AutoScan: Leveraging Computer Vision to Improve Comic Book Translations and Editing

Kino Leonhart
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, USA
leonhartkino@gmail.com

Lynn Vonderhaar
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, USA
vonderhl@my.erau.edu

Juan Couder
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, USA
ortizcoj@my.erau.edu

Charles Walker
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, USA
walkec45@my.erau.edu

Omar Ochoa
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, USA
ochoao@erau.edu

***Abstract*— As Machine Learning (ML) becomes ever more ubiquitous, it offers promise in automating many previously time-consuming processes. The applications are widespread, including the potential for making literature more accessible. This work focuses on the challenges facing the manga community regarding translation and widespread dissemination of the thousands of available comics. Not only is translating and editing manga time-consuming, but translated copies are often unofficial, so they may be subject to individuality in translation as opposed to an official and standard translation. Although some existing literature addresses automatic translation of manga, it lacks some details that lend additional meaning to the story, e.g., varying fonts and image context. The purpose of this work is to apply computer vision technology to advance the domain of automated manga translation by maintaining variating fonts and leveraging generative Artificial Intelligence (AI) tools to help automatic translations be truer to the original.**

Keywords—machine learning, manga, machine translation, comic book translation, generative artificial intelligence

I. Introduction

The use of Machine Learning (ML) in literature offers a solution to time-consuming translations, including those of manga series. There are tens of thousands of manga series, but due to the high cost of translation, most have not been translated from their original language [1, 2]. This not only limits how many readers can enjoy the untranslated series but also leads to many unofficial translations that may be viewed millions of times [3]. Machine Translation (MT) offers the ability to quickly put official translations on the market.

Existing literature discusses various solutions in making MT a reality including character tracking, translations utilizing scene context, and even fully automated translation pipelines [1, 3, 4]. However, MT still faces many performance challenges in staying true to the original comic because of the need to utilize rich contextual information for a truer translation [2]. Many of these challenges arise from needing to not only translate words, but also incorporate visual context into the translation process, making manga especially difficult to translate.

This work introduces AutoScan, which brings the manga community one step closer to fully automated and accurate MT of series. AutoScan not only offers fully automatic MT by utilizing computer vision technology, but also brings attention to additional visual cues, e.g., font consistency, to make the translations better representations of the original comics. The contributions of this work are as follows:

1. Summarizing the leading contributions to automatic manga translation.
2. Presenting AutoScan, an automatic manga translator based on the compilation of leading contributions.
3. Implementing font consistency using multiple classes for speech bubbles.
4. Exploring generative fill technology for improving the cleaning process before translation.
5. Identifying current challenges and future directions for automatic manga translation.

II. Background

Before describing the approach, it is critical to first explain object detection and Optical Character Recognition (OCR), two critical components of AutoScan.

A. Object Detection

Object detection is an ML classification task that aims to locate specified classes within an image. Older methods use a Convolutional Neural Network (CNN) with a sliding window to locate the class instances within an image, but this can be relatively time-consuming [5]. By contrast, the You Only Look Once (YOLO) models are CNNs that divide the image into a grid and any cell that includes the center of an object must detect that object with a confidence score [5]. This means, as the name implies, that the YOLO model only needs to scan the

image once to locate the desired classes within it, making it much faster than other popular methods. This process is shown in Figure 1.

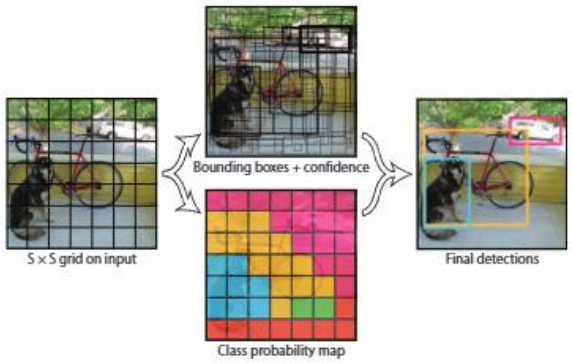

Figure 1. The YOLO model's method [5].

B. Optical Character Recognition

OCR is a method by which an ML model extracts text from an image [6, 7]. There are many potential applications of OCRs including converting the text in scanned documents to a digital format and, in the case of this work, finding text within a manga page for translation. OCRs may be trained to pick up different alphabets and languages, or even handwritten text [6, 7]. OCR is accomplished in four main steps:

1. Pre-processing: During this phase, the OCR reduces noise within the image. This phase is occasionally placed after the other three phases.
2. Segmentation: This is the process of detecting text lines within the rest of the image.
3. Feature extraction: This phase combines text lines into feature vectors for classification.
4. Classification: The OCR classifies the feature vectors into recognizable characters within an alphabet [6, 7].

