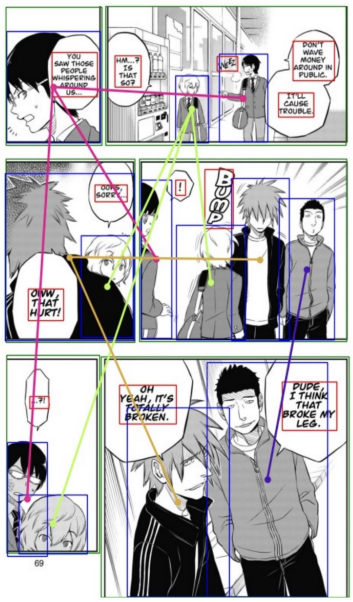
Manga / Comic Parser

Project Proposal, Spring 2026

(a) Panel ordering task



(b) Speech bubble bounding box detection and speaker attribution

(c) Panel edge detection

Introduction

The task of parsing manga and comic books has been a classic problem in computer vision/document parsing since the early 2000s. Given the highly stylized and asymmetric layout of many manga and comic pages, robustly detecting panels and speech bubbles has been a difficult challenge for computer vision systems. Given a manga or comic page, can you extract panel edges and bounding boxes? How can you account for occluded or oddly shaped panels, objects that spill over panel edges, or panels that are entirely contained within other panels? Furthermore, can you detect bounding boxes for characters and speech bubbles?

Methods

Feel free to form your project from the following computer vision tasks (or possibly combine them into a pipeline!)

- **Panel Edge Detection:** Are there any simple edge detection methods you can build on? Quantitatively compare the results of each of your approaches. You may find how Ishii et al. [1] use sobel operators and simple line heuristics as a baseline. A simple extension is the Canny edge detection algorithm [1]—compare your methods on a dataset of your choice (think

[1] Read this wikipedia page for info on Canny edge detection and how it's implemented in this paper [2]. The first two steps involve 1) applying a gaussian filter and 2) finding the intensity gradient of the image—all concepts we have learned in class!

carefully about what evaluation metrics would be relevant for this problem). How can you modify your approaches to be robust to oddly shaped, occluded, or implied panel edges?

- **Text bubble and character bounding boxes:** Predicting bounding boxes is a classic segmentation task for manga—can you take a lightweight, pretrained CNN model and finetune it to predict panel edges, text bubbles, and characters? You may find a class of models called YOLO (You Only Look Once) to be useful for this task: you can read about this model family here [3]. You can read about a transformer-based approach here [4], although be wary of computational feasibility of your project. Can you extend this pretrained model to fulfill some of the limitations mentioned in the paper without retraining?
- **Character attribution, diarization, and panel ordering:** Finally, as a more open ended and challenging direction, can you thing of how you would attribute speech bubbles to speakers? Once you have bounding boxes for panels, speech bubbles, and characters, can you think of any heuristics to match speech to characters? How could you match characters between panels? Finally, what approach could you use to determine the reading order of manga? Once again, pay careful attention to what evaluation metrics would be relevant to this problem, and use a carefully chosen dataset that labels the right kind of data.

Datasets

- Manga109: Contains 109 full manga volumes with over 500,000 bounding boxes for bodies, faces, panels, and text.
- Manga109Dialog: Built on Manga109, this dataset maps over 130,000 explicit speaker-to-text pairs by directly connecting character bounding boxes to text boxes.
- COMICS Dataset: Consisting of over 1.2 million extracted panels from Golden Age American comics paired with automated OCR transcriptions.
- MangaSeg: Contains over 700,000 egmentation masks spanning 10,130 double-sided manga pages. Annotation categories include frames, speech balloons, text/dialogue, character faces, and character bodies.

References

References

[1] D. Ishii and H. Watanabe, "A study on frame position detection of digitized comics images," in Proc. Workshop on Picture Coding and Image Processing, PCSJ20J0lIMPS20I0, Nagoya, Japan, 2010, pp. 124-125.

[2] D Liu, Y Wang, Z Tang, L Li, L Gao, "Automatic comic page image understanding based on edge segment analysis" in Document Recognition and Retrieval XXI, 2014

[3] J Terven, DM Córdova-Esparza, JA Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas", Machine learning and knowledge extraction, 2023

[4] R Sachdeva, A Zisserman, "The manga whisperer: Automatically generating transcriptions for comics" Proceedings of the IEEE/CVF Conference on Computer Vision, 2024