Blind Dates: Examining the Expression of Temporality in Historical Photographs

Alexandra Barancová[1,†], Melvin Wevers[2,*,†] and Nanne van Noord[3]

[1]Faculty of Humanities, Media Studies, University of Amsterdam, the Netherlands
[2]Faculty of Humanities, Amsterdam School of Historical Studies, University of Amsterdam, the Netherlands
[3]Faculty of Science, Informatics Institute, University of Amsterdam, the Netherlands

Abstract

This paper explores the capacity of computer vision models to discern temporal information in visual content, focusing specifically on historical photographs. We investigate the dating of images using OpenCLIP, an open-source implementation of CLIP, a multi-modal language and vision model. Our experiment consists of three steps: zero-shot classification, fine-tuning, and analysis of visual content. We use the *De Boer Scene Detection* dataset, containing 39,866 gray-scale historical press photographs from 1950 to 1999. The results show that zero-shot classification is relatively ineffective for image dating, with a bias towards predicting dates in the past. Fine-tuning OpenCLIP with a logistic classifier improves performance and eliminates the bias. Additionally, our analysis reveals that images featuring buses, cars, cats, dogs, and people are more accurately dated, suggesting the presence of temporal markers. The study highlights the potential of machine learning models like OpenCLIP in dating images and emphasizes the importance of fine-tuning for accurate temporal analysis. Future research should explore the application of these findings to color photographs and diverse datasets.

Keywords

Image dating, Computer vision, Temporal analysis, Historical photographs, OpenCLIP

1. Introduction

Time plays a crucial role in shaping our understanding and interpretation of the world around us. Our perception of duration, the sequence of our memories, and the authority with which historical records organize the past all contribute to our lived experiences and memory. This perception extends to our interpretation of visual content, where an image's materiality, content, and style can convey critical temporal information. Despite its significance, this aspect of image understanding remains underexplored in artificial intelligence research [18, 16].

AI models, typically trained on limited data periods, possess a narrow understanding of temporality due to their lack of 'awareness' of historical variations. Although efforts have been made to integrate historical data into language models [11] and even encode time explicitly [20],

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

*Corresponding author.

†These authors contributed equally.

a.barancova@uva.nl (A. Barancová); m.j.h.f.wevers@uva.nl (M. Wevers); n.j.e.vannoord@uva.nl (N. van Noord)

http://www.melvinwevers.nl (M. Wevers); https://nanne.github.io/ (N. van Noord)

0000-0001-8177-4582 (M. Wevers); 0000-0002-5145-3603 (N. van Noord)

CEUR Workshop Proceedings (CEUR-WS.org)

these methods primarily focus on text, leaving visual content interpretation largely uncharted territory. In this paper, we experiment with the task of 'dating' images, predicting when an image was taken based on its visual content.[1] We examine how different image aspects influence a multimodal AI model's predictive accuracy, uncover structural biases in pre-trained computer vision models, and explore their effects on predictions. Our research aims to extend our understanding of the visual representation of time and its influence on image interpretation. This experiment is situated within a broader goal of developing more temporally-aware computer vision and multimodal models. For this, cross-pollination between AI and humanities scholarship on cultural heritage, archiving, and temporality will be needed; as [4] show, interdisciplinary work in this area has been limited, yet it has the potential to be mutually beneficial.

2. Background

The challenge of automating dating has been addressed across a variety of historical objects, spanning from photographs [18] and artworks [13, 8] to archaeological sites [9, 25]. With the increasing digitization of historical documents, many of which lack publication dates, computational methods have been employed to estimate their creation dates, primarily analyzing writing styles [6], focusing on both the text and the visual content of the writing. The automatic dating of historical photographs offers substantial value to archives and museums, but also domains like temporal forensic analysis, where dating can serve as evidence. Forensic applications typically concentrate on an image's material aspects, using techniques that identify specific camera models or devices [12, 1]. While such methods may be overly meticulous for large-scale dating of historical images, they underscore the importance of material information in establishing an image's capture date.