III. Existing Solutions

The use of ML to aid manga translation has garnered much attention in the research community with contributions spanning various challenges facing automatic translation. Because of the complexity of the task, many authors focus on improving one aspect of the pipeline at a time. For instance, some authors focus on character and speech recognition [8, 9, 10, 11, 12]. Al-Ibrahim, et al. provide a method for extracting text from non-flat manga where bubbles may overlap frames and cause confusion for ML models [8]. Rigaud, et al. show a method for speaker association between bubbles and characters [9]. Dutta, et al. use segmentation to locate speech bubbles as opposed to considering this a classification problem because of the varied shapes of speech bubbles in manga [12]. Other authors, e.g., Saeed Sharif, et al., improve the translations by using a post processing grammar corrector [13]. Meanwhile Kovanen and Aizawa improve the ordering of extracted text bubbles [14].

Some authors do provide approaches to automatic manga translation [1, 3, 15]. Hinami, et al. propose a context-aware pipeline that translates comics utilizing the context of the frame and other bubbles to make a more accurate translation [1]. Sachdeva and Zisserman provide a pipeline for extracting text and contributing it to characters for automatically writing manga transcripts for the visually impaired [15]. Meanwhile, Kiano, et al. performs AT utilizing more context to improve the accuracy of the translated comic [3]. The authors accomplish this in two ways. The first is by using a longer contextual window using previous scenes to improve the translation. The second approach is by utilizing bibliographic information including the title of the comic and the name of the comic's author. Lippmann, et al. also use context to improve the

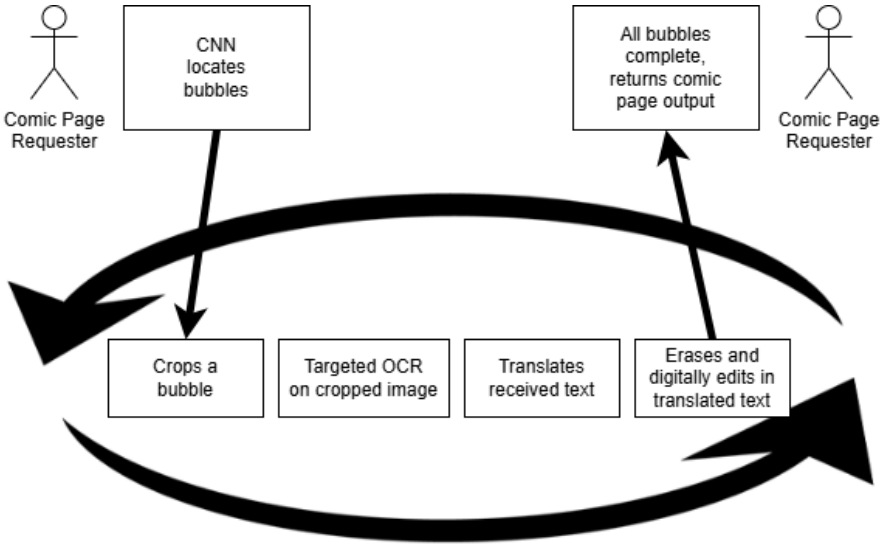

Figure 2. The AutoScan pipeline.

accuracy of the translation but do so with a multimodal Large Language Model (LLM) [2].

All of this existing work helps with the process of accurate automatic translation of manga. However, none have maintained the varying fonts within the translated version, thereby losing some vital meaning.

IV. Approach

MT for manga can be broken down into four main steps, which are shown visually in Figure 2:

1. locating text within the frames,
2. extracting text from the frames,
3. translating the text, and
4. editing the frames to include the new translations.

In this work, locating the text within the frames utilizes the YOLOv8 model by Ultralytics to classify the speech bubbles [16]. For this application, 1062 images of manga pages were annotated using Roboflow to adapt the model to this application via transfer learning [17]. There are several types of speech bubbles which often have different fonts to reflect the sentiment of the scene. The following types of speech bubbles were tracked as different classes as part of font maintenance and some examples are shown in Figure 3:

- Regular speech bubbles,
- shout bubbles,
- narrative bubbles,
- happy bubbles,
- evil bubbles,
- thought bubbles, and
- scared bubbles.

