A Novel Application of Optical Character Recognition for Product Image Compliance

Abhinav Chanda, Maharshi Dutta, Samir Husain, Tirthankar Mukhopadhyay, Matthew A. Lanham
Purdue University, Krannert School of Management

KRANNERT SCHOOL OF MANAGEMENT
PURDUE UNIVERSITY

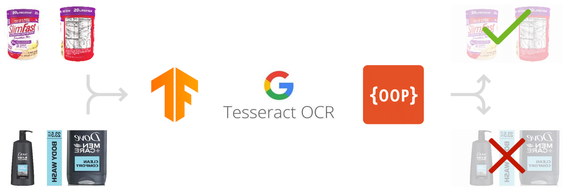
Abstract

Product images on a digital platform have multiple legal/internal compliances that need to be satisfied. Our work is an attempt at automating the audit process. Our solution aims at cutting down on the manual effort for a major US retailer while saving potential losses due to lawsuits. The project is implemented using Python. The images are scraped from the digital platform. Next, object detection techniques crop backgrounds, and a custom OCR algorithm extracts text. Finally, a scalable business rule framework validates the text. The solution can be extended to any industry facing a similar challenge.

Introduction

The Americans with Disabilities Act (ADA) states that product images on a digital platform must clearly show warnings, nutritional information and supplement facts.

Violations to these rules incur hefty fines for the organization per listing. Hence, retail firms must ensure products are compliant based on the aforementioned checks.

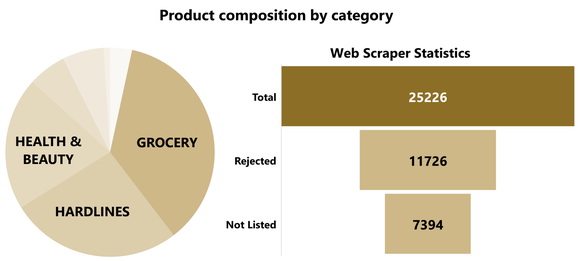


Our solution automates the compliance testing process, by sourcing images, using machine learning algorithms to extract text, and finally passing a verdict on each listing.

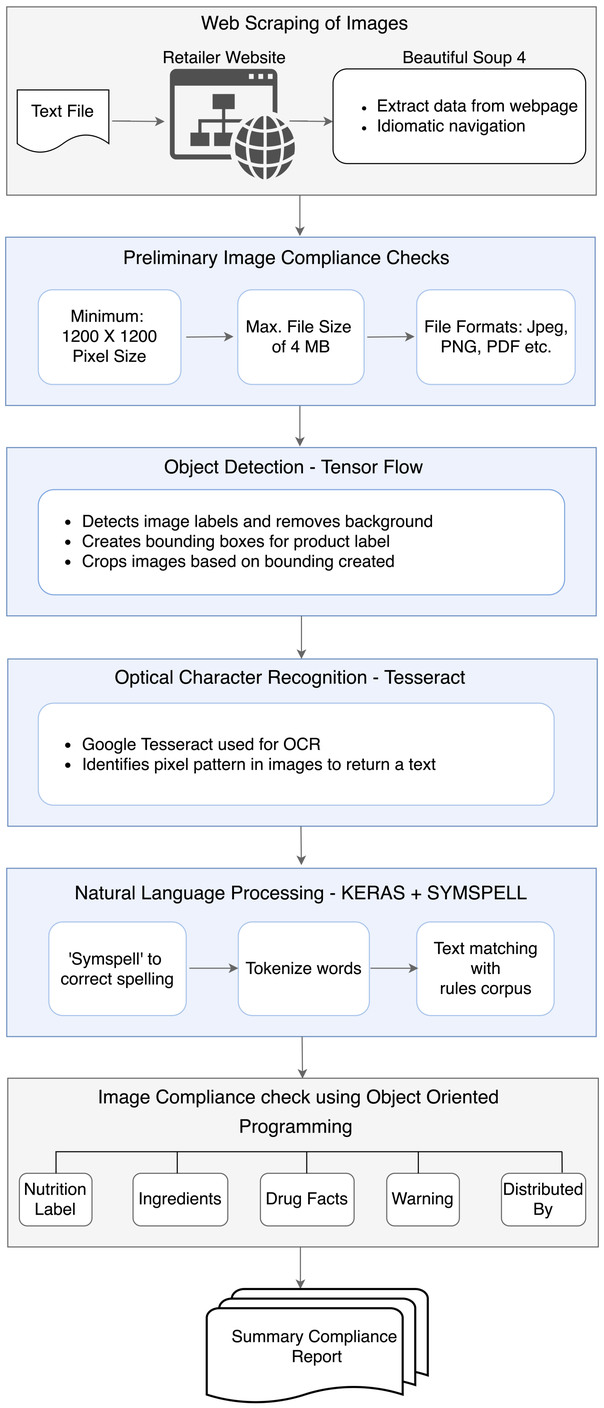
Data Summary

The digital platform of US retailer currently has approx. 360K images. Each of these images are broadly divided into 8 categories like Health and Beauty, Grocery etc.