Beyond material aspects, others show that low-level image features like RGB color derivatives and color angles carry temporal information. Models trained on these features often surpass human accuracy in dating photographs [18, 3]. Research in this domain has seen the adoption of neural networks for dating photographs, treating it as an ordinal [10], regression [15], classification [21, 5, 24], or retrieval task [14]. Studies have also started to pay attention to image content, emphasizing the connection between time and visual elements or semantic cues. Research indicates that temporal cues can be derived from human appearance features, such as clothing, hairstyles, and glasses [21, 5], or even from architectural elements like windows to estimate the age of buildings [24]. A recent study on a family photo album dataset found that the accuracy of the model used for the dating task improved as the number of faces and/or people in a photograph increased [23] – this suggests that certain high-level image features and visual elements may carry more temporal information than others. Finally, [2] recently turned to generative approaches to synthesize portrait images for specific decades between 1880 and the present day to distinguish visual markers for these periods.

As we transition our focus from photograph materiality to content, an essential challenge arises: deciphering how models interpret higher-level input features to predict dates. This exploration aims to yield deeper insights into the ways in which temporality is encoded in

[1] All code used for this study is available on GitHub: https://github.com/CANAL-amsterdam/dating-images/

visual content and how we can enhance computer vision systems' ability to interpret cultural artifacts.

3. Image Dating

We use OpenCLIP [7], the open source implementation of CLIP [19], to predict when a photograph was taken. CLIP is a multi-modal language and vision model that has been shown to have a strong zero-shot capability on diverse vision tasks [19] and to outperform a number of domain-specific models on various vision and language tasks following task-specific fine-tuning [22]. Our interest lies in understanding how visual features, particularly objects, are leveraged for dating purposes, while also evaluating the model's aptitude for the dating task. Among the various models that have been used for dating images, we have not yet seen experiments with large models that have zero-shot capabilities. Experimenting with the potential of such models is interesting due to the lesser need for training data, their broader generalizability and the possibility to examine a multimodal, based on textual and visual data, perspective on tasks like dating.

Data Collections of press photographs are available with relatively reliable dates, making them well-suited for examining the visual representation of time. For this experiment, we have chosen to use the *De Boer Scene Detection* dataset, which contains 52,160 digitized historical press photographs from the *De Boer* newspaper agency spanning from 1945 to 2005.[2] The images are scanned and cropped photo negatives, the vast majority of which are gray-scale. [27] The dataset contains relatively mundane photographs, rather than iconic ones, as well as a wide variety of different scenes ranging from sporting events to landscapes [26]; this makes it an interesting case for exploring the visual elements that carry temporal information. Besides the exact year, each image has a label describing the depicted scene. We excluded images lacking date information and those taken before 1950 and after 1999, resulting in 39,866 photographs. These cut-off points were chosen to include data spanning complete decades in our analysis. Figure 1 illustrates the dataset distribution per year. We split the dataset in a train and test set using stratified sampling based on the year, with the aim of reducing uneven distributions across the splits to prevent biases, in 80% and 20% respectively.

We structure our experiment in three steps: examining zero-shot classification capabilities, fine-tuning the model, and assessing the impact of visual content on the model's dating ability.

Zero-shot Classification. To investigate to what extent OpenCLIP can be used for dating we apply zero-shot classification to the test set to predict the photograph's date. This process uses the prompt 'a photograph from the year *x*' where *x* ranges from 1950 up to 2000. We employ Mean Absolute Error (MAE) for performance evaluation, following [3, 15]. We find an MAE of 15.8, which indicates relatively poor performance considering both the 50 year range of our dataset and the comparison to results others have demonstrated on the dating task — [15] and [14], for example, attained MAE values of up to 7.12 and 7.48 respectively in their experiments with the Date Estimation in the Wild (DEW) dataset that covers the period

[2] The dataset is available at https://zenodo.org/record/7137452

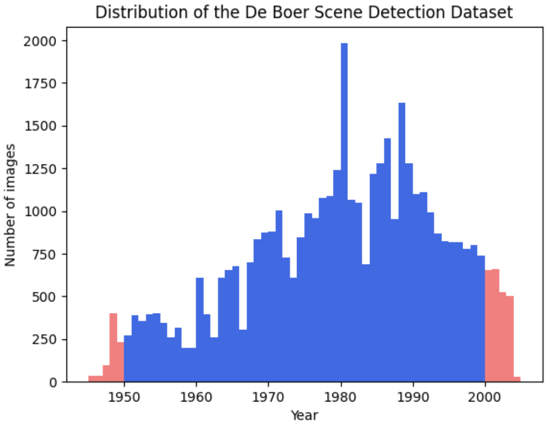


Figure 1: Distribution graph of the *De Boer Scene Detection* dataset by year. We utilized the data marked in blue in our experiments.