After training, the model achieved a 0.724 mean Average Precision (mAP) at Intersection over Union (IoU) 0.50. This means that the bubble location is accepted when the predicted box overlaps with the ground truth box with an IoU of 0.50 or greater. The bubble location step also saved the (x, y) coordinates for each corner of the bounding box within the page, which are used in a later step for cropping and editing the translated text. The bubble is then cropped from the image and saved for the OCR step.

Manga OCR is then used to extract the text from the regions classified as bubbles by the YOLOv8 model [18]. Several OCRs were tried, but Manga OCR had the best success rate. Table I shows the full list of OCRs attempted. In this step, Manga OCR extracted the text from each cropped bubble and put it in a script file. It was found that constraining the sections of text was the best way to get a clear OCR extraction as extraneous input could lead to other sections of the image being picked up as random text. The tested OCRs are all multi-lingual to improve accuracy on manga, specifically they had to pick up Japanese and English for this application.

As was mentioned, AutoScan uses Manga OCR, which is a custom model trained for manga based on Microsoft's TrOCR [18, 19]. Manga OCR is only trained for Japanese, but it handles Japanese well, including text with Furigana and text on top of images [18]. However, it is important to note that experiments with Manga OCR do show some difficulty in handling

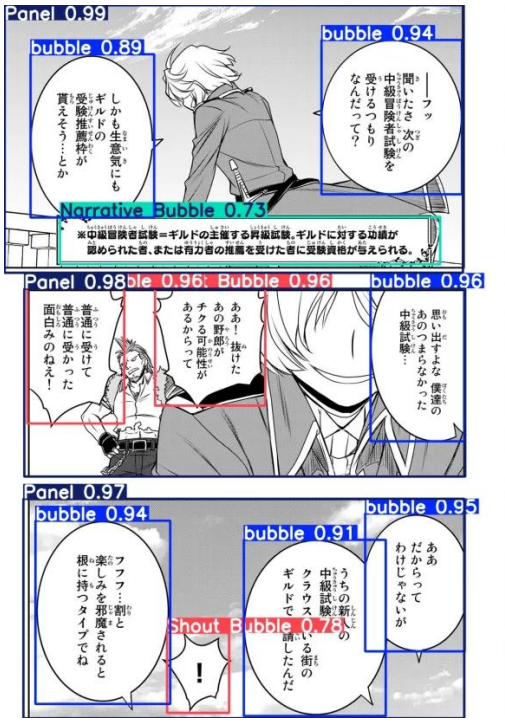

Figure 3. YOLO results [30].

TABLE I. FULL LIST OF OCRS

OCR	Reference	Performance Notes
Tesseract	[27]	• Available in many languages • Gave some basic results, but not good enough results
EasyOCR	[28]	• Supports over 80 languages • Not good results
Google Cloud Vision (GCV)	[29]	• Improvement on other previously tried OCRs • Tracked the speech well and received precise data on the location of the text itself • Collected word-level bounding polygons inside of each cropped speech bubble • Grouping the polygons enables building a mask that covers the text, which enables precise cleaning for a translated version of the text • Did not work quite as well as Manga OCR
Manga OCR	[18]	• Works well with Japanese characters, including Furigana • Trained specifically on manga • Only works with Japanese manga

Furigana-heavy text, which are small characters next to the main text that offer pronunciation guidance. In some instances, it understands the Furigana as normal speech, as opposed to the pronunciation guide that it is. Additionally, Manga OCR does not track text metrics, e.g., location of the text, which is helpful for masking and editing the text in the final editing step. GCV does this, but its performance is worse than Manga OCR. Therefore, adding this function to Manga OCR in the future would be beneficial.

Step three is the translation step. There are many available translators including Google Translate and DeepL [20, 21]. AutoScan uses DeepL because of its performance [21]. Currently, AutoScan takes each cropped bubble, runs Manga OCR on it, translates the characters to English, and then sends them to the editing step. However, a more accurate translation would come from a complete script that incorporates context into the translation. This will be discussed further in the discussion and future work sections.

