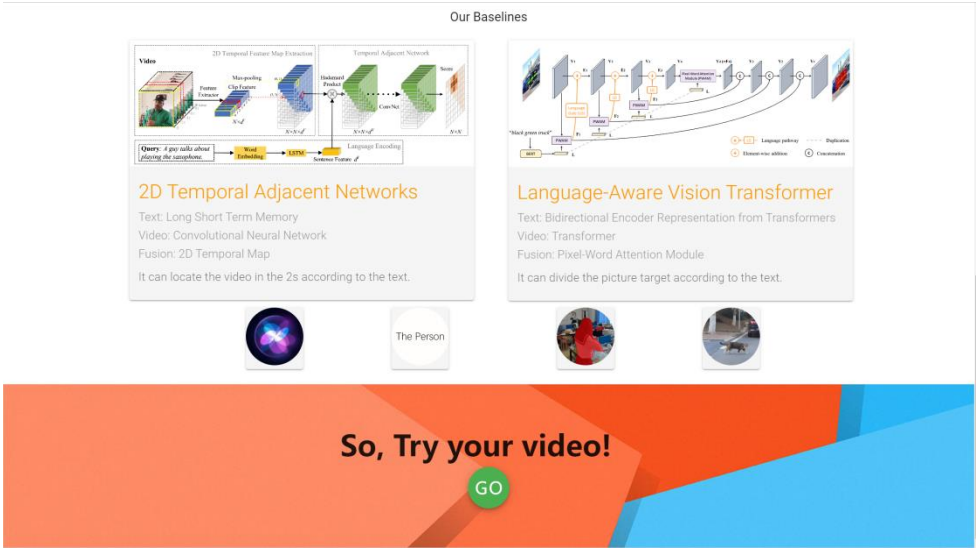
当发生事故时，交警需要及时了解事故发生的时间和位置，因此此软件需要对交通事故的发生进行检测，如：事故发生的时间段，涉事车辆的方位和事故发生全过程，并提供给交警方便后续分析。



三、Fusion Finders 系统架构设计

1.Fusion Finders 系统设计目标和原则

此系统使用 BS 架构，前端利用 Vue 作为核心框架，使用 vuetify、sweetalert 以及 echarts 进行页面设计美化；后端使用 python flask 作为核心框架，并使用 LAVT 和 2D-TAN 模型作为 baseline，设计 Fusion Finders 模型，完成多模态的融合。



2.Fusion Finders 系统的关键功能流程

2.1 语音识别功能

语音识别部分采用的基础模型为 Web Speech，并在此基础上进行优化，添加数据信息和指令信息和可视化语音助手。运行时，当用户输入一段语音后，系统将调用 Web Speech 对此段语音进行解析识别，生成相应文本，随后系统对文本进行分析，若是指令，系统将执行相应操作；若是描述语言，系统将作为输入数据显示。具体的指令情况如下：

```
reset：清空输入内容，重新输入。
location：根据输入内容对视频片段进行定位。
detection：根据输入内容对指代的目标进行检测。
stop：停止语音识别，语音助手退出。
其他：将用户所描述内容识别到输入框。
```

2.2 文本翻译功能

文本翻译功能使用的关键技术为 CaiYun AI，它可以将输入文本翻译为目标语言。与此同时，我们将此功能进行优化，以减少运行时间。我们在进行翻译前将检索此文本是否为使检索结果最优的语言，若是，则取消翻译，否则将其翻译为系统会返回最优结果的语言。

2.3 视频片段定位功能。

此功能的主要目的为接收用户的视频输入和文本输入，并返回可能的视频片段。当系统接收到视频输入和文本输入后，系统将对视频

和文本进行预处理，并将数据以数据流的方式传输到后端服务器上。后端服务器接收到视频和文本后对其进行特征提取及算法计算，并生成 5 段最符合要求的结果返回给前端。前端接收到 5 段结果后，将结果按照置信度进行排序，随后截取每个结果开始端对应的视频帧，连同结果数据一同显示到前端页面。

2.4 指代目标检测功能

此功能的主要目的是对指定的目标进行分割检测，以方便后续的处理。在此过程中，系统首先会将视频当前帧截取出来，随后处理为 RGB 数据，连同用户提问一起传至后端服务器中，后端服务器获得信息后调用 Fusion Finders 模型进行检测，并将检测渲染完成的结果返回给前端。前端解析 RGB 数据后将结果展示到页面中。

