Advances in Digital Music Iconography

Benchmarking musical instrument localization in non-photorealistic images

Matthia Sabatelli, Nikolay Banar, Marie Cocriamont, Eva Coudyzer, Karine Lasaracina, Walter Daelemans, Pierre Geurts and Mike Kestemont*
* = presenting author

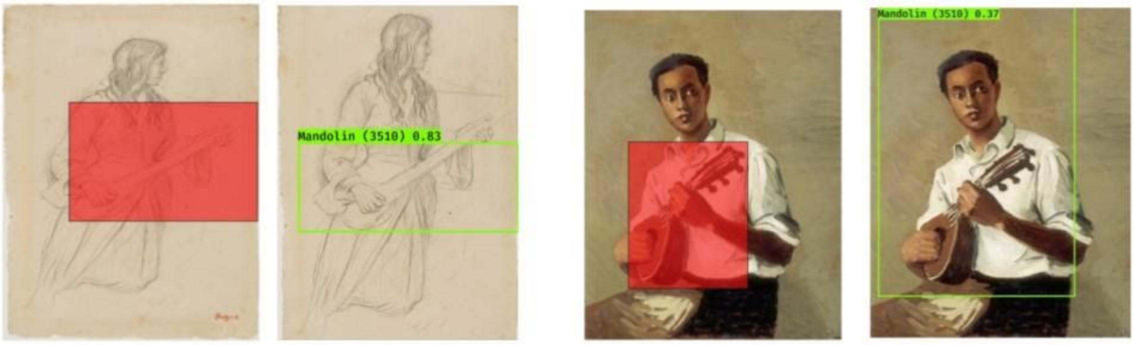
Paper abstract for DHBenelux 2020, 3rd - 5th June 2020