The final step is the editing step. Editing uses digital technology to target and erase areas in which the previous text is present. This involves removing the Japanese characters and replacing them with the translated characters. AutoScan uses Adobe's generative fill tool for the editing step, as is shown in Figure 4. This is helpful because it does not ignore the background but rather extends it to make the new image more natural [22]. This process is currently manual, but future iterations of AutoScan will automate this step using models from Google's Vertex AI [23]. It is important to note that although Manga OCR is more accurate than GCV for this application, GCV also outputs a text mask, which would offer an ideal way to target areas for erasing. The current pipeline sometimes leaves some pixels of Japanese characters after the cleaning, so utilizing a mask may help with the cleaning process.

The translated text from step three is then inputted into the newly cleaned area in the image with the bounding box being the maximum size for the text to fit. By utilizing the various types of speech bubbles, the pipeline is able to input the English text in font types that match the original text, thereby maintaining additional context from the original comic. An example of a fully translated page with two types of fonts is shown in Figure 5.

V. DISCUSSION

The four-step process described in the approach section is a conglomerate of the state-of-the-art in MT for manga. This work then builds on the state-of-the-art to add font consistency and generative filling outside of speech bubbles.

However, there are still challenges and areas for future improvement. For instance, there is some existing literature on ordering speech bubbles so that scenes and longer scripts can be translated together [3, 1, 14]. This helps to improve the

Figure 4. Generative fill for text outside of a speech bubble [22, 30].

Figure 5. Fully translated page with two types of fonts [30].

transcription as opposed to translating a single speech bubble at a time without any context. Currently, AutoScan translates bubble-by-bubble, but incorporating methods for bubble ordering would be beneficial. This would likely include additional steps, e.g., frame tracking.

As Figure 4 shows, text may occur outside of speech bubbles. In this case, Adobe's generative fill tool was used to continue the background before overlaying the new, translated text. With the rise of generative Artificial Intelligence (AI), e.g., diffusion models, this step could easily be automated in the future.

VI. Challenges and Future Directions

AutoScan offers a promising direction for comic book transcription. However, there are some clear directions for future work. Sachdeva and Zisserman discuss tracking characters [15]. However, this is mostly regarding automatically generating a transcription for the purpose of aiding people with visual impairments in enjoying comics. Vachmanus et al. analyze the use of a model to track faces in manga, stating that the model can be adapted for manga emotion analysis [24]. Adding character tracking to AutoScan would enable the building of character personas, potentially with the help of LLMs, which can be used to generate fan-based stories about the characters. The use of LLMs for the creation of fanfiction and other fan-based stories is well documented already [25, 26]. Combining character tracking and emotional analysis would help in generating better translations because it adds further context about the speaker, the subject, and the storyline.

VII. Conclusion

The purpose of this work was to add to the existing literature on the subject of automatic translation and editing for manga series. Existing literature spans many aspects of this area, but this work consolidates the state-of-the-art in the area into a new pipeline called AutoScan. AutoScan follows the four main steps of automated MT and editing for comics and brings attention to font consistency as important visual context in producing a high quality comic translation. This work also discusses tools for generative filling of background when cleaning a page for translation. Finally, this work discusses future work in character tracking to build character personas and improve the overall quality of the translations.

References

[1] R. Hinami, S. Ishiwatari, K. Yasuda and Y. Matsui, "Towards Fully Automated Manga Translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtual, 2021.

[2] P. Lippmann, K. Skublicki, J. Tanner, S. Ishiwatari and J. Yang, "Context-Informed Machine Translation of Manga using Multimodal Large Language Models," *arXiv,* 2024.

[3] H. Kaino, S. Sugihara, T. Kajiwara, T. Ninomiya, J. B. Tanner and S. Ishiwatari, "Utilizing Longer Context than Speech Bubbles in Automated Manga Translation," in *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 2024.

[4] Z. Zhang, Z. Wang and W. Hu, "Unsupervised Manga Character Re-identification via Face-body and Spatial-temporal Associated Clustering," *arXiv,* 2022.

[5] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[6] K. Abdulwahhab Hamad and M. Kaya, "A Detailed Analysis of Optical Character Recognition Technology," *International Journal of Applied Mathematics, Electronics, and Computers,* vol. 4, no. Special Issue, pp. 244-249, 2016.

[7] R. Mithe, S. Indalkar and N. Divekar, "Optical Character Recognition`," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 2, no. 1, pp. 72-75, 2013.

[8] N. Al-Ibrahim, M. Al-Ibrahim and W. Al-Awadhi, "Text Extraction and Recognition in Manga Comics Using Image Processing Techniques," in *Science and Information Conference (SAI)*, London, UK, 2024.

