

Finding a Needle in a Haystack: Computer vision and machine learning techniques for extracting comics from Belgian Illustrated Periodicals in the Interwar Period

For submission to: DHBenelux

Presentation category: Short papers

Authors:

- Julie M Birkholz (UGent/KBR): <Julie.Birkholz@kbr.be>;
- Benoît Crucifix (KULeuven/KBR) : <benoit.crucifix@kuleuven.be>;
- Erwin Dejasse (ULB/KBR): <Erwin.Dejasse@ulb.be>;
- Sébastien Hermans (KBR): <Sebastien.Hermans@kbr.be>;
- Krishna Kumar Thirukokaranam Chandrasekar (UGent): <krishnakumar.tc@ugent.be>;
- Bas Vercruysse (UGent) : <Bas.Vercruysse@UGent.be>

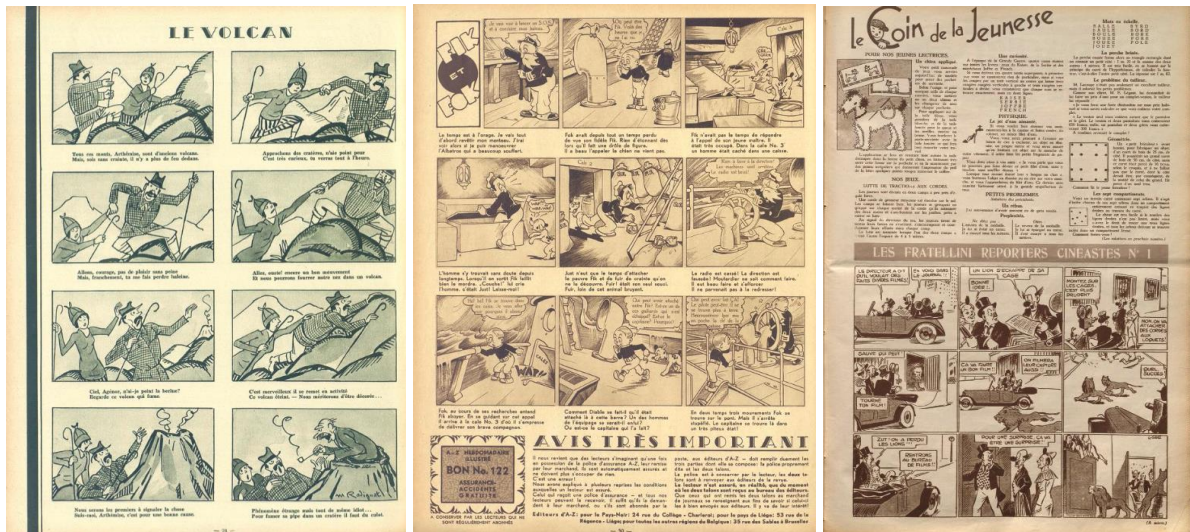
Abstract

Comics in the popular press provide illustrative markers of the state of a specific society, denoting notable events and people through often wordless commentary on life. The publication and circulation of comics in general-audience magazines remains a largely overlooked avenue of research for the humanities for literary history, comics and periodical studies. This is largely due to the fact that comics as parts of periodicals and newspapers are not always indexed or cataloged, making it a tedious process to generate systematic corpora. The increasing digitization of the press provides new opportunities for implementing computational tools for identifying comics.

