Retrieving and Analyzing Translations of American Newspaper Comics with Visual Evidence

Jacob Murel[1] and David A. Smith[1]

Northeastern University, Boston, MA, USA
j.murel@northeastern.edu
dasmith@ccs.neu.edu

Abstract. Research on image classification and text translation for comics have transpired largely independent of one another. Machine translation tools focus on comics' text features, thereby largely ignoring comics' heavily visual dimension. Image classification applications for comics focus primarily on genre and artist attribution. This paper bridges the gap between these areas by investigating image classification model accuracy for identifying translations of American newspaper comic strips. How might machine learning algorithms leverage comics' distinguishing visual features in order to identify pre-existing translations? To what extent do textual differences affect classification model accuracy in identifying otherwise identical comics? Using a dataset of ~~1~~8,000 English and Spanish comics, we generate embeddings from three CNNs and a Vision Transformer. We generate additional embeddings from binarized images and images with text redacted using an OCR model. We compute the cosine distance between given pairs of comics and evaluate its accuracy at retrieving translations. The best models rank the true translation first for 97% of queries, falling to 94% when the language is not known.

Keywords: Information Retrieval · Translation · CNNs · OCR

1 Introduction

Machine translation tools for comics utilize optical character recognition (OCR) and other text-based approaches. These have shown promising results. Many textual features in comics are simultaneously visual however. One notable example is onomatopoeia. In many comics, the size, shape, and color of words—i.e., the visual components of text—convey important semantic information. Machine translation for comics, then, may need to integrate image-based approaches for translation. Our paper makes initial steps in this direction.

The interplay of text and image is a distinguishing feature of comics. In attempting to define their craft, cartoonists frequently highlight comics' visual features as the medium's defining quality [17,8,19]. Additionally, comics historians criticize tendencies to focus on comics' textual-linguistic elements and neglect comics' formal visual qualities, wherein they claim much of a given comics'

meaning resides [18]. Coinciding with this, machine learning research notes the need for integrating text and image-based analyses and has begun to make steps in this direction [1,20].

With the larger aim of developing a more holistic approach to comics translation, this paper explores image classification methods for identifying translations of comics images. Specifically, we employ pre-trained image classification models to generate embeddings for several thousand comic strips and their pre-existing translations. We then measure machine performance in identifying translation pairs based solely on visual properties. We institute several modifications to examine how text affects purely image-based retrieval accuracy. In this way, we here lay groundwork for a larger project that analyzes how comics' visual and textual features interact in translation.

2 Related Work

Research on text translation and image classification for comics have transpired largely independent of one another. After reviewing machine translation for comics (§2.1) and the use of image classification for entire strips (§2.2), we also note prior work on identifying text within panels (§2.3), as our experiments explore text's effect on translation retrieval performance.

2.1 Machine Translation

Open-access machine translation tools for comics—such comic-translate[1], ImageTrans[2], and manga-image-translator[3]—all operate using OCR engines to extract text and generate translations thereof. The language learning platform Manga Vocabulometer uses Google Translate to source translations of manga text [13]. Current comics translation tools thus use exclusively textual information to generate translations.

One recent study explores context-aware, multimodal translation for manga text. This uses Faster R-CNN and illustrate2vec in order to group comics panels into scenes and extract visual semantic information to generate more robust text translations. In this way, the researchers propose to develop a fully automated approach to comics translation [11]. This is the only study of which we are aware that incorporates comics' visual features in generating translations.

2.2 Image Classification

A number of studies explore approaches to comics classification using convolutional neural networks (CNNs). Many of these studies use pre-existing labels to categorize manga pages among discrete genres, after which they train CNNs

[1] `https://github.com/ogkalu2/comic-translate`
[2] `https://www.basiccat.org/imagetrans/`
[3] `https://github.com/zyddnys/manga-image-translator`

using this labeled data for image-based genre classification tasks [3,26,31]. User-labeled data for CNNs have also been used for attribution tasks. Essentially, artists are considered categories, to which CNNs attribute a given comics image according to visual features. Experiments for author classification show promise on page-level but not panel-level manga data. Author and series attribution have also been addressed in digital humanities research that explores computational methods for distinguishing artistic styles [16].

