Searching for cultural relationships through deep learning models

Lorenzo Stacchio[1], Alessia Angeli[2], Giuseppe Lisanti[2] and Gustavo Marfia[3,*]

[1]University of Bologna, Department for Life Quality Studies
[2]University of Bologna,Department of Computer Science and Engineering
[3]University of Bologna, Department of the Arts

Abstract

Family album photo collections may reveal historical insights regarding specific cultures and times. In most cases, such photos are scattered among private homes and only available on paper or photographic film, thus making their analysis very cumbersome. Their study may also become difficult because of the number of photos that such collections contain. It would be exceedingly long to manually verify the characteristics of more than a few hundred photos, considering that often no associated descriptions are available. This work falls in the described domain, addressing the problem of dating an image resorting to the analysis of an analog family album photo dataset, namely IMAGO, containing photos shot in the 20th century. Thanks to the IMAGO dataset, it was possible to apply different deep learning-based architectures to date images belonging to photo albums without needing any other sources of information. In addition, with the implementation of cross-dataset experiments, which also involved models previously presented in the literature, it was possible to observe temporal shifts which may be due to known intercultural influences. Concluding, deep learning models revealed their potential not only in terms of their performance but also in terms of their possible applications to intercultural research.

Keywords

family album, analog photographs, date estimation, intercultural influences, deep learning,

1. Introduction

Family albums represent an example of vernacular photography that has drawn the attention of researchers and public institutions. Scholars from different fields agree in identifying such collections as capable of capturing salient features regarding the evolution of local communities in space and time. However, contributions in this field usually base their findings on the study of small corpora of photos [1, 2], since a large-scale analysis is often impeded as they are too many to be processed manually. Many research initiatives have addressed the problem of processing and analyzing digital images. It is more difficult to find initiatives focused on analog ones,

VIPERC2022: 1st International Virtual Conference on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, 12 September 2022

*Corresponding author.
lorenzo.stacchio2@unibo.it (L. Stacchio); alessia.angeli2@unibo.it (A. Angeli); giuseppe.lisanti@unibo.it (G. Lisanti); gustavo.marfia@unibo.it (G. Marfia)
0000-0002-9341-7651 (L. Stacchio); 0000-0002-3572-2076 (A. Angeli); 0000-0002-0785-9972 (G. Lisanti); 0000-0003-3058-8004 (G. Marfia)

mainly because printed images are scattered in numerous public and private collections, of variable quality, and worn out due to their prolonged use in time. In essence, any analysis employing image processing and computer vision algorithms requires the time-consuming and potentially degrading initial digitization step. Despite the complications and challenges brought on by analog photographs, they represent an unparalleled source of information regarding the recent past [3, 4]. The different clothes that people wear, their haircut styles, the tools and machinery, the natural landscape, the overall environment, etc., may exhibit the culture of a given time and place. All of these visual features may amount to important cues to estimate the shooting year [5]. This work addresses the problem of dating an image, exploiting the IMAGO collection of family album photos, started in the year 2004 at the University of Bologna [2]. Such collection contains digitized versions of analog prints with specific characteristics. Each image portrays at least one person, and the lion's share of such photos has been shot in a given area of Italy by Italian citizens.

We here perform a dating analysis of the IMAGO collection, exploiting different deep learning-based architectures, without using any other source of information. Differently from [6], we here perform a more thorough analysis, comparing different Convolutional Neural Network (CNN) architectures for the dating task; we then trained a model which combines different salient image regions together to estimate the date; finally, we also attempt to verify possible intercultural influences (i.e., the adoption of different customs and habits in different epochs and countries) by analysing the differences in dating, resulting from a cross-dataset experiment, in which we employ the datasets from [7, 8].

The rest of the paper is organized as follows: in Section 2 we review the state of the art that falls closest to this contribution. Section 3 describes the considered dataset, along with its pre-processing and splitting. Sections 4 and 5 present and validate several deep neural networks models applied to the proposed dataset. In Section 6, we report and discuss cross-dataset experiments from an intercultural influence perspective. Finally, in Section 7 an overall discussion is carried out, along with possible future works.

2. Related Work

Only a few works have proposed so far the dating of collections of vernacular photographs, also taking into account analog ones [9, 7, 8, 10, 5]. In [7] the authors employed a deep learning approach to analyze and date 37,921 historical frontal-facing American high school yearbook photos taken from 1928 to 2010. Here, a CNN architecture was trained to analyze people's faces and predict the year in which a photo was taken. In addition, the authors observed a gender-dependency in the performance of dating models. Along the same line, the authors of [8] presented a dataset containing images of students taken from high school yearbooks, covering the 1950 to 2014 time span (considering 1,400 photos per year). They also resorted to CNNs to estimate the date of an image, to evaluate the quality of color vs. grayscale, considering the following features: faces, torsos (i.e., people's upper bodies including faces), and random regions of images. The best performance was obtained with the torsos of people. In addition, their results provide cues that human appearance is related to time. In [10], instead, dating was implemented through the analysis of images belonging to the years 1930 through 1999. Vernacular and

landscape photos were considered, including at most than 25,000 pictures per year. The authors proposed different baselines relying on CNNs, using regression and classification approaches. Differently, in [5], the authors formulated the date estimation task as an image-retrieval one where, given a query, the retrieved images are ranked in terms of date similarity. For their study, they analyzed the same public dataset employed in [10].