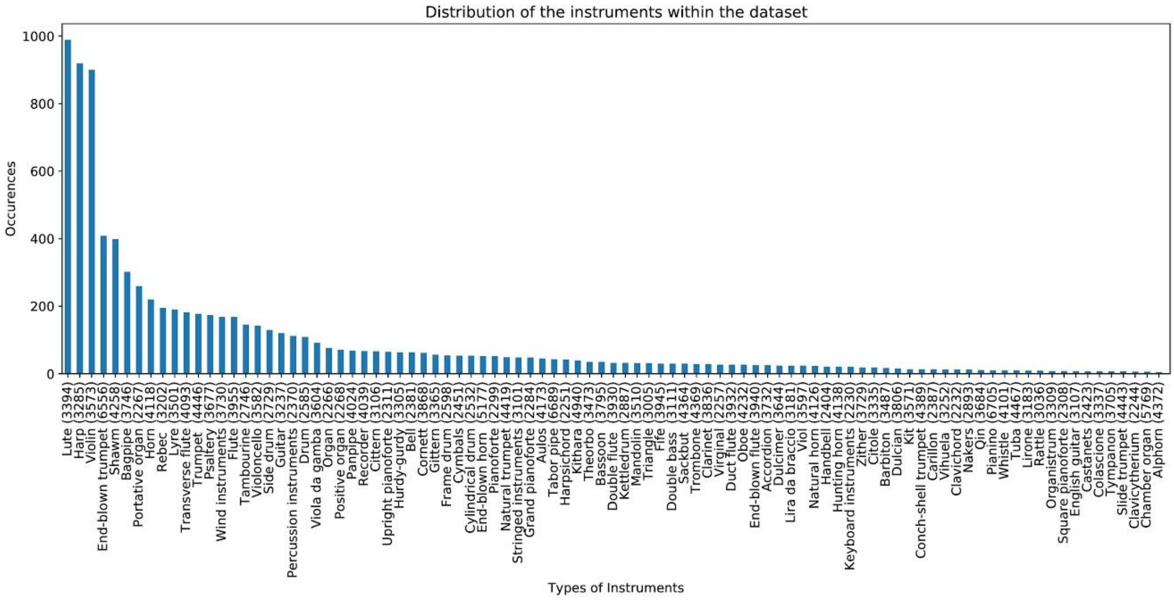
Introduction

In the past decade, deep neural networks have significantly pushed the state of the art in computer vision [LeCun et al. 2015]. In the Digital Humanities too, the potential of computer vision is nowadays increasingly recognized [Wevers & Smits 2019; Arnold & Tilton 2019]. These new methodologies have indeed a privileged role to play in the exploration of large heritage collections. This paper is situated in a multidisciplinary project in which we investigate how modern artificial intelligence can support GLAM institutions (galleries, libraries, archives, and museums) in cataloguing and curating their rapidly expanding digital assets. One major hurdle is that modern computer vision gravitates towards photo-realistic material, i.e. digital images that do not actively attempt to distort the reality they depict -- such as the influential ImageNet dataset [Russakovsky et al. 2015], that offers highly realistic photographic renderings of everyday concepts. While some more recent heritage collections abound in such photorealistic material (e.g. advertisements in historic newspapers), traditional photography does not take us further back in time than the nineteenth century [Hertzmann 2018]. Additionally, the Humanities study many other visual arts that prioritize much less photorealistic representation or even completely 'fictional' renderings of historical realities.

Minerva

We present Minerva ('Musical INstrumEnts Represented in the Visual Arts'): a novel benchmark data set in the field of object detection, focused on the detection of musical instruments in non-photorealistic image collections. This task can be situated in the field of music iconography, a discipline on the brink of musicology and art history, studying the object, themes, and subject matter relating to music as they are represented in the visual arts [Baldassarre 2004]. The discipline was professionalized with the development of the *Répertoire International d'Iconographie Musicale* (RIDIM) in 1971, and the establishment of the Center for Musical Iconography (RCMI) at the City University in New York in 1972. RIDIM now functions as a reference image database, designed to facilitate efficient yet powerful description and discovery of music-related art works [Green & Ferguson 2013]. Using the conventional method of rectangular bounding boxes, we have manually annotated around 16,000 musical instruments in more than 10,000 images within the Cytomine software environment [Marée et al. 2016]. To increase the interoperability of this data set, we have identified the instruments using MIMO keywords, drawn from an international database of musical instruments that aggregates metadata from musical instrument museums [mimo-international.com]. Using this publicly available benchmark data, we have stress-tested the available technology for the identification and detection of objects in images.





Distribution of instrument categories in the Minerva dataset

Classification

We have investigated whether convolutional neural networks are able to correctly classify the instrument depictions in the dataset. To this end, we have extracted the various patches delineated by the bounding boxes in Minerva as stand-alone instances. Next, we tackled this task as a standard machine-learning classification problem for which we applied a representative selection of established network architectures, pretrained on the Rijksmuseum dataset [Mensink and Van Gemert 2014; Sabatelli et al. 2018]. Below, we report the results in terms of Accuracy and F1-score for the Minerva test sets. For the individual instruments, we do so for four versions of the dataset of increasing complexity: top-5/top-10/top-20 instruments and the entire dataset. Analogously, we report the scores for a classification experiment where the object detector is trained on instrument hypernyms (e.g. ‘string instrument’ instead of ‘violin’) as class labels (‘Granular’).

	Top5		Top10		Top20		All		Granular	
CNN	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
R-Net	62.30	48.51	44.52	17.29	33.62	10.43	22.80	0.87	**71.82**	**55.41**
V3	**71.07**	**64.54**	**49.52**	**25.90**	34.07	8.18	**25.12**	**1.79**	71.55	53.80
V19	53.33	38.28	40.66	16.01	**36.29**	**10.49**	21.64	0.23	66.07	40.36