四、系统详细设计

1.新手引导模块

新手引导模块的主要功能是辅助用户熟悉整个系统的详细功能，让用户更快地上手操作。此模块分为如下几个部分：

（1）引导者：在此页面中，设计了“引导者”对新手进行使用指引。用户可先于页面左下角上传视频，完成此步骤或点击“引导者”时，即可跳转到下一步。

（2）语音助手：系统内置了语音识别功能和语言翻译功能，前者可将用户的指令转为操作，后者则方便用户使用母语提问。语音助

手将持续接收内容或指令并响应。具体的指令如下：

reset：清空输入内容，重新输入。

location：根据输入内容对视频片段进行定位。

detection：根据输入内容对指代的目标进行检测。

stop：停止语音识别，语音助手退出。

其他：将用户所描述内容识别到输入框。

（3）视频片段定位：在上传视频并进行相关描述后，用户可通过语音助手或右侧“定位片段”按钮进行视频片段定位。视频片段定位的时间取决于云服务器的拥挤程度。当计算完成后，在页面左侧将显示视频中 5 个与描述相符合的片段，点击对应片段即可播放。

（4）指代目标检测：在上传视频并进行相关描述后，用户可通过语音助手或右侧“检测目标”按钮进行视频片段定位。指代目标检测的时间取决于云服务器的拥挤程度。点击检测后，系统将截取当前时刻的视频帧，并在页面右侧显示；检测结束后，系统将当前帧替换为检测分割后的图片，图片支持放大查看。

（5）翻译助手：用户可以使用母语在输入框进行条件描述，系统会将其实时翻译为英文来调用模型，提高所得结果的精确度。

2.监控检索模块

（1）应用场景：通过前期需求分析调查，发现大约有 87%的问卷参与者有过丢失物品后需要调取监控的经历，而对监控的查找往往需要耗费较大的时间和人力。因此，我们推出针对视频监控的检索系统，

系统通过使用监控数据训练后的模型对视频内容进行检索，可以极大降低人工查找的时间成本。

（2）操作方式：为了降低学习成本，此页面结构与新手入门的页面具有相似性，用户可以通过语音或文字输入想要检索的片段，如“一个人进来且拿走了一本书”，系统将对输入的监控视频片段进行检索并返回相关结果。与此同时，用户可以对特定视频帧进行提问，截取出期望目标。

3.远程关怀模块

（1）应用场景：近年来，独居老人的看护问题一直处于社会焦点，绝大部分老年人在独居时都有可能会因自身疾病或意外事故而身处险境。因此，我们模拟如下场景，期冀能够改善此类情况：在特定房间中布置摄像头采集老人的活动数据，当老人遇到危险或困难时，如：摔倒，连接着摄像头的模态探索者就可以检测到这一情况并及时通知家属。这样可以大大缩短滞后时间，为独居老人提供了宝贵的救援时间。

（2）操作方式：为了降低学习成本，此页面结构与新手入门的页面具有相似性，用户可以通过语音或文字输入想要检索的片段，如“一个人摔倒了”，系统将对输入的监控视频片段进行检索并返回相关结果。与此同时，用户也可以对特定视频帧进行提问，截取出期望目标。

4.动物检测模块

（1）应用场景：对生态研究者来说，他们需要获取野生动物或珍稀物种的视频图片，因此需要长时间将相机置于他们可能的出现区域进行录制，而录制之后对视频的检索同样耗费人力。因此，此系统应该实现对一段视频片段中指定动物的检测，并及时返回结果。

（2）操作方式：为了降低学习成本，此页面结构与新手入门的页面具有相似性，用户可以通过语音或文字表达筛选条件，如“一只猫出现”，系统将对上传的监控视频进行检索并返回相应结果。与此同时，用户可以对特定视频帧进行筛选，截取出期望目标。

5.事故分析模块

（1）应用场景：模态探索者可以对道路交通摄像头的数据进行采集和分析。举例如下：当执法者需要调查特定事故的发生时间时，无需人工检索，只需要在页面输入指令，如“两辆车相撞”，系统就将返回相应的结果，并可以根据后续提供的细节视频对场景进行重现。

（2）操作方式：为了降低学习成本，此页面结构与新手入门的页面具有相似性，用户可以通过语音或文字输入想要检索的片段，如“两辆车相撞”，系统将对输入的监控视频片段进行检索并返回相应结果。与此同时，用户可以对特定视频帧进行条件限制，截取出期望目标。

