Proof Of Concept

Mechatronics & Software

Team 3 - EasyRead: Instant Translation with Smart Wearable Device

Ding (Chris) Hao
Yimin (Jane) Pang
Yucheng Yao
Wenyu (Winnie) Yin
Taoming Yu
Xiang (Shawn) Zhang

Date	Developers	Change
Oct 17, 2024	Group 3	First version of the document

Table 1: Revision History

1 Top Challenge Overview

One of the system's key processes is extracting various text segments from captured images. To achieve this objective, two main multifaceted challenges are identified:

1. Ensuring efficient image transfer between the camera hardware and the software translation application.
2. Identifying and correctly extracting various text segments from the images.

Selecting appropriate hardware components for image transfer is critical, requiring a balance between a compact design, adequate image quality, and seamless connectivity. Moreover, the image transfer process must be robust, ensuring minimal latency and enough quality for a lightweight system.

Meanwhile, a simple machine learning model explicitly trained for text extraction is insufficient for accurately detecting and locating separated characters' positions from an image. This inability hinders effective text extraction and translation, as it prevents proper alignment and display of translated text in the original layout, impacting readability and user experience.

Therefore, a tailored solution for the above challenges is required to meet the project's objectives.

2 Challenge Rationale

Challenge 1

In designing an effective wearable device, one of the critical difficulties lies in selecting the appropriate hardware components. These components include a camera, connectivity module, microcontroller and battery; each must be chosen based on specific constraints such as dimensions, weight, functionality, performance and modularity. Wearable devices must be lightweight and compact to ensure user comfort and convenience, which greatly limits the available space for hardware design. For example, the camera should be compact and lightweight while still offering sufficient resolution and image quality for text extraction. Similarly, the connectivity module should support reliable wireless communication between captured images and machine learning models while minimizing power consumption. Balancing these factors makes hardware selection for this system a significant challenge.

Selecting an appropriate image transfer method that enables the camera or microcontroller to push notifications to backend machine learning models is a significant challenge. This challenge arises from ensuring efficient communication while balancing speed, stability and power consumption. Wearables must transmit data fast enough to provide a smooth user experience, but high-speed methods can sometimes compromise reliability. Therefore, finding the right solution involves evaluating communication protocols to achieve the best balance.

Challenge 2

There are a few reasons that performing text extraction and page segmentation at the same time could be challenging for this project.

First of all, OCR and page layout analysis are commonly understood as two separate tasks in most use cases. Therefore, training a machine-learning model that is capable of handling both simultaneously is not a common practice due to the lack of accuracy. However, utilizing more than one machine-learning process leads to an inevitable growth in the system's processing time, reducing its throughput, which eventually lowers customer satisfaction. To achieve maximum system efficiency while maintaining an acceptable prediction accuracy, the team decided that two machine learning models with optimal processing time should be trained or adapted. This decision leads to the next concern that training a machine learning model requires a significant amount of resources, including images and computational power. Given the limited budgets for this project, it would be unrealistic to create a training/testing data set from scratch. Hence, a dedicated solution is required to streamline the machine-learning procedure used in this project.

Another fact that makes the task a challenging one is the broad purpose of image segmentation. In the most common user scenarios, image segmentation is used in object detection and scene understanding, so the available pre-trained models for image segmentation are insufficient for examining the boundary of related text pieces. Instead, more customizations should be incorporated when training the model.