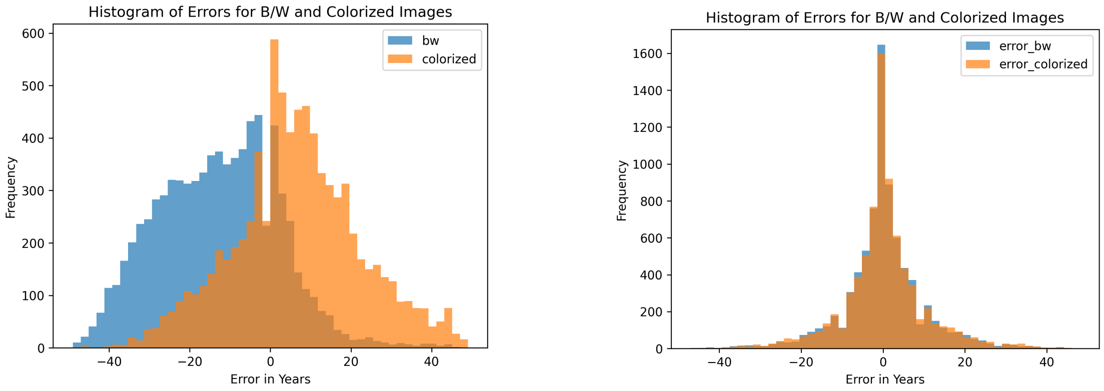
1930-1999. Additionally, [15] showed that human participants had an MAE of 10.9 on the DEW dataset and as such were on average about 11 years off in their predictions. Comparatively, the almost 16 years that the zero-shot model achieves is quite poor.

The error rate distribution reveals a preference in the zero-shot model for predicting dates earlier than the actual date. Suspecting a correlation with the images' gray-scale nature, we tested the zero-shot classification with a colorized version of our dataset,[3] resulting in the error distribution leaning slightly more toward the future (see Figure 2a).[4] Figure 3 shows two sample images for which colorization had a large effect on decreasing date prediction error. One depicts an outdoor view of a church, the other a group of people in formal wear. Both images show a large prediction error in the gray-scale variant, albeit in different directions, i.e. overestimate or underestimating the actual date. We find that colorizing the images improves the overall zero-shot capabilities of OpenCLIP, however, with an MAE of 13.2 it is still relatively ineffective for the dating task.

Fine-tuned Classifier. To overcome the zero-shot limitations we explore whether fine-tuning OpenCLIP improves performance on the dating task and eliminates the bias found in the initial experiment. To this end, we train a logistic classifier using the OpenCLIP image embeddings. Training a logistic classifier also allows us to focus solely on the temporal information in the visual content, removing possible confounding temporal bias in the model from text, through prompting. When using textual prompts for zero-shot classification, the words used might be better suited for specific historical periods, thereby introducing a bias in the images corresponding to this prompt. Fine-tuning reduced the error and the bias, with the MAE being 6.65 for the classifier trained and evaluated on the original gray-scale images and 6.79 for the colorized images. The bias between gray-scale and colorized images found in the zero-shot

[3] We colorized all the photographs using DeOldify https://github.com/jantic/DeOldify

[4] A KS-test (KS-statistic: 0.49, p-value: 0.0) supports the difference between the distributions.