（3）三维重建：在系统首页，通过“事故处理”按钮可以找到相应的“场景还原”按钮，进入此页面后，用户可以上传细节视频和自己的

联系方式，系统将对事故现场进行三维重建，有助于事故纠纷解决和事故原因分析等。由于技术限制，三维重建所需时间较长，所以系统重建完成后会将结果通过邮件的方式发给用户。

五、Fusion Finders 实现与测试

1.Fusion Finders 系统的总体实现

此系统主要通过前后端编程以及模型训练微调实现。

在前后端编程方面，前端主要使用 Intellig IDEA 开发工具，采用 Vue 框架，使用 Vuetify、SweetAlert、Echarts 等插件进行系统构建。后端主要使用 VS Code 进行编程，python flask 框架进行实现。

在模型训练和微调方面，主要使用 2D-TAN 视频片段定位模型和 LAVT 指代目标检测模型作为基础模型，随后进行了模型速度和精度上的优化，使其更适合于指定的部分场景。

2.Fusion Finders 系统的测试

（1）前期准备

数据准备：首先需要准备一些用于测试的数据集。数据集应该包含不同种类的目标物体和不同场景下的图像和视频。可以使用公共数据集，例如 COCO、PASCAL VOC、KITTI 等，也可以使用自己收集的数据集。

测试环境：需要准备一个稳定的测试环境，包括计算机、操作系统、图像和视频输入设备等。测试环境应该能够满足 Fusion Finders

系统的运行要求，并且尽可能排除干扰因素。

测试流程：在测试过程中，需要按照一定的流程执行测试。例如，可以按照以下步骤进行测试：

使用 Fusion Finders 系统对测试数据集中的图像和视频进行目标检测和跟踪；

记录系统的检测和跟踪结果，包括目标的类别、位置、大小、姿态等信息；

对系统的检测和跟踪结果进行评估，包括准确率、召回率、精确度、漏检率等指标；

对系统的性能进行分析，包括运行速度、内存占用、稳定性等方面的指标。

测试结果：最后需要对测试结果进行分析和总结。根据测试结果，可以评估 Fusion Finders 系统的性能和稳定性，并找出系统的优点和不足之处。如果需要改进系统的性能，可以根据测试结果进行相应的调整和优化。

（2）测试流程

单元测试：

对语音识别模块、翻译模块、视频片段定位模块以及指代目标检测模块进行路径测试和覆盖测试，实现了对每一条代码执行路径和代码分支的全覆盖测试。

集成测试：

按照自下而上的测试方法，先测试底层的模块，再集成起来测试

顶层集成模块。在测试底层模块时，设计驱动程序模拟顶层调用，最终集成为 Fusion Finders 系统。

系统测试：

功能测试：系统能够按照规格说明书或需求文档中的要求完成各项功能，并且能够准确地响应用户的输入和操作。

性能测试：系统的性能表现，包括响应速度、吞吐量、并发性和负载能力等方面的指标优秀，可以承载大视频、大数据。

兼容性测试：系统在不同硬件和软件平台上的兼容性，以评估系统的可移植性和可扩展性。系统在 chrome、edge 等多个浏览器上皆可以稳定地运行。

可用性测试：系统的易用性和用户体验优秀，拥有较高的可用性和用户满意度。

六、总结

此系统设计耗时 28 天，其中大约利用 5 天时间对基础模型进行选取及复现，利用 5 天时间设计 Fusion Finders 模型，利用 2 天时间进行系统的整体设计和单元设计，利用 8 天时间进行系统的代码设计及前后端与模型的交互，最后 8 天时间进行系统的优化和测试，努力让系统做到稳定、耐用、简便。

通过此次实训，小组成员学习到了如何利用软件工程完成一项功能完善的系统，并将系统创新性地融合多种模态，实现有效的人机交互。在整个过程中，我们需要互相协作、交流，确保每个人都能发挥

自己的专长和潜力，也体会到了软件开发过程中需求分析的重要性。在项目开始之前，我们需要明确清晰的开发目标和需求规格。只有清楚明确的需求才能指导我们的开发工作，并确保最终的软件产品符合期望。最后，软件项目实训的过程是一个不断学习和提升的过程。在实训中，我们遇到了各种技术难题和挑战，但是通过学习和尝试，我们也不断充实自己的知识和技能，这对我们未来从事软件开发工作非常有帮助。