3 Pioneering Solutions

Challenge 1

1. Integrated ESP32-CAM Module (Determined solution)
    - The ESP32-CAM is an all-in-one module that combines a camera, microcontroller, and built-in WiFi capabilities. Its compact and integrated design makes it highly suitable for mounting on a wearable device. The module can capture high-resolution images and transmit them wirelessly to a paired device for processing and translation.
    - Pros
        - Compact and lightweight
            - Minimal additional components make it ideal for wearable applications.
        - Integrated connectivity
            - Built-in Wi-Fi and Bluetooth simplify the design and reduce the need for external modules.
        - Cost-effective
            - Affordable solution with sufficient functionality for text capture.
    - Cons
        - Limited processing power
            - It may require offloading tasks to an external device.
        - Battery life
            - High Wi-Fi usage can drain the battery quickly.
2. ArduCam Mini 5MP with ESP8266 Wi-Fi module
    - This solution has the ArduCam Mini 5MP for higher-resolution image capture, paired with an Arduino Nano for processing and an ESP8266 for Wi-Fi connectivity. The combination allows for efficient image transmission to software for translation.
    - Pros
        - Higher resolution
            - A 5MP camera provides clearer images, enhancing text extraction accuracy.
        - Flexibility
            - Separate modules allow for customization and selective upgrades.
    - Cons
        - Increased complexity
            - Multiple modules require careful integration and wiring
            - Additional components will add to the overall size and weight of the device.
3. OV7670 with HC-05 Bluetooth module and Raspberry Pi zero W

- The low-cost solution utilizes the OV7670 camera module for basic image capture, combined with an Arduino Pro Mini for processing and an HC-05 Bluetooth module for wireless data transfer.
- Pros
    - Lightweight and compact
        - Minimal mass makes it comfortable for extended wear
    - Low power consumption
        - Bluetooth connectivity extends battery life
- Cons
    - Lower resolution
        - VGA resolution may limit translation accuracy for small text
    - Slower transfer speed
        - Bluetooth's limited bandwidth can result in longer transmission times.

Image transfer method

1. Websocket communication
    - Websockets allow bidirectional communication between ESP32 and the web application in real-time.
    - Real-time notifications and efficient communication.
2. Push notifications
    - ESP32 sends a request to a cloud service when an image is taken. Then the cloud service pushes a notification to the web application.
    - Works in the background and is reliable.
3. MQTT (Message Queue Telemetry Transport)
    - ESP32 can send notifications to a central MQTT broker, and the phone can subscribe to the broker to receive the notification.
    - Lightweight, suitable for IoT systems.

Determined solution

The ESP32-CAM module paired with WebSocket communication is chosen because it offers a compact, integrated design with camera and connectivity features. This makes it highly suitable for the wearable device, where space and weight are strictly limited. It can also capture images sufficient for the text extraction process while being lightweight and easy to integrate into the overall system. Compared to other options, the ESP32-CAM is more cost-effective and does not require multiple separate modules for different functionalities. Alternatives such as ArduCam Mini with ESP8266 require more components, which increases the complexity, size and weight of the device. Websocket communication was chosen because it allows efficient, real-time, bidirectional data transfer between the camera and the software. This ensures that data can be transferred promptly without significant delays, which is crucial for a good user

experience. Unlike push notifications or MQTT, WebSocket provides a direct, continuous connection with immediate response and less latency. This combination of ESP32-CAM and WebSocket meets the requirements for speed and reliability while reducing complexity and minimizing the need for additional hardware components, making it ideal for wearable applications.

Challenge 2

To address the challenge regarding text segment identification and isolated extraction, the following technologies are explored:

1. **Tesseract**
    - A text extraction API powered by a pre-trained machine-learning model
    - Pros
        - Light-weight and sufficient for simple and clear documents/images
        - Supports layout analysis to some extent
    - Cons
        - Limited layout handling capability, less flexible, and likely to struggle with complicated text

2. **EasyOCR**
    - Another simple text extraction API implemented with machine learning algorithms
    - Pros
        - Simple to use
        - Slightly better performance
    - Cons
        - Minimal configurability

3. **Layout Parser (requires detectron2)**
    - A machine learning API developed for analyzing text regions in a document
    - Pros
        - Supports a wide variety of document types, including ancient literature, newspapers, and other common textual mediums
        - Utilizes a pre-trained model to perform document layout analysis
    - Cons
        - Limited support for text regions on non-document-based images
        - Requires detectron2, a computer-vision library supported by Mac and Linus operating systems

4. **YOLO from Ultralytics** (self-trained using a dataset in Hugging Face)

- A general-purpose object detection library that can be used, for instance, for segmentation and bounding box prediction
- Pros
    - Optimized for real-time object detection, but faster than the other APIs in general
    - Supports custom model training for various types of object detection
- Cons
    - Extra cost for training and model tuning to keep the model accurate (i.e. reduce the likelihood of false positive/negative).
    - Not initially designed for real-life pictures, which does not 100% fit our user scenarios.