The contributions so far presented focused on the dating of vernacular photographs shot in heterogeneous settings (e.g., landscapes, portraits). To the best of our knowledge: (i) none has considered a dataset solely containing analog pictures depicting at least one person and belonging to 20th-century family albums, and (ii) no other works have also considered a cross-dataset and intercultural perspective when approaching the dating task.

3. Dataset, pre-processing and splitting

The IMAGO collection was introduced in [6][1]. It represents a digital collection of Italian analog family album photos composed by 16,642 labeled images taken between 1845 and 2009, to focus on image dating. Fig. 1 reports the number of labeled images available per year in the 1930 to 1999 time frame, exhibiting the unbalance in terms of the number of photos per year (most fall between 1950 and 1980).

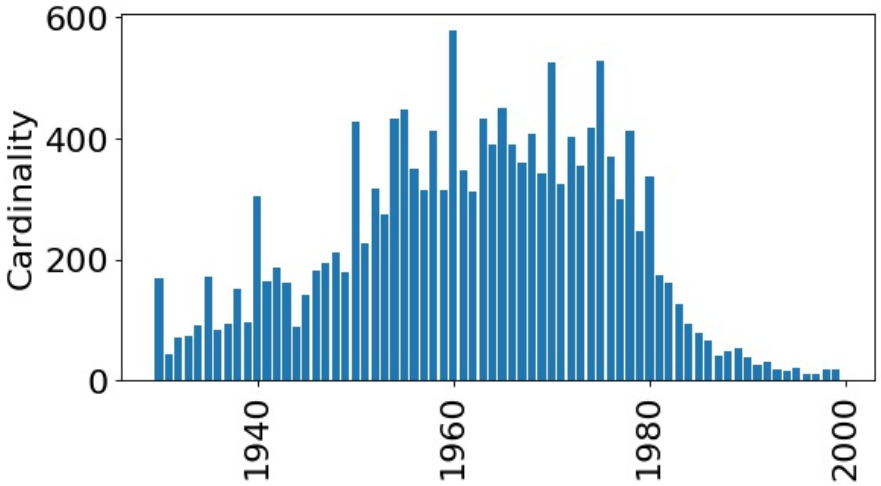


Figure 1: IMAGO classes distribution

The overall available images in this interval amount to 15,673. Out of such time intervals, the number of available images is too little to be considered. Fig. 2, shows four exemplar images from the IMAGO dataset, belonging to different decades. Here, it is possible to appreciate what characterizes each photo (e.g., number of people, clothing, colors, and location), highlighting one of the main ones, i.e., each portrays at least one person.

The pre-processing phase carried out on the IMAGO dataset aimed at isolating the regions of interest which could enhance the performance of the deep learning models (more details in

[1]The IMAGO dataset is available upon request.

Figure 2: IMAGO image samples

Sections 4). Following insight from [8, 7] we extracted from each image of the IMAGO dataset, referred to as FULL-IMAGES, all the faces and full figure crops of the people portrayed, gathered in FACES and PEOPLE sets. Important to note that such patches are always present since we are dealing with photos that always include at least one person. In particular, for FACES and PEOPLE images, we processed each image of the IMAGO dataset using an open-source implementation of YOLO-FACE [11] and YOLO [12], respectively. Then, the FACES and the PEOPLE images have been constructed accounting for the number of people portrayed in a photo. Indeed, adopting a fixed size bounding box may result in the possible loss of pixels related to the faces or people's full figures. For this reason, we rescale the provided bounding boxes used to crop a face/people depending on the number of people portrayed in a photo, i.e., the greater the number of people, the smaller the bounding box. Fig. 3 shows an IMAGO full-image sample with the respective crops taken from FACES and PEOPLE.

full-image face crop person crop

Figure 3: FULL-IMAGES, FACES and PEOPLE image samples

It is possible to appreciate that PEOPLE images include details that are not present in FACES ones (e.g., the clothing of a person). Finally, we verified the utility of performing denoising [13, 14] and super resolution [15, 16] operations, as all the images derive from scans of analog prints. Nevertheless, since the overall improvement obtained adopting such strategies were revealed to be negligible, we hence opted for an analysis based on the original scans.