(a) Zero-shot classification

(b) Fine-tuned classification

Figure 2: Histograms showing the error distribution for the zero-shot and fine-tuned classification.

approach disappears after fine-tuning (Figure 2), displaying a more normal error distribution.[5]

Content Analysis. Upon training a model to predict dates, we investigated the content of the images to determine whether specific visual features improved or hindered the predictions. An initial analysis using the available scene labels proved to be inconclusive as there were large differences in the within-scene error rates, as well as a large variety of visual features represented in individual scenes. We opted to examine the visual features at the object level, using Detectron2 outputs [28]. For the detection task, we used the 80 default objects, as defined in COCO, default ROI threshold (0.5). Next, we only selected detection with a confidence threshold above 0.8 and types that appeared more than 200 times in the entire data set, in order to shorten the long tail. The confidence threshold was output by Detectron2 per image. Finally, we picked 12 classes representing modes of transport and living beings to focus this experiment on.[6]. Our motivation in picking these classes was to reduce the granularity of the available categories so as to identify larger trends; classes like ‘tie‘ for example, might be closely related to ‘person’, essentially functioning as a sub-class thereof. An additional motivation for excluding some of the MS COCO object classes, is that they did not all suit the context and/or time span of our dataset, especially technology like ‘laptop’, ‘cell phone’ or ‘microwave’.

A Bayesian regression analysis was conducted to measure the effect of object presence and absence on the error rate.[7] The regression model was defined as follows:

$$prediction\_error = 1 + object\_presence$$

Where *prediction_error* is the outcome variable, *object_presence* is the binary prediction variable indicating each object’s presence. To model the distribution, we assumed a negative binomial. We model the errors as counts, where the event is the counts of predictions with a

[5] Also supported by KS-test (K-S statistic: 0.00, p-value: 0.92).

[6] ‘bicycle’, ‘boat’, ‘bus’, ‘car’, ‘motorcycle’, ‘train’, ‘truck’, ‘bird’, ‘cat’, ‘dog’, ‘horse’, ‘person’

[7] The analysis was performed using the Python library Bambi and the NumPyro nuts sampler.

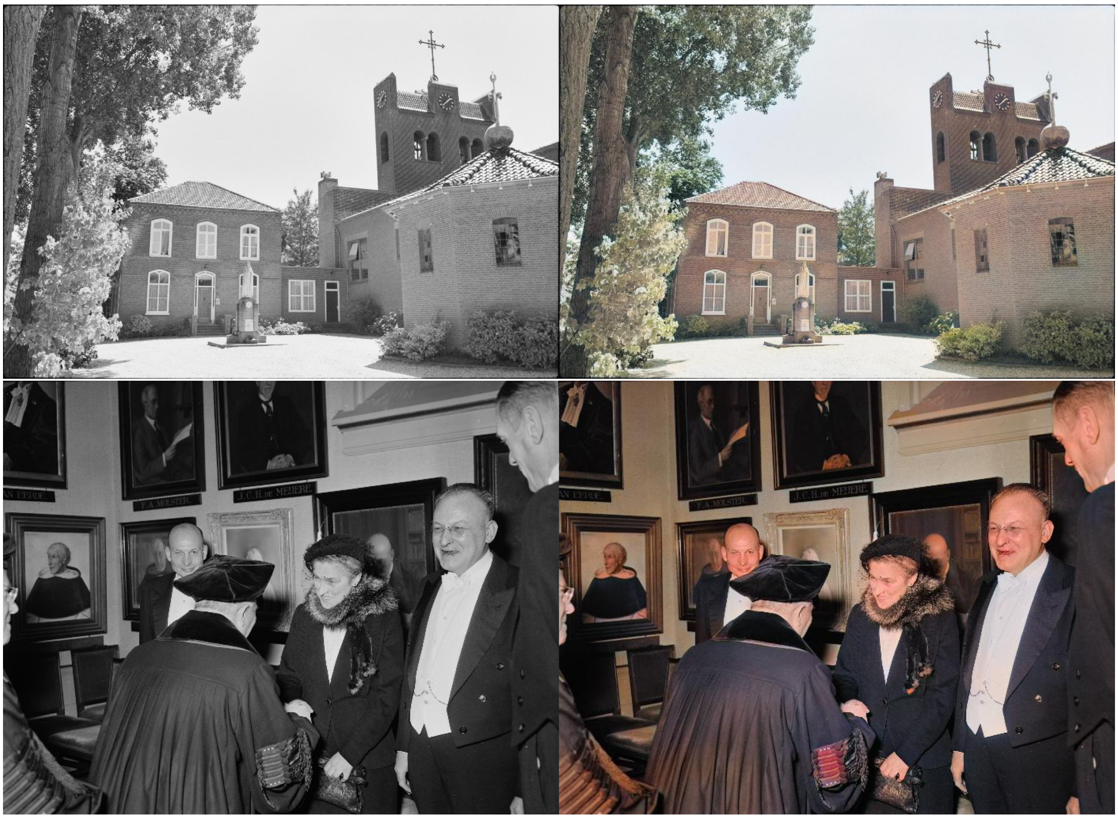
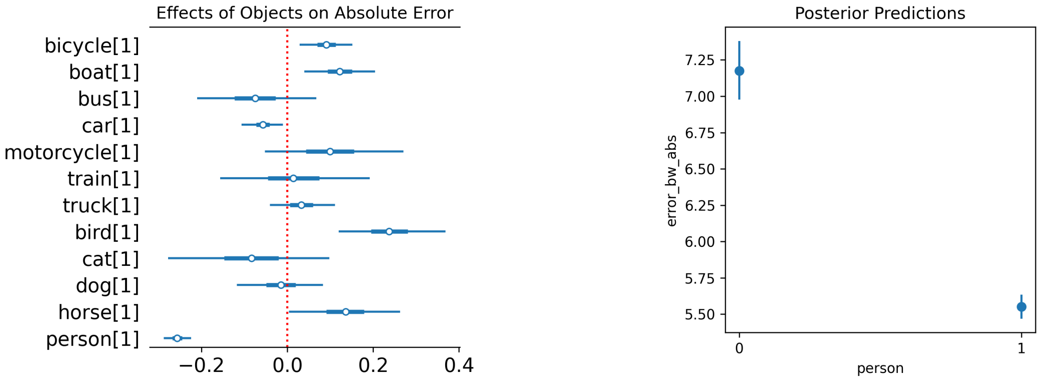