2.3 Text Segmentation

Text segmentation poses an acute problem for comics. One reason is the heavily stylized nature of comics typography, which can impede OCR models not familiar with the wide, seemingly limitless, array of text forms that appear in comics. Region detection is another issue, as OCR models often interleave transcriptions from separate speech bubbles.

An early unsupervised approach to manga text detection groups text according to the distance between detected text characters [15]. A later approach utilizes visual features (e.g. color, shape, topology, etc.) identify connected text blocks from which to produce independent transcriptions [25,24]. While this latter approach shows a marginally decreased performance in terms of F1-score (approx. 2%), it is tested on a larger dataset of Japanese *manga* and French *bandes dessinées*. Pixel-level text detection is another approach, albeit one with less success[4]. Research shows CNNs perform well for detecting speech balloons and other visual text markers in Bangla comics [7].

3 Method

We investigate methods based on image classification for identifying links between comics and their translations. As we shall see, translations of comics and even reprints in the same language, may change not only the text but also the visual layout.

Comics are a primarily visual medium. In light of this, text segmentation and translation experiments have asked how comics' visual components may improve text extraction. For instance, Hinami et al. use character detection and other image detection methods to improve text translations [11]. Text segmentation methods also leverage comics visual components (e.g., balloon boundaries, color, etc.) to improve text extraction. We examine the reverse: how do textual features affect image-based comics retrieval tasks? More specifically, we ask: to what extent do textual differences affect a model's performance in identifying otherwise identical comics?

Image classification applications for comics focus primarily on genre and artist attribution. Our paper breaks new ground by investigating the effect of text variants and extraction on image classification accuracy for American newspaper comic strips. In this way, we explore comics translation and classification through the lens of information retrieval. Our research thereby examines the interplay of

text and image in comics, an issue which has been a primary concern of comics scholars [17,9,14,22].

3.1 Datasets

We compile a dataset of 19,526 images of American newspaper comic strips. These images are divided among five series-based classes: *Calvin & Hobbes*, *Luann*, *Nancy*, *Peanuts*, and *Garfield*. The set contains colorized and binary strips in landscape and portrait orientations according to how the strip was initially published. The set contains daily and Sunday strips. In each class, half of the the images are English-language comics with the other half being their Spanish-language translations. We compile metadata to readily identify comics image files as linguistic pairs.

We generate a parallel dataset to examine how text extraction affects comics translation retrieval. Using the Tesseract OCR engine[4], we automatically identify and remove text with whitespace from the entire corpus of comics images.

Additionally, we create a third and final dataset to examine how more strictly visual features—specifically, color—affect comics translation retrieval. We use the Pillow library[5] to create a binarized version of the original comics dataset.

Test Data We use a set of 500 comics images for evaluating our models. This set consists of 100 images from each of the *Calvin & Hobbes*, *Luann*, *Nancy*, *Peanuts*, and *Garfield* classes (500 images total). It contains both daily strips and Sunday strips. All 500 test images are unique English-language comics.

We use corresponding versions of the same 500 strips for evaluating models and embeddings across each dataset. This means that, whichever 500 comics we use for evaluating the original comics embeddings, we use those same comics' text-stripped versions for evaluating the redacted embeddings, and their binarized versions for evaluating the binarized embeddings.

3.2 Models & Embeddings

We use four pre-trained neural networks: ResNet50 [10], MobileNet-v2 [27], EfficientNet-B0 [29], and Vision Transformer (ViT) [5]. The former three are all CNNs that have been widely tested in image classification research, although, to our knowledge, ResNet is the only one of these to have been used in comics image classification research [31,3]. We include ViT for two reasons: 1) a number of recent experiments in computer vision compare vision transformers and CNNs with mixed results [32,23,2]; and 2) to our knowledge, no research investigates ViT's applications with regard to comics classification and retrieval.

We generate twelve sets of image embeddings. Specifically, we create three sets of embeddings for each of the pre-trained image classification models: one from the dataset of original comics, one from the dataset redacted, and one from the binarized dataset.

