

Supplementary Material for COMICS: More Details of Post-Processing and Dataset Creation

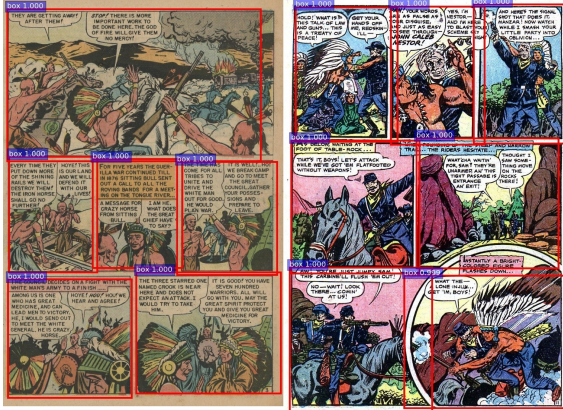


Figure 1. Two examples of panel-segmentation. On the left, we have a typical case where the panels are perfect rectangles, which results in perfect panel segmentation. Whereas on the right, panel layouts are atypical (the bottom-right panel is circular, and dialog-boxes spill over from one panel to another). Therefore, panel and subsequent textbox segmentation could possibly contain errors.

1. Panel detection using Faster R-CNN

Panels were detected from raw comic book page scans using Faster R-CNN 1. With regards to how we select the number of training examples used for training the R-CNNs, we observe that the shapes of panels (and textboxes) in our dataset follow a long-tailed distribution, with the majority of them being fairly standard (i.e., rectangular). Thus, only 500 manual annotations were sufficient to obtain robust panel detection. We ran an experiment on a small held out set of 20 random pages with 124 panels, and found the mean intersection-over-union overlap between ground-truth panel boxes and RCNN predicted boxes to be 0.911 (1.0 is perfect overlap).

2. OCR Post-Processing and Advertisement Removal

OCR makes systematic mistakes on our textboxes. We target two types of these mistakes using PyEnchant:¹ 1) where the OCR system fails to recognize the first letter of a

¹<http://pythonhosted.org/pyenchant/faq.html>

particular word (e.g., *eleportation* instead of *teleportation*), and 2) where the OCR system transcribes part of a word as a single alphabetical character. To eliminate errors of the first type, we start by tokenizing the OCR output using NLTK’s Punkt Tokenizer.² We then sort the vocabulary of the tokenized OCR output in decreasing order of frequency and pick words ranked from 10,001 to 100,000, because most misspelled words are also rare. For each of these words that is length three or longer, we look up the most likely suggestion offered by PyEnchant. If the only difference between the most likely suggestion and the original word is an additional letter in the first position of the suggestion, then we replace the word with the suggestion everywhere in our corpus. To correct the second type of errors, we simply delete all single character alphabetical tokens that are not one of ‘a’, ‘d’, ‘i’, ‘m’, ‘s’, ‘t’ - characters which can plausibly occur by themselves quite frequently (some occur after an apostrophe).

In addition to spelling errors, the books in COMICS contain many advertisements that we need to remove before generating data for our tasks. While most dialogue and narration boxes contain less than 30 words, longer textboxes frequently come from full-page product advertisements (e.g., Figure 2). However, detecting ads from page images is not easy. Some ads are deceptively similar to comic pages, containing images and even containing faux mini-comics. Aside from ads, there are also other undesirable pages; many books contain text-only short stories in addition to comics. We remove these kinds of pages using features from OCR transcriptions. We annotate each page of 100 random books with a label indicating the presence or absence of an invalid page as our training set and each page of 20 random books as our test set. Out of 6,117 annotated pages, 697 of them are either advertisements or text-only stories (11.4%). We train a binary classifier using Vowpal Wabbit:³ which takes the OCR text for all the panels of a pages as lexical features (unigrams and bigrams). We improve our model by adding features like total count of words in the page and a count of non-alphanumeric char-

²<http://www.nltk.org/>

³https://github.com/JohnLangford/vowpal_wabbit/wiki

acters. Our model gives us a total misclassification error of 8% and a false negative error of 17.3%, which means it misses one invalid page out of every six. The model has a negligible false positive error of 0.2%. Using this model to filter the entire dataset of 198,657 pages yields 13,200 invalid pages.