[9] C. Rigaud, N. Le Thanh, J. C. Burie, J. M. Ogier, M. Iwata and E. Imazu, "Speech balloon and speaker association for comics and manga understanding," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2025.

[10] X. Qin, Y. Zhou, Y. Li, S. Wang, Y. Wang and Z. Tang, "Progressive deep feature learning for manga character recognition via unlabeled training data," in *ACM TURC '19: Proceedings of the ACM Turing Celebration Conference - China*, Chengdu, China, 2019.

[11] K. Arai and H. Tolle, "Method for Real Time Text Extraction of Digital Manga Comic," *International Journal of Image Processing (IJIP),* vol. 4, no. 6, pp. 669-676, 2011.

[12] A. Dutta, S. Biswas and A. Kumar Das, "CNN-based segmentation of speech balloons and narrative text boxes from comic book page images," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 24, pp. 49-62, 2021.

[13] M. Saeed Sharif, B. Auwal Romo, H. Maltby and A. Al-Bayatti, "An Effective Hybrid Approach Based on Machine Learning Techniques for Auto-Translation: Japanese to English," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Zallaq, Bahrain, 2021.

[14] S. Kovanen and K. Aizawa, "A layered method for determining manga text bubble reading order," in *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, Canada, 2015.

[15] R. Sachdeva and A. Zisserman, "The Manga Whisperer: Automatically Generating Transcriptions for Comics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024.

[16] "Explore Ultralytics YOLOv8," Ultralytics, 2025. [Online]. Available: https://docs.ultralytics.com/models/yolov8/. [Accessed 2025].

[17] "Everything you need to build and deploy computer vision applications.," Roboflow, 2025. [Online]. Available: https://roboflow.com/. [Accessed 2025].

[18] kaiserinn, "Manga OCR," GitHub, 2025. [Online]. Available: https://github.com/kha-white/manga-ocr. [Accessed 2025].

[19] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li and F. Wei, "TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington D.C., USA, 2023.

[20] "Google Translate," Google, 2025. [Online]. Available: https://translate.google.com/?sl=es&tl=en&op=translate. [Accessed 2025].

[21] "Translate Text," DeepL, 2025. [Online]. Available: https://www.deepl.com/en/translator. [Accessed 2025].

[22] "Next-level Generative Fill. Now in Photoshop.," Adobe, 2025. [Online]. Available: https://www.adobe.com/products/photoshop/generative-fill.html. [Accessed 2025].

[23] "Imagen 2 for Generation and Editing," Google, 2025. [Online]. Available: https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/imagegeneration?_gl=1*1nn5iu5*_up*MQ..&gclid=CjwKCAiAwqHIBhAEEiwAx9cTeeuJmp0C23C4lfY-wBV2JEPeucLAxKiijy4v1Tvq7jP8CXCK7NGEoxoCD-sQAvD_BwE&gclsrc=aw.ds. [Accessed 2025].

[24] A. Kusmiatun, D. Efendi, M. J. Al Pansori, L. Judijanto, Y. Sariasih and K. Saddhono, "The Power of AI in Generating Novel Based and Impactful Character Development for Fiction Story," in *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2024.

[25] S. Vachmanus, N. Phinklao, N. Phongsarnariyakul, T. Plongcharoen, S. Hotta and S. Tuarob, "Automating Manga Character Analysis: A Robust Deep Vision-Transformer Approach to Facial Landmark Detection," *IEEE Access,* vol. 12, pp. 131284 - 131295, 2024.

[26] D. Stanko, "Integrating Artificial Intelligence in Naruto Fan Fiction Writing: A Case Study," *Arab World English Journal (AWEJ) Special Issue on ChatGPT,* 2024.

[27] 0xflotus and stweil, "Tesseract OCR," GitHub, 2025. [Online]. Available: https://github.com/tesseract-ocr/tesseract. [Accessed 2025].

[28] rkcosmos, "EasyOCR," GitHub, 2024. [Online]. Available: https://github.com/JaidedAI/EasyOCR. [Accessed 2025].

[29] "Cloud Vision API Documentation," Google, 2025. [Online]. Available: https://cloud.google.com/vision/docs. [Accessed 2025].

[30] LA-Gun and R. Nakajima, Dame Skill “Jidō Kinō” ga Kakusei Shimashita: Are, Girūdo no Sukauto no Minasan, Ore wo “Irai” tte Iimasendeshita?, 2021.