[4] `https://github.com/tesseract-ocr/tesseract`
[5] `https://pypi.org/project/pillow/`

3.3 Evaluation

Our primary concern is how well each model identifies a given comic's corresponding translation compared to other comics within and outside the same series and language. In other words, we ask: if provided a comic strip from a given linguistic domain, how often will the model return the strip's corresponding language variant rather than an unrelated comic?

To answer this question, we evaluate model retrieval performance using our test set of 500 comics. For each image in these two test sets, we compile the ten-closest image-vectors in a given model's embeddings. Upon compiling the ten-closest vectors for a given English-language test comic, we then determine for how many comics in both of our test sets the model identified the corresponding Spanish-variant as the closest vector.

We evaluate for two different comics retrieval scenarios. First, we evaluate for situations in which each comic's language is known beforehand. In this approach, when compiling the ten-closest image-vectors to a provided English-language comic, we consider only those comics from a different language. That is, since all of the comics in our test set are English-language comics, we only consider Spanish-language comics when retrieving the closest image-vectors. We also evaluate for situations in which the language of each comic is unknown beforehand. If we provide a given English-language comic, how well can the model survey all comics regardless of language and identify its corresponding Spanish-language variant? In this way, we evaluate model performance for different information retrieval scenarios.

Note that we never test for class-wise restrictions. That is, in both unrestricted and language-restricted search tasks, we evaluate model's performance in identifying a corresponding comic strip out of given every comic across all available classes.

We calculate retrieval accuracy for each test comic by using cosine similarity scores to determine recall at rank one and at rank ten.

Cosine similarity We use cosine similarity for comparing two image-vectors in a given model's embeddings. For two embedding vectors $\mathbf{x}$ and $\mathbf{y}$ of images of comic strips, cosine similarity is computed as:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2} \sqrt{\sum_{i=1}^{n} \mathbf{y}_i^2}} \tag{1}$$

Cosine similarity is the cosine of the angle (i.e., distance) between two points in a vector space, which we use to signify the model's judged degree of similarity between two comics strips. The higher two strips' cosine value, the more visually similar the model determines the strips to be.

Recall The standard recall formula used in classification tasks is:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

TP signifies the total number of true positives in a model output, and FN is the total number of false negatives. In our task, a true positive is the corresponding Spanish-language comic for a given English-language comic. Everything else is a true negative.

We are interested in whether a model identifies corresponding translations as visually similar. We therefore calculate recall at one and recall at ten for each image in our text set.

Recall at one (R@1) is a modification of the standard recall metric for information retrieval that asks: how often is the first returned item a true positive? Using cosine similarity scores to determine the Spanish-language comic closest to a test comic, we calculate how often the closest Spanish-language comic is a true positive, i.e. the true corresponding translation of a given English-language comic. We also calculate recall at ten (R@10). This measures how often a true positive is among the ten highest ranked items in terms of similarity.

4 Results

Tables 1 and 2 show class-wise recall at 1 and 10 for a search task where the language of each strip is known. When given a query strip in English, therefore, the retrieval system will only consider candidate strips in Spanish. Tables 3 and 4 show class-wise recall at 1 and 10 for an search task not restricted by language. All tables show recall on three different versions of the comics: the original image, comics with redacted text, and binarized images. All tables display model recall for each comics series and average recall across all five series.

Table 1: Class-wise R@1 for language-restricted search task

	Calvin & Hobbes	*Luann*	*Nancy*	*Peanuts*	*Garfield*	**Total**
ResNet50						
original comics	1	.99	.12	.73	.98	.76
redacted	.99	.94	.16	.65	.97	.74
binarized	.97	.96	.82	.50	.94	.84
MobileNet-v2						
original comics	1	1	.94	.92	.98	.97
redacted	1	1	.88	.93	.97	.96
binarized	1	1	.99	.85	.97	.96
EfficientNet-b0						
original comics	1	1	.98	.87	.97	.96
redacted	1	1	.99	.78	.97	.95
binarized	1	1	.99	.88	.97	.97
ViT						
original comics	1	1	.11	.43	.97	.70
redacted	1	1	.11	.53	.97	.72
binarized	1	1	.98	.17	.97	.82