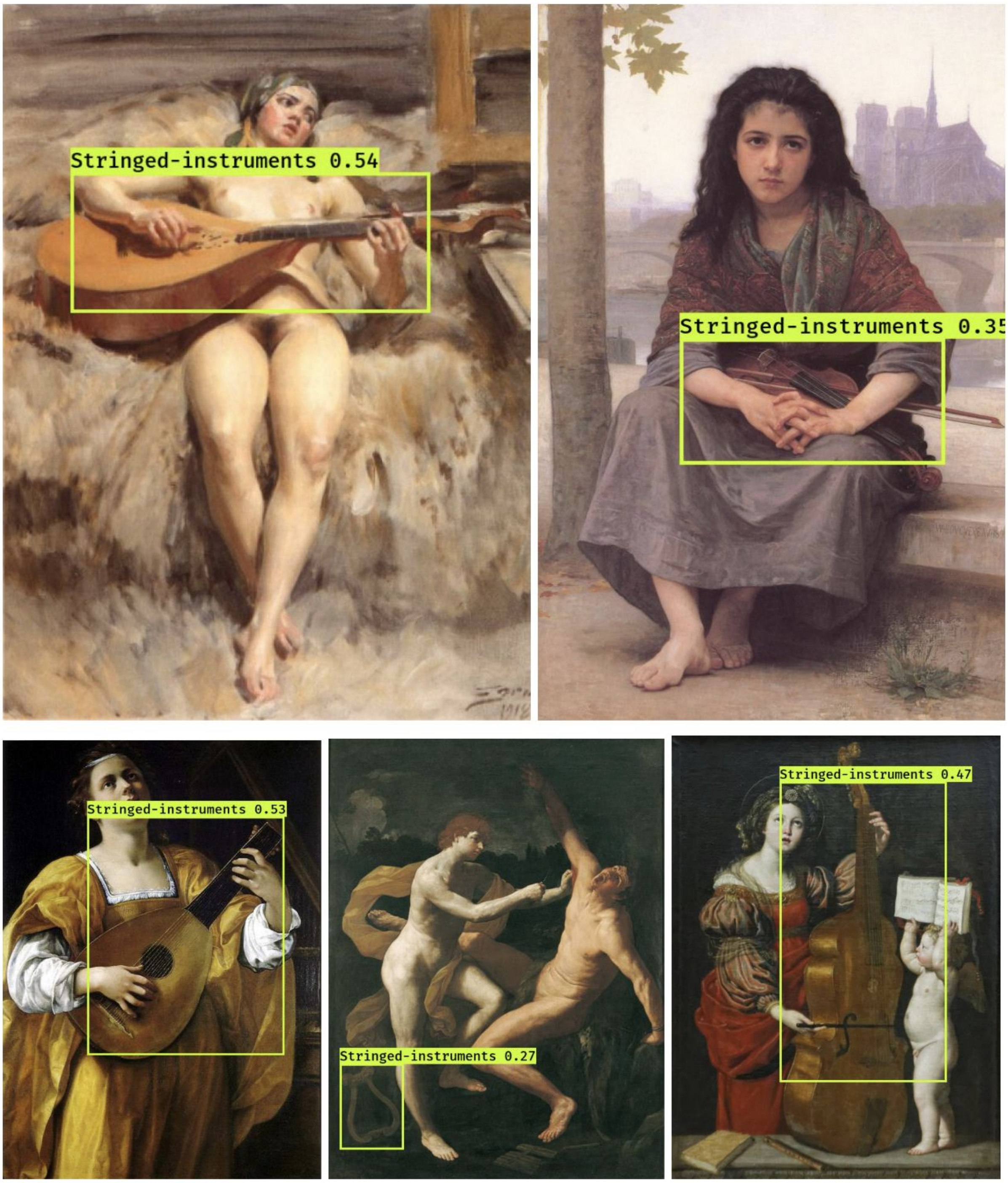
Detection

We also report the results for the detection benchmarks introduced in Minerva, using the popular YOLO-V3 architecture [Redmon and Farhdi 2018]. To assess the network's performance, we follow an evaluation protocol based on the "Intersection over Union" (IoU), whereby detected bounding boxes are compared to the ground truth box on the Cytomine platform. The table below lists precision, recall and average precision (AP) scores for a number of representative instrument "hypernyms".