Figure 3: Examples of original (left) and colorized (right) photographs from the *De Boer Scene Detection* dataset for which colorization had the largest impact on decreasing error for zero-shot classification. Colorization decreased the error from -37 to 0 for the top image (actual year: 1999, prediction original: 1952, prediction colorized: 1999), and from 37 to 2 for the bottom image (actual year: 1952, prediction original: 1999, prediction colorized: 1954).

specific error.[8]

Figure 4a shows that for modes of transport, the presence of 'bicycle', 'boat', 'motorcycle', and 'train' increase the absolute error, whereas 'bus' and 'car' decrease the absolute error. We hypothesize that these vehicles are more prone to exterior changes in this period. Of the animals, we see that 'bird' and 'horse' increase the error and 'cat' decreases the error. Our hypothesis here is that cats might be depicted more often together with humans and in interior environments, which may include more temporal markers. Finally, we see that having a 'person' in an image has the strongest effect, decreasing the MAE from approximately 7.2 to 5.5 (Figure 4b), indicating that depictions of people convey visual cues about time. These results and hypotheses need further examination, which we intend to undertake in future work.

[8] Since the variance and mean are not equal a zero-inflated Poisson was not warranted. See the GitHub for more information on the models.



(a) Estimated effects of objects on absolute error. HDI .95, meaning that there is a 95% chance that the true value lies within this range.

(b) Posterior predictions for the class ‘person’. A person in the image reduces the absolute error from 7.2 to 5.5.

Figure 4: Output of the regression model. The MAE values are based on the entire dataset, without taking into account random effects, which results in higher reported MAE values. However, for this analysis the change in MAE is of primary interest.

4. Conclusion and Discussion

Our exploration with OpenCLIP to date historical press photographs yielded several findings.

Ineffectiveness of Zero-shot Classification. Our first finding is that the zero-shot classification capability of OpenCLIP does not perform well in dating images. The model demonstrated a distinct bias towards predicting earlier dates, which we attribute to the gray-scale nature of our images. This suggests that OpenCLIP may have learned to associate gray-scale with older photographs, as [17] also concluded in exploring the concept of history in foundation models including CLIP. We attempted to counteract this bias by colorizing the images, which mildly improved the model’s accuracy and shifted the bias towards predicting more recent dates. However, despite these adjustments, the efficacy of zero-shot classification for this task remained limited.