3. Examples from Dataset Creation

OCR transcription is the final stage of our data creation pipeline (panel extraction \rightarrow textbox extraction \rightarrow OCR). Therefore, faulty outputs in any of the preceding steps can lead to faulty OCR outputs. In Figure 3, there are only minor errors in OCR extraction due to understandable misinterpretations of the text in the dialog boxes. For example, the OCR interprets the letters “IC” as “K”, which leads to incorrectly predicting the word “QUICKLY” as “QUKKLY”. However, in Figure 4, we observe a more critical error due to missing pixels in the panel extraction process. Due to the layout of the textbox in the panel, crucial portions of the text are trimmed from view; while the OCR does a valiant job of predicting the contents of the textbox, its output is gibberish.

NOW FLY LIKE A BIRD

**With Wings Made From
The Original Sketch of
Leonardo Da Vinci's Flying Wings!**

Now any adventure loving boy can build Da Vinci's flying wings with just ordinary carpenter's tools.

OFFERED FOR THE FIRST TIME

People said it couldn't be done but Leonardo went right ahead and built the wings and then carted them to a nearby hill and took off. What happened is excitingly told in **THE BIRDMAN**, The Story of Leonardo Da Vinci. See the actual original sketch Leonardo used to build his flying wings with just ordinary tools.

EXTRA SPECIAL TREAT

Also in **THE BIRDMAN**: The diagram of the parachute which Leonardo invented. Yes, you too can make a parachute out of cloth and string by just following Leonardo's drawing.

EXCITING — THRILLING

Whether you build the flying wings, the parachute or other of Leonardo's inventions, one thing is sure, you will enjoy the exciting and thrilling story, **THE BIRDMAN**, which is illustrated in color with the kind of pictures you like to look at. You don't have to buy **THE BIRDMAN** which is only 98c because you can send for it for a 10-day trial and if you don't get a real kick out of **THE BIRDMAN** the cost will be nothing.



SEND NO MONEY...Try 10 Days

MAIL COUPON NOW

STRAVON PUBLISHERS Dept. B-9212
113 West 57th St., New York 19, N. Y.

I want to try **THE BIRDMAN** 10-days. I will deposit with postman only 98c plus postage. After trying 10-days I may return **THE BIRDMAN** for a full refund of the purchase price.

NAME _____

ADDRESS _____

CITY _____

ZONE _____

STATE _____

☐ Check if you enclose 98c. Stravon pays postage. Same refund.

HOW TO HYPNOTIZE

**IT'S EASY TO
HYPNOTIZE...**
when you know how!



Mail Coupon Today

STRAVON PUBLISHERS, Dept. H-9212
113 West 57th St., N. Y. 19, N. Y.

Send **HOW TO HYPNOTIZE** in plain wrapper.

☐ Send C.O.D. I will pay postman \$1.98 plus postage.

☐ I enclose \$1.98. Send postpaid.

If not delighted, I may return it in 10 days and get my money back.

Name _____

Address _____

City _____

Zone _____

State _____

Canada & Foreign—\$2.50 with order

Want the thrill of imposing your will over someone? Of making someone do exactly what you order? Try hypnotism! This amazing technique gives full personal satisfaction. You'll find it entertaining and gratifying. **HOW TO HYPNOTIZE** shows all you need to know. It is put so simply, anyone can follow it. And there are 24 revealing photographs for your guidance.

SEND NO MONEY

FREE ten days' examination of this system is offered to you if you send the coupon today. We will ship you our copy by return mail, in plain wrapper. If not delighted with results, return it in 10 days and your money will be refunded. Stravon Publishers, 113 West 57th St., New York 19, N. Y.

DEATH VALLEY, DEC., 1953, Vol. 1, No. 2. Published bi-monthly by ALLEN HARDY ASSOCIATES, INC., 500 Fifth Ave., New York 36, New York. Subscription rates: 12 issues \$1.50 in U. S. Possessions and Canada. Foreign: \$2.00 International Money Order, U. S. Funds. Application for entry as second-class matter pending at the Post Office in New York, N. Y. Additional entry at Syracuse, N. Y. Copyright 1953 by ALLEN HARDY ASSOCIATES, INC. No similarity between any of the names, characters, persons, or institutions appearing in this magazine with those of any living or dead person or institution is intended and any such similarity which may exist is purely coincidental. Advertising representative: Leonard Greene and Associates, 45 West 45th Street, New York City, New York. Printed in U. S. A.

Figure 2. An advertisement from the dataset. The juxtaposition of text and image causes it to slightly resemble a comics page.