Table 2: Class-wise R@10 for language-restricted search task

	Calvin & Hobbes	*Luann*	*Nancy*	*Peanuts*	*Garfield*	**Total**
ResNet50						
original comics	1	1	.19	.90	.98	.81
redacted	1	.97	.16	.84	.98	.79
binarized	1	.98	.86	.75	.96	.91
MobileNet-v2						
original comics	1	1	.97	.98	.98	.99
redacted	1	1	.96	.97	.98	.98
binarized	1	1	.99	.92	.97	.98
EfficientNet-b0						
original comics	1	1	.99	.91	.97	.97
redacted	1	1	.99	.89	.98	.97
binarized	1	1	.99	.93	.97	.98
ViT						
original comics	1	1	.12	.61	.97	.74
redacted	1	1	.11	.71	.98	.76
binarized	1	1	.98	.36	.98	.86

MobileNet and EfficientNet markedly outperform ResNet and ViT on all three input conditions. Each model consistently performs well with the *Calvin & Hobbes*, *Luann*, and *Garfield* classes. By comparison, models evidence more variable performance with respect to both the *Nancy* and *Peanuts* classes.

Comparing Tables1 and 2 with Tables 3 and 4 shows that all models perform better in the language-restricted search task rather than the unrestricted search task. If the model's search is limited to Spanish-language comics in the former, it has a more limited set of data with which to compare a given English-language comic. While R@10 is by definition higher than R@1, the gap is not very large in any of the comics series or models evaluated.

5 Discussion

The four tables reveal notable differences between model architectures in comics translations retrieval tasks. MobileNet and EfficientNet markedly outperform ResNet and ViT. One potential explanation for this may be their respective design purposes. Both MobileNet and EfficientNet are designed to work with limited data and computing constraints, which may have a regularizing effect on their representations of input images.

As previously mentioned, all four tables reveal class-wise differences in model performance. For instance, all models across all three input conditions reveal high accuracy in identifying translation pairs for *Calvin & Hobbes*, *Luann*, and *Garfield*.

Our modifications to comics' visual features (i.e., text redaction and binarization) have a variable effect on model performance. ResNet and ViT model

Table 3: Class-wise R@1 for unrestricted search task

	Calvin & Hobbes	*Luann*	*Nancy*	*Peanuts*	*Garfield*	**Total**
ResNet50						
original comics	.98	.99	.10	.54	.98	.72
redacted	.98	.95	.10	.46	.97	.69
binarized	.97	.96	.80	.21	.94	.78
MobileNet-v2						
original comics	1	1	.82	.78	.97	.91
redacted	1	1	.60	.67	.97	.85
binarized	1	1	.99	.57	.97	.91
EfficientNet-b0						
original comics	1	1	.96	.76	.97	.94
redacted	1	1	.96	.64	.97	.91
binarized	1	1	.99	.67	.97	.93
ViT						
original comics	1	1	.11	.15	.97	.65
redacted	1	1	.12	.20	.97	.66
binarized	1	1	.97	.01	.94	.78

Table 4: Class-wise R@10 for unrestricted search task

	Calvin & Hobbes	*Luann*	*Nancy*	*Peanuts*	*Garfield*	**Total**
ResNet50						
original comics	1	.99	.11	.80	.98	.78
redacted	1	.95	.10	.67	.98	.74
binarized	1	.98	.83	.52	.94	.85
MobileNet-v2						
original comics	1	1	.95	.89	.98	.96
redacted	1	1	.88	.88	.98	.95
binarized	1	1	.99	.83	.97	.96
EfficientNet-b0						
original comics	1	1	.99	.85	.97	.96
redacted	1	1	.99	.80	.97	.95
binarized	1	1	.99	.78	.97	.95
ViT						
original comics	1	1	.11	.87	.97	.79
redacted	1	1	.11	.44	.97	.70
binarized	1	1	.98	.04	.97	.80