Improvement through Fine-tuning. Our second finding is that fine-tuning OpenCLIP using a logistic classifier significantly enhances the model’s performance. The fine-tuned model effectively eliminates the bias towards past dates seen in the zero-shot approach and offers comparable accuracy levels for both gray-scale and colorized images. This indicates that the presence of color in images becomes less significant for dating them when the model is trained to focus on visual content. Future research could look into generating captions rather than labels or scenes to provide a more enriched context for each image.

Objects as Temporal Markers. Our third finding, coming from the post-hoc regression analysis, is that the presence of people in images generally leads to more accurate date predictions. This echoes the findings of [23]. We posit that this could be attributed to the time-dependent markers humans tend to carry, like fashion and hairstyles, as has previously been shown in studies on yearbook portraits by [5] and [21]. Moreover, we see that the presence of animals often kept as house pets also reduces the error, we hypothesize that this might be due them being photographed indoors or in proximity to humans, which might carry more temporal markers than animals, such as horses or birds that are captured in nature. Finally, certain modes of transportation increase the error rate while others decrease it. We need to explore to what extent this is related to innovations that lead to visual changes over time. All in, further investigation is necessary to validate these hypotheses; considering our findings on the influential role of human figures in the images, it would be worthwhile to explore datasets containing fewer human figures – in our dataset only 8,674 of the 39,866 photographs contained no people.[9] This could shed light on whether the presence of humans is generally advantageous for image dating, or if this is a specific characteristic of our dataset or a manifestation of model bias.

In conclusion, this study deepens our understanding of how computer vision models interpret and extract temporal information from historical visual material. It highlights the potential of OpenCLIP for image dating tasks. It also underscores the importance of model fine-tuning to counter biases. Future work could test our findings' generalizability to color images and datasets from various periods and geographical regions. Such work can be a means to identifying and using temporal information in visual material better, with the aim of creating more temporally-aware computer vision and multimodal models. To this end, we see case studies engaging with specific computer vision/ temporal tasks like image dating, as important steps in testing what works in terms of both models and data.

References

[1] F. Ahmed, F. Khelifi, A. Lawgaly, and A. Bouridane. "Temporal Image Forensic Analysis for Picture Dating with Deep Learning". In: *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE).* Southend, United Kingdom: Ieee, 2020, pp. 109–114. DOI: 10.1109/iCCECE49321.2020.9231160.

[2] E. M. Chen, J. Sun, A. Khandelwal, D. Lischinski, N. Snavely, and H. Averbuch-Elor. "What's in a Decade? Transforming Faces Through Time". In: *Computer Graphics Forum* 42.2 (2023), pp. 281–291. DOI: 10.1111/cgf.14761.

[3] B. Fernando, D. Muselet, R. Khan, and T. Tuytelaars. "Color features for dating historical color images". In: *2014 IEEE International Conference on Image Processing (ICIP).* Paris, France: Ieee, 2014, pp. 2589–2593. DOI: 10.1109/icip.2014.7025524.

[4] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James. "Machine Learning for Cultural Heritage: A Survey". In: *Pattern Recognition Letters* 133 (2020), pp. 102–108. DOI: 10.1016/j.patrec.2020.02.017.

[9] This is based on outputs from Detectron2 at an object confidence threshold of 0.8.

[5] S. Ginosar, K. Rakelly, S. M. Sachs, B. Yin, C. Lee, P. Krähenbühl, and A. A. Efros. "A Century of Portraits: A Visual Historical Record of American High School Yearbooks". In: *IEEE Transactions on Computational Imaging* 3.3 (2017), pp. 421–431. DOI: 10.1109/tci.2017.2699865.

[6] A. Hamid, M. Bibi, M. Moetesum, and I. Siddiqi. "Deep Learning Based Approach for Historical Manuscript Dating". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. Sydney, Australia: Ieee, 2019, pp. 967–972. DOI: 10.1109/icdar.2019.00159.

[7] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. *OpenCLIP*. 2021. DOI: 10.5281/zenodo.5143773.

[8] S. Khan and N. van Noord. "Stylistic Multi-Task Analysis of Ukiyo-e Woodblock Prints". In: *British Machine Vision Conference*. 2021, p. 14.

[9] S. Klassen, J. Weed, and D. Evans. "Semi-supervised machine learning approaches for predicting the chronology of archaeological sites: A case study of temples from medieval Angkor, Cambodia". In: *Plos One* 13.11 (2018), e0205649. DOI: 10.1371/journal.pone.0205649.