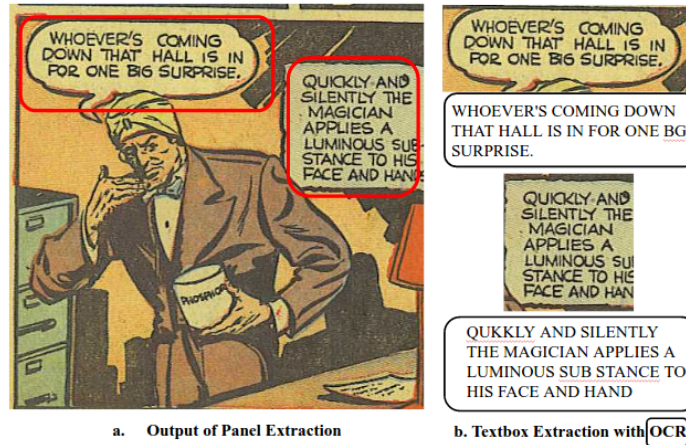


Figure 3. A minor OCR error. Mistakes such as predicting “BG” for “BIG” are understandable, since the ‘I’ in “BIG” is barely visible. Similarly, the “IC” in “QUICKLY” looks a lot like “K” in this font. Finally, “SUB STANCE” is predicted rather than “SUBSTANCE”, due to an end-of-line word break.

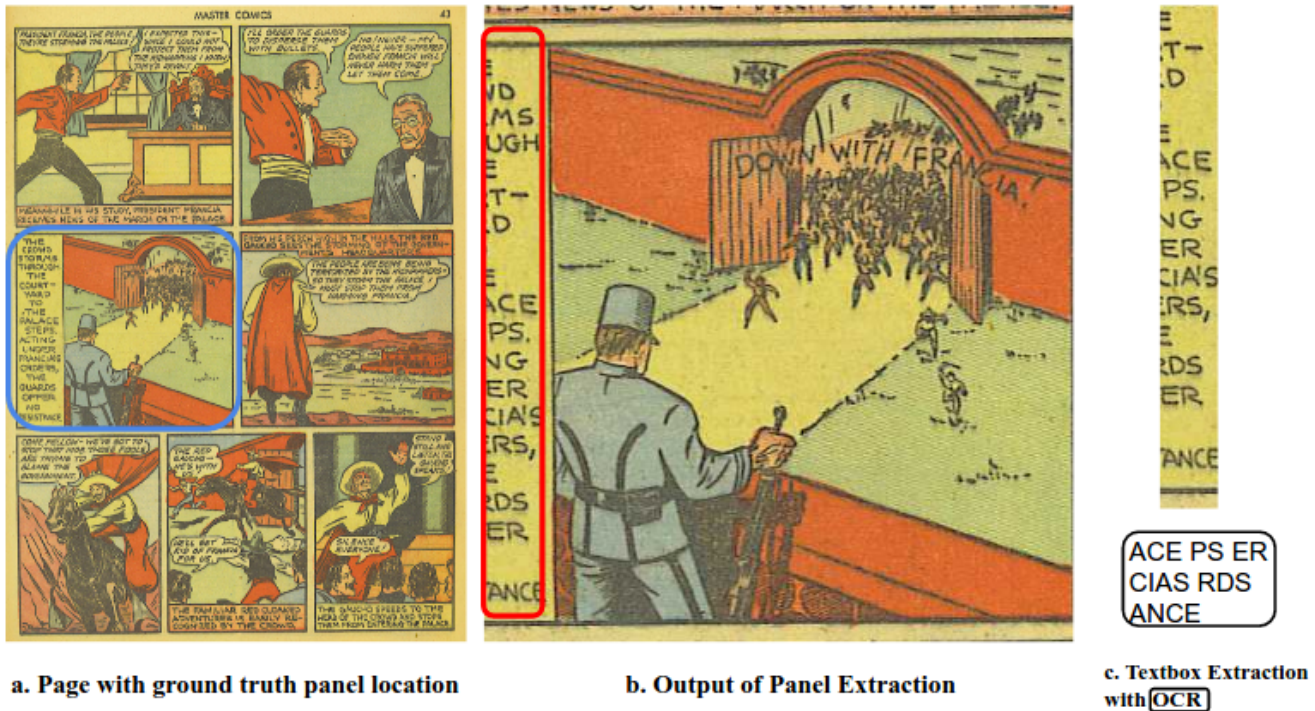


Figure 4. A major OCR error. In part a) of the figure, note the location of the panel in the page. b) gives us the panel as predicted by the RCNN, but a critical portion of the text is missing. As a consequence, the textbox extraction is also faulty, rendering the OCR completely meaningless.