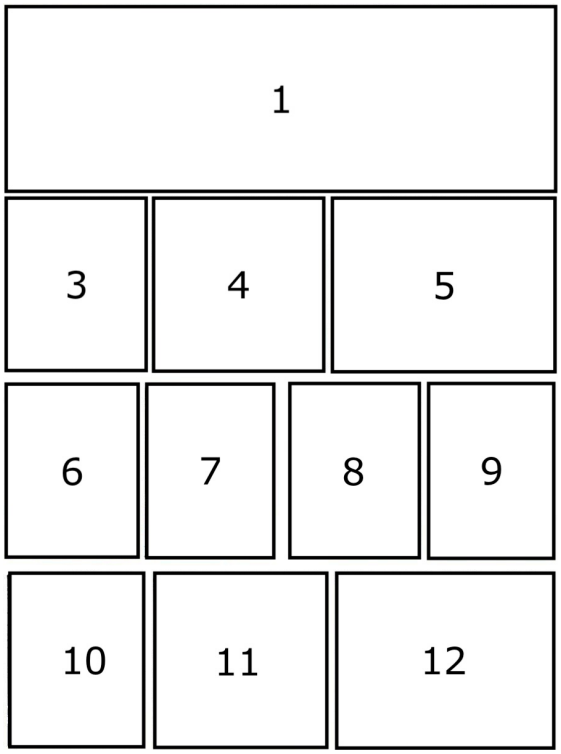
performance improves on *Nancy* comics once we binarize the dataset. This is undoubtedly due to that fact that, unlike other series, language variants of *Nancy* comics are colored differently. That is, English-language daily *Nancy* strips are colorized while their Spanish-language counterparts are black-and-white. ResNet, MobileNet, and ViT recall for *Peanuts* comics decreases with binarized comics however.

Text removal does not significantly affect model performance for any class. Nevertheless, model performance marginally decreases on the redacted dataset. We believe this slight decrease in performance may be attributed to the introduction of new visual features between language variants introduced by the text removal process. The Tesseract OCR engine does not identify all of the text in each comic, and so a select few words are left behind at times in our redacted dataset. But which lines and portions of text remain differ between languages. The Tesseract OCR engine does not fully strip corresponding regions and lines of text from different language variants. Thus, while Tesseract removes large portions of text, it nevertheless leaves different remnants of text. In this way, our text removal process essentially introduces new visual differences between linguistic variants of a single comic strip. We suspect these increased visual differences account for the marginal decrease in model performance between the original comics and redacted datasets. With access to more accurate text-detection models for comics, however, it might be possible to achieve more meaningful results with text redaction.

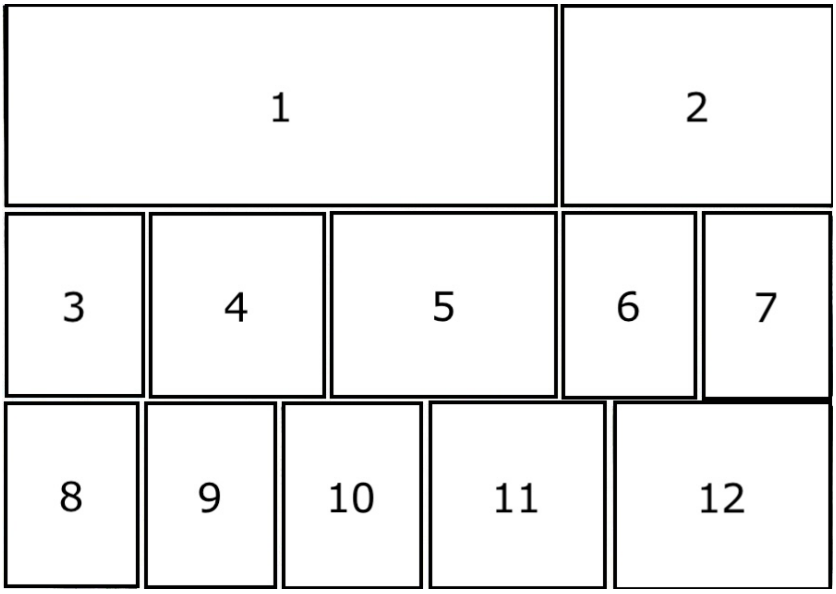
Differences in panel layouts between English and Spanish variants inconsistently affect model performance. Figure 1 shows the panel layouts for a given *Peanuts* Sunday strip in both its English and Spanish-language versions. The panels are numbered according to their corresponding content. The content of corresponding panels are the same; the only difference between these copies (aside from their language) is the panel arrangement The English variant is in portrait while the Spanish variant is in landscape. The landscape arrangement of panels also includes an additional panel (panel 2) not included in our English copy (revealing how some of our comics have been modified from their original syndication appearance). All of our models, however, correctly identify these two *Peanuts* strips as a translation pair.