[10] Y. Liu, A. W.-K. Kong, and C. K. Goh. "Deep Ordinal Regression based on Data Relationship for Small Datasets". In: *Ijcai* (2017), pp. 2372–2378.

[11] E. Manjavacas and L. Fonteyn. "Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950)". In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. 2021, pp. 23–36.

[12] J. Mao, O. Bulan, G. Sharma, and S. Datta. "Device temporal forensics: An information theoretic approach". In: *2009 16th IEEE International Conference on Image Processing (ICIP)*. 2009, pp. 1501–1504. DOI: 10.1109/icip.2009.5414612.

[13] T. Mensink and J. Van Gemert. "The rijksmuseum challenge: Museum-centered visual recognition". In: *Proceedings of International Conference on Multimedia Retrieval*. 2014, pp. 451–454.

[14] A. Molina, P. Riba, L. Gomez, O. Ramos-Terrades, and J. Lladós. "Date Estimation in the Wild of Scanned Historical Photos: An Image Retrieval Approach". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by J. Lladós, D. Lopresti, and S. Uchida. Vol. 12822. Cham: Springer International Publishing, 2021, pp. 306–320. DOI: 10.1007/978-3-030-86331-9_20.

[15] E. Müller, M. Springstein, and R. Ewerth. ""When Was This Picture Taken?" – Image Date Estimation in the Wild". In: *Advances in Information Retrieval*. Ed. by J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, and J. Tait. Vol. 10193. Cham: Springer International Publishing, 2017, pp. 619–625. DOI: 10.1007/978-3-319-56608-5_57.

[16] N. van Noord, M. Wevers, T. Blanke, J. Noordegraaf, and M. Worring. *An Analytics of Culture: Modeling Subjectivity, Scalability, Contextuality, and Temporality*. 2022. DOI: 10.48550/arxiv.2211.07460.

[17] F. Offert. “On the Concept of History (in Foundation Models)”. In: *Image* 37.1 (2023), pp. 121–134. DOI: 10.1453/1614-0885-1-2023-15462.

[18] F. Palermo, J. Hays, and A. A. Efros. “Dating Historical Color Images”. In: *Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science.* Vol. vol 7577. Florence, Italy: Springer, 2012, pp. 499–512. DOI: 10.1007/978-3-642-33783-3_36.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Pmlr, 2021, pp. 8748–8763.

[20] G. D. Rosin, I. Guy, and K. Radinsky. “Time masking for temporal language models”. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 2022, pp. 833–841.

[21] T. Salem, S. Workman, M. Zhai, and N. Jacobs. “Analyzing human appearance as a cue for dating images”. In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV).* Lake Placid, NY, USA: Ieee, 2016, pp. 1–8. DOI: 10.1109/wacv.2016.7477678.

[22] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. *How Much Can CLIP Benefit Vision-and-Language Tasks?* 2021. DOI: 10.48550/arxiv.2107.06383.

[23] L. Stacchio, A. Angeli, G. Lisanti, D. Calanca, and G. Marfia. “IMAGO: A family photo album dataset for a socio-historical analysis of the twentieth century”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 18.3s (2022), pp. 1–23. DOI: 10.1145/3507918.

[24] M. Sun, F. Zhang, F. Duarte, and C. Ratti. “Understanding architecture age and style through deep learning”. In: *Cities* 128 (2022), p. 103787. DOI: 10.1016/j.cities.2022.103787.

[25] G. Toner and X. Han. *Language and chronology: text dating by machine learning*. Brill, 2019.

[26] M. Wevers. “Scene Detection in De Boer Historical Photo Collection:” in: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence.* Vienna, Austria: SCITEPRESS - Science and Technology Publications, 2021, pp. 601–610. DOI: 10.5220/0010288206010610.

[27] M. Wevers, N. Vriend, and A. de Bruin. “What to Do with 2.000.000 Historical Press Photos? The Challenges and Opportunities of Applying a Scene Detection Algorithm to a Digitised Press Photo Collection”. In: *TMG Journal for Media History* 25.1 (1 2022), pp. 1–24.

[28] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.