5. **PaddleOCR**
    - Pros
        - Better accuracy in recognizing printed text compared to traditional OCR with deep learning models
        - Supports multiple languages
        - Optimized for speed, with mobile and light-weight versions for portability
    - Cons
        - Requires more computational resources for optimal performance
        - Might not perform well on hand-written text

Determined Solution

After evaluating various technologies for text segment identification and isolated extraction, the determined solution is to utilize YOLO (You Only Look Once) from Ultralytics. This choice stems from YOLO's general-purpose object detection capabilities, which allow for efficient segmentation and bounding box prediction. Its optimization for real-time object detection makes it significantly faster than other APIs, ensuring a responsive user experience. Additionally, YOLO supports custom model training, allowing us to tailor the model to our specific text extraction needs. This flexibility is crucial for handling diverse document types and complex layouts, providing an advantage over more rigid solutions.

With training analysis, it shows approximately ideal results by P-C and P-R curves (Fig. 1). It is evident that the overall accuracy of the model can be balanced under a large variety of recall rates and confidence values while reaching a high degree of precision that is acceptable. As a result, YOLO reduces the adjustment and modification costs. The image processing will easily reach a level with fewer false positives (better precision) and more possibly true positives (less needed recall). Besides, it also shows a high Mean Average Precision (mAP) after 10 epochs of training for demonstration (Fig. 2).

Given these benefits, YOLO stands out as the ideal choice for our project, balancing performance and adaptability while addressing the challenges of accurate text extraction.

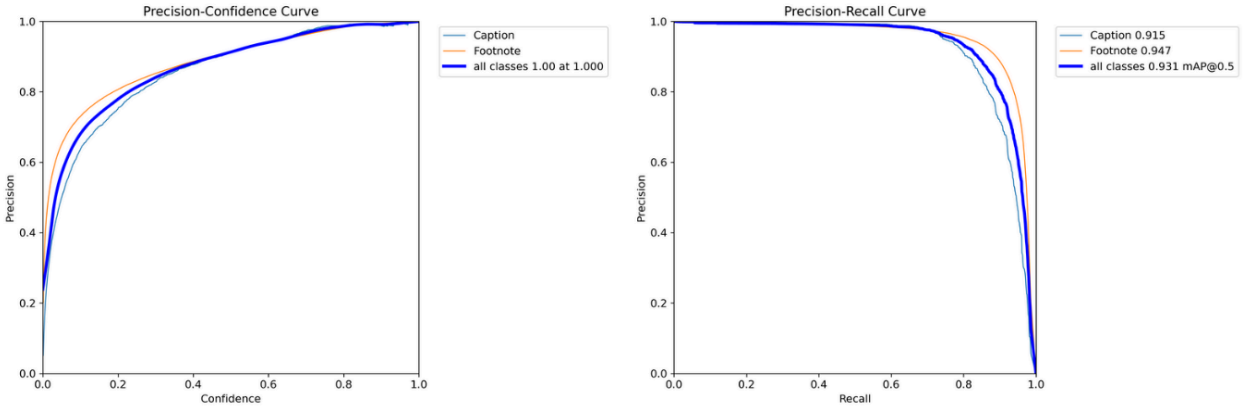

Fig. 1 Precision-Confidence Curve & Precision-Recall Curve

```
val: Scanning F:\Capstone\smart-translation-glasses\proof-of-concept\software-demo\dataset\l
               Class     Images  Instances      Box(P          R      mAP50  mAP50-95): 10
                 all       6489      99816        0.9      0.851      0.931      0.761
             Caption       1067       1763      0.897      0.814      0.915      0.795
            Footnote       6480      98053      0.903      0.888      0.947      0.726
```

Fig. 2 Validation result: Recall rate (R) and Mean Average Precision (mAP)