Instrument ≥ IoU	**Precision**	**Recall**	**AP**
Single-instrument ≥ 10	0.63	0.47	40.26%
Single-instrument ≥ 50	0.48	0.36	26.28%
Stringed-Instruments ≥ 10	0.60	0.46	37.99%
Stringed-Instruments ≥ 50	0.51	0.39	28.94%
Wind-Instruments ≥ 10	0.52	0.36	27.11%
Wind-Instruments ≥ 50	0.33	0.13	7.07%
Percussion-Instruments ≥ 10	0.33	0.11	4.64%
Percussion-Instruments ≥ 50	0.29	0.10	3.98%
Keyboard-Instruments ≥ 10	0.60	0.35	25.07%
Keyboard-Instruments ≥ 50	0.45	0.26	17.59%

Discussion

To illustrate the broader relevance of our approach, we have also applied the trained benchmark system for object detection 'in the wild', on out-of-sample heritage data, such as the *WikiArt* collection, followed by a quantitative and qualitative evaluation of the results (informally illustrated in the figure below). During the talk, we will present an in-depth error analysis, which has yielded a number of unexpected insights into the contextual cues that trigger the detector. The iconography surrounding children and musical instruments, for instance, shares some core properties with the depiction of musical instruments, such as an intimacy in body language. All in all, our benchmark experiments highlight the feasibility of the classification and detection tasks under scrutiny but also, and perhaps primarily, the significant challenges that state-of-the-art machine learning systems are still confronted with on this data, such as the "long-tail" of the instruments' distribution and the staggering variance in depiction across the images in the dataset.


Stringed-instruments 0.54
Stringed-instruments 0.35
Stringed-instruments 0.53
Stringed-instruments 0.27
Stringed-instruments 0.47

Cherry-picked examples of successful detections in WikiArt for “stringed instruments”.

Bibliography

- [Arnold and Tilton 2019] Arnold, T., and Tilton, L., Distant viewing: analyzing large visual corpora. Digital Scholarship in the Humanities, https://doi.org/10.1093/digitalsh/fqz013 (2019): advance access.
- [Baldassarre 2008] Baldassarre, A. "Music Iconography: What is it all about? Some remarks and considerations with a selected bibliography", Ictus: Periódico do Programa de Pós-Graduação em Música da UFBA, 9 (2008), 55-95.

- [Green and Ferguson 2013] Green, A., and Ferguson, S. "RIDIM: Cataloguing music iconography since 1971", Fontes Artis Musicae, 60 (2013), 1-8.
- [Hertzmann 2018] Hertzmann, A. "Can Computers Create Art?", Arts, 7 (2018) doi:10.3390/arts7020018.
- [LeCun et al. 2015] LeCun, J., Bengio, Y., and Hinton, G., “Deep Learning”, Nature, 521 (2015): 436–444.
- [Marée et al. 2016] Marée, R., Rollus, L. Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J., Kainz, P., Geurts, P., and Wehenkel "Collaborative analysis of multi-gigapixel imaging data using Cytomine", Bioinformatics, 32 (2016): 1395–1401.
- [Mensink and Van Gemert 2014] Mensink, T. and Van Gemert, J. "The Rijksmuseum challenge: Museum-centered visual recognition". In Proceedings of International Conference on Multimedia Retrieval, page 451. ACM, 2014.
- [Russakovsky et al. 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, K., Khosla, A., Bernstein, M. et al. "Imagenet large scale visual recognition challenge". International journal of computer vision, 115(3):211–252, 2015.
- [Sabatelli et al. 2018] Sabatelli, M., Kestemont, M., Daelemans, W. and Geurts, P. "Deep transfer learning for art classification problems". In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–646, 2018.
- [Wevers & Smits 2019] Wevers M., and Smits, T. "The visual digital turn: Using neural networks to study historical images", Digital Scholarship in the Humanities, https://doi.org/10.1093/llc/fqy085 (2019): advance access.