Specifically, the use of computer vision techniques combined with machine learning pipelines (Ali et al. 2023) provides reliable approaches for automatically identifying comics in a digitized corpus. Machine learning approaches for historical documents have largely focused on the extraction of textual information from OCR results (Paaß & Konya, 2011; Bhatt et al., 2021); yet visual elements in these documents are also of interest to researchers. Visual object detection has focused on object detection (Lee et al., 2020), using a model with seven classes: photographs, illustrations, maps, comics, editorial cartoons, advertisements and headings. Smits & Wevers (2020) implemented convolutional neural networks (CNNs) to categorize digitized historical visual sources for example in detecting differences in photographs from illustrations. Other visual approaches considered segmentation and classification of newspaper elements using a (Almutairi & Almashan 2019) a Mask-RCNN model. Du, Le and Honig (2023) implemented a YOLO model to compare the similarity of images of printed images and original photographs. These approaches allow us to create new and unique “collections as data” datasets for research (Always Already Computational: Collections as Data); yet in order to implement this there is a need to develop a workflow that integrates knowledge and information retrieval (Jose et al., 2020).

Thus, we question what is a comic and which approaches computer vision & machine learning techniques can be implemented to identify said comics. Drawing on the massive digitization of mass-market illustrated periodicals from the Interwar period in Belgium, carried out at KBR in the framework of the ARTPRESSE project (<https://www.artpresse-researchproject.com/>); we take a selection of six digitized periodicals in Dutch and French (*Hebdo, A-Z/Bonjour, Le Soir illustré, Ons Volk, ABC, Zondagsvriend*), between 1934 and 1940, to define the characteristics of comics of this period. The selected period covers a moment of transformation for comics forms, under the influence of the increased import of American comics (as of 1934) until the outbreak of WWII.

We then consider a number of approaches including YOLOv7 (Chien-Yao, Alexey & Hong-Yuan Mark, 2022) and the Distant Viewing Toolkit which integrates a number of visual models (Arnold & Tilton 2020) to identify a set of comics for this period. The use of these pre-trained models allows us to computationally identify comics which encompasses both illustrations and comics included in other parts of the periodicals, for instance, advertisements. Therefore, this initial corpus is then manually annotated by two comic experts and two novices to identify a labeled dataset for retraining. This annotated dataset is then used to retrain the model to identify a comics corpus that characterizes this period and presents various genres. We combine these results with related metadata to sketch a picture of comics in this time period. The paper will detail the methodological development of this project and share preliminary results, considering the affordances and obstacles of a computational approach to identifying comics in the magazines.



Works Cited

- Ali, D., Milleville, K., Verstockt, S., Van de Weghe, N., Chambers, S., & Birkholz, J. M. (2023). "Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections", *Journal of Documentation*. <https://doi.org/10.1108/jd-01-2022-0029>.
- Almutairi, A. and Almashan, M. (2019). "Instance segmentation of newspaper elements using mask R-CNN", *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Boca Raton, FL, pp. 1371-1375, December.
- Arnold, T., & Tilton, L. (2020). Distant viewing toolkit: A python package for the analysis of visual culture. *Journal of Open Source Software*, 5 (45), 1800.
- Bhatt, J., Hashmi, K.A., Afzal, M.Z. and Stricker, D. (2021), "A survey of graphical page object detection with deep neural networks", *Applied Sciences*, 11 (12), p.5344.
- Du, L., Le, B., & Honig, E. (2023). Probing Historical Image Contexts: Enhancing Visual Archive Retrieval through Computer Vision. *ACM Journal on Computing and Cultural Heritage*.
- Jose, J.M., Yilmaz, E., Magalhaes, J., Castells, P., Ferro, N., Silva, M.J. and Martins, F. (Eds), (2020). *Advances in Information Retrieval: 42nd European Conference on IRResearch, ECIR2020*, Lisbon, Portugal, April 14-17, 2020, Springer Nature, Proceedings ,Part I, Vol. 12035.
- Paaß, G. and Konya, I. (2011). "Machine learning for document structure recognition", in *Modeling, Learning, and Processing of Text Technological Data Structures*, Springer, Berlin, Heidelberg, pp.221-247.
- Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019). Always already computational: collections as data. Available at: <https://collectionsasdata.github.io/>.
- Wang Chien-Yao, Bochkovskiy A., and Liao Hong-Yuan Mark (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.
- Wevers, M., & Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35 (1), 194-207.