Panel re-arrangement seems to negatively affect other cases, however. Across all four tables, model recall never goes about .99 for the *Nancy* class and .98 for the *Garfield* class. Indeed, every model consistently misidentifies the same one *Nancy* strip and two *Garfield* strips. Both are Sunday strips in which panel arrangements differ between the English and Spanish-language variants, and an additional title-panel is added to the top of the Spanish-language version.

In fact, panel layout provides some explanation for why the *Calvin & Hobbes* and *Luann* classes achieve perfect recall scores. Unlike many other American newspaper comics strips, *Luann* Sunday strips are a single row of panels. Thus, as with every comics' daily strips, *Luann* Sunday strip panels are not rearranged. *Calvin & Hobbes* Sunday strips are likewise unmodified, albeit for different reasons. Bill Watterson, the creator of *Calvin & Hobbes*, explicitly designed many



(a) Layout of English-language *Peanuts* Sunday strip



(b) Layout of Spanish-language *Peanuts* Sunday strip

Fig. 1: Comparison of panel layouts for a pair of English and Spanish-language *Peanuts* Sunday strips

Sunday strips so that the panels cannot be rearranged by avoiding a traditional grid format [30]. Moreover, as revealed through correspondence with Andrews McMeel while attempting to acquire rights for republishing comics in this paper, Watterson has stipulated that republications of *Calvin & Hobbes* cannot alter their original created appearance. Unlike other newspaper comics strips, then, *Luann* and *Calvin & Hobbes* are less open to panel rearrangement, perhaps accounting for universally high model performance with these classes.

6 Conclusion

In this paper, we explore an image classification method for identifying translation pairs of American newspaper comic strips. We use four popular image classification models to generate embeddings of comics images. We use series/author as image classes. We use recall at one and ten to evaluate for different search tasks. Our experiments show that pre-trained image classification models can successfully identify translations of comics images, with some variable performance across classes.

Visual differences between linguistic variants seems to the central factor affecting model performance. Thus, binarizing comics images appears to improve model performance overall, while removing the comics text with a standard OCR engine hinders model performance. Panel rearrangements between translations remains a key factor negatively affecting models.

7 Future Work

One area for future work is developing a method for identifying comics translations despite rearranged panel layouts. One potential method for addressing panel rearrangements is to search for translation pairs at the panel level, or to re-rank the top k candidates at the strip level by using panel-level alignments. Panel segmentation is an ongoing and key research area in machine learning applications for comics [21,28,6,12]. Given the architectural differences between linguistic variants of comics, a comics retrieval pipeline could segment panels within a given strip and match one strip's panels with those of another. The projected similarity between strips—and so the probability that they are translations of one another—would then be a function of the degree of similarity between one strip's individual panels and those of another strip. One could also consider the interaction between text translation and visual content. Do some translators move text among panels to balance the layout? Or do they modify each panel's appearance to accommodate a panel-by-panel translation? What can be left unspecified in the text translation because of visual context? These are only a few avenues for further research in comics translation opened up by the methods presented here.