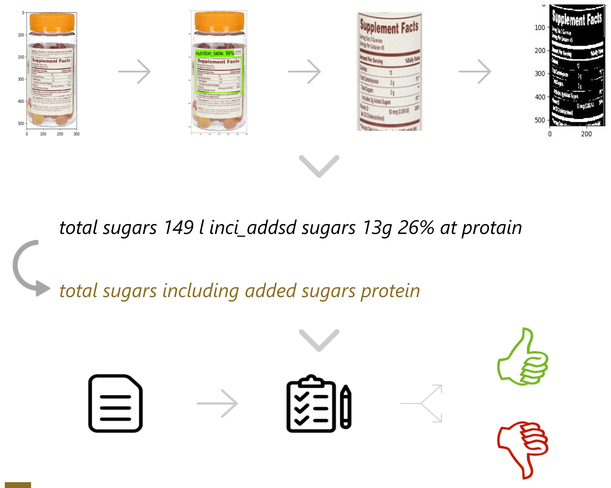
All images data are scraped checking for:
- Image listing on the website
- Front Facing image availability



Methodology

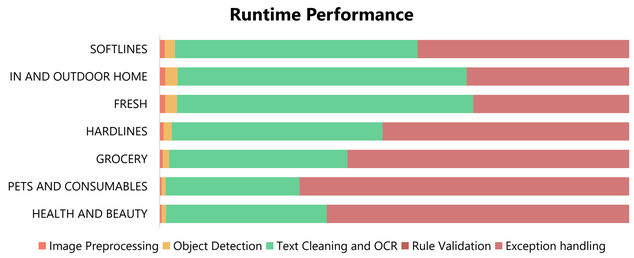


Product Journey

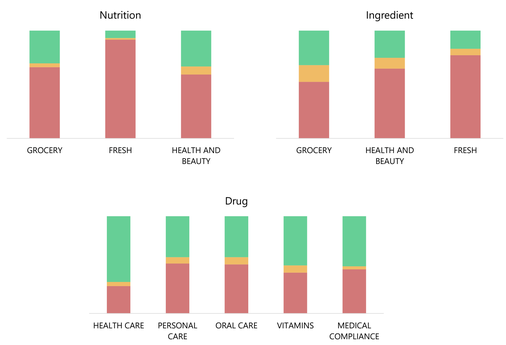


Results

In order to test the performance of our proposed solution, we generated a dataset containing 25000 listings. The category wise breakdown for performance at each step is shown below.



Each category (nutrition, ingredients, etc.), products are classified as Red, Amber and Green.



The above visuals show the composition of compliant, unsure, non-compliant images based on presence of Nutrition facts, Ingredients and Drug facts.

Performance Metrics

We manually flagged 3,000 images for the rules as specified. This was compared against the flags from the algorithm. Following are the results.

Nutrition facts	+	-
+	218	106
-	34	324

91%

Ingredients	+	-
+	260	60
-	147	369

72%

Drug facts	+	-
+	265	62
-	63	201

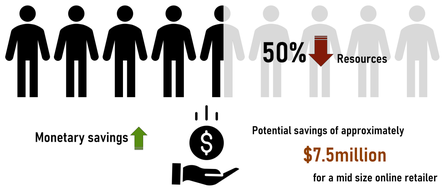
76%

SPECIFICITY

Conclusion

Image compliance testing for digital product listings requires considerable manual effort. Large retail organizations maintain hundreds of thousands of products on their websites, and hence, the effort can lead to several man hours.

Our solution automates the manual process of tagging non-compliant images. Time saved with this can go up to thousands of hours every audit cycle.



ADA non compliance can attract heavy penalties on a per case basis and thus the overall value of this project can be placed between $7.5 Mn to $10 Mn for a large retail organization.

The solution is designed such that new compliance requirements and corpus can be added with no code change.

Future Scope

- GUI to upload images and corpus, change rules
- Automate emails for non-compliant listing
- Extend OCR model to cater other languages
- Accommodate video feeds in warehouses

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.