References

1. Augereau, O., Iwata, M., Kise, K.: A survey of comics research in computer science. Journal of Imaging **4**(7) (2018)
2. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? In: Advances in Neural Information Processing Systems. vol. 34, pp. 26831–26843 (2021)
3. Daiku, Y., Iwata, M., Augereau, O., Kise, K.: Comics story representation system based on genre. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 257–262 (2018)
4. Del Gobbo, J., Herrera, R.M.: Unconstrained text detection in manga: A new dataset and baseline. In: Proceedings of the 16th European Conference on Conputer Vision (ECCV) Workshops. pp. 629–646 (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
6. Dutta, A., Biswas, S.: Cnn based extraction of panels/characters from bengali comic book page images. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 1, pp. 38–43 (2019)
7. Dutta, A., Biswas, S., Kumar Das, A.: Cnn-based segmentation of speech balloons and narrative text boxes from comic book page images. International Journal on Document Analysis and Recognition **24**, 49–62 (2021)
8. Eisner, W.: Comics and Sequential Art. Poorhouse Press (1985)
9. Groensteen, T.: The System of Comics. University Press of Mississippi (2009)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
11. Hinami, R., Ishiwatari, S., Yasuda, K., Matsui, Y.: Towards fully automated manga translation. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence. vol. 35, pp. 12998–13008 (2021)
12. Ho, A.K.N., Burie, J.C., Ogier, J.M.: Panel and speech balloon extraction from comic books. In: 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 424–428 (2012)
13. Kato, J., Iwata, M., Kise, K.: Manga vocabulometer, a new support system for extensive reading with japanese manga translated into english. In: Proceedings for the 25th International Conference on Pattern Recognition, Workshops and Challenges. pp. 223–235 (2021)
14. Lambeens, T., Pint, K.: The interaction of image and text in modern comics. In: Texts, Transmissions, Receptions: Modern Approaches to Narratives, p. 240–257. Brill (2015)
15. Li, L., Wang, Y., Tang, Z., Lu, X., Gao, L.: Unsupervised speech text localization in comic images. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1190–1194 (2013)
16. Manovich, L.: How to compare one million images? In: Berry, D. (ed.) Understanding Digital Humanities, p. 249–278. Palgrave Macmillan (2012)
17. McCloud, S.: Understanding Comics: The Invisible Art. Kitchen Sink Press (1993)
18. Miodrag, H.: Comics and Language: Reimagining Critical Discourse on the Form. University Press of Mississippi (2013)
19. Mitchell, W., Spiegelman, A.: Public conversation: What the

20. Nguyen, N.V., Rigaud, C., Burie, J.C.: Digital comics image indexing based on deep learning. Journal of Imaging **4**(7) (2018)
21. Nguyen Nhu, V., Rigaud, C., Burie, J.C.: What do we expect from comic panel extraction? In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 1, pp. 44–49 (2019)
22. Postema, B.: Narrative Structure in Comics: Making Sense of Fragments. RIT Press (2013)
23. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: In: Advances in Neural Information Processing Systems. vol. 34, pp. 12116–12128 (2021)
24. Rigaud, C., Burie, J.C., Ogier, J.M.: Segmentation-free speech text recognition for comic books. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 03, pp. 29–34 (2017)
25. Rigaud, C., Nguyen, N.V., Burie, J.C.: Text block segmentation in comic speech bubbles. In: Proceedings for the 25th International Conference on Pattern Recognition, Workshops and Challenges. pp. 250–261 (2021)
26. Rishu, Kukreja, V., Sharma, V.: Automated classification of comics into genres using cnn-svm model: A study on visual storytelling. In: Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS 2023). pp. 122–127 (2023)
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4510–4520 (2018)
28. Stommel, M., Merhej, L., Müller, M.: Segmentation-free detection of comic panels. In: Proceedings of the International Conference on Computer Vision and Graphics (ICCVG). pp. 633–640 (2012)
29. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114 (2019)
30. Watterson, B.: Calvin and Hobbes: Sunday Pages 1985-1995. Andrews McMeel (2001)
31. Xu, C., Xu, X., Zhao, N., Cai, W., Zhang, H., Li, C., Liu, X.: Panel-page-aware comic genre understanding. IEEE Transactions on Image Processing **32**, 2636–2648 (2023)
32. Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Liu, X., Liu, Z.: Delving deep into the generalization of vision transformers under distribution shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7277–7286 (2022)