The FULL-IMAGES dataset has been then partitioned into three subsets of pictures: 80% for training and 20% for testing. In addition, 10% of the training set is used as a validation subset. In particular, for each image in the train, validation, and test sets of IMAGO, the faces and the people there portrayed are extracted and added to the corresponding FACES and PEOPLE sets, respectively. This process guarantees that no faces or people crops from the validation or test sets are observed during the training phase.

4. Model architectures and training settings

In this work, we considered single and multi-input deep learning architectures. The former analyzed the FULL-IMAGES, FACES, and PEOPLE images in isolation, while, the latter, instead, their combination. More in detail, we employed three well-known CNN architectures pre-trained on ImageNet [17]: ResNet-50 [18], InceptionV3 [19] and DenseNet121 [20]. Each model was modified replacing the top-level classifier with a new classification layer, whose structure depends on the number of output classes, with randomly initialized weights. In addition, the pre-trained convolutional layers were fine-tuned with the given input data. For what regards the single-input classifiers, one has been trained per each set of images, and named following the analyzed patches: *full-image*, *faces* and *people*. For the FACES and PEOPLE images we evaluated the accuracy, not for a single face or person, but agglomerating the activations for all of those who appeared in a picture. This means that if a picture contained n persons, the final prediction would obtained by passing to the softmax function the average of the activations coming from each face or person in that image.

For the multi-input classifiers, instead, we defined the Merged model which combines together the single-input classifiers introduced before, with the aim not only to exploit different sources of information but also to learn how. Hence, a new training session was carried out as the newly introduced network was asked to learn how to perform such a combination. In particular, the pre-trained single-input classifiers were employed, but the classification layer was removed, preserving the CNN backbone as feature extractors. Adopting such architecture, the cardinality of the different extracted feature vectors depends on the number of faces/people portrayed in an image, and the average of such feature vectors was computed to combine them with the vector obtained from the full image. As a picture could contain more than one person, multiple FACES and PEOPLE images could stem from a single one in FULL-IMAGES. The three resulting feature vectors were linearly combined employing a weighted sum, whose weights were a set of three real scalars learned during the training phase. The final vector, resulting from the linear combination, is fed to a fully connected layer with a softmax activation, yielding the final probability vector used for the classification.

Moving to the training settings, we applied for all the considered patches random cropping, and horizontal flipping. Each model was fine-tuned using a weighted cross-entropy loss and an Adam optimizer with a learning rate of $1e-4$ and a weight decay of $5e-4$. We set the batch size to 32 for the training on the *full-images* classifier and to 64 for *faces* and *people*.

5. Experimental Results

The results are expressed in terms of time distances, as in [7, 8]. The time distance defines the tolerance accepted in predictions concerning the actual year. For example, if a photo was labeled with the year 1942 and the model returned 1937 (or even 1947) this would be considered correct for if the time distance is set to be equal or greater than 5, otherwise it represent an error. In this work, model accuracies were computed considering temporal distances of 0, 5, and 10 years. The results are reported in Table 1.

It is possible to appreciate that different baseline models (i.e., ResNet-50, InceptionV3,

Table 1
Model accuracies for different time distances (d = 0, d = 5, d = 10)

	Single-input classifier		
	ResNet-50	**InceptionV3**	**DenseNet121**
time distance	***full-image***		
d = 0	**11.31**	10.45	10.68
d = 5	**62.56**	61.38	60.77
d = 10	82.54	**82.82**	82.47
time distance	***faces***		
d = 0	**15.01**	14.60	12.91
d = 5	**58.09**	56.95	57.81
d = 10	78.39	78.46	**79.70**
time distance	***people***		
d = 0	**15.77**	12.56	13.99
d = 5	**62.40**	60.04	59.69
d = 10	**82.47**	81.39	81.42

	Multi-input classifier		
	ResNet-50	**InceptionV3**	**DenseNet121**
time distance	**Merged**		
d = 0	**18.71**	17.14	16.22
d = 5	**67.59**	67.56	66.67
d = 10	86.17	**86.30**	86.07

DenseNet121) return similar accuracies. In addition, Table 1 exhibits different accuracies when different single input classifiers, hence different image patches, are considered. In particular, the *faces* and the *people* classifiers slightly outperform the *full-image* one. These results could be firstly explained by the model averaging obtained from the ensembling of multiple regions when FACES and PEOPLE images are considered, as the use of more data allows controlling the uncertainty and reducing the prediction error [21]. These results may also be due to the fact that each model exploits different salient cues from people's appearance (e.g., dresses, hairstyle, earrings, trousers). When comparing the results of the different approaches, the multi-input model (Merged) improves compared to the single-input classifiers. In this case, the performance

improvement can be explained by both the ensembling of multiple regions and the fact that the Merged model has learned to fuse the features from different classifiers.

In the analyses that follow, the ResNet-50 was selected as reference backbone for the models, since it provided the best trade-off between accuracy and model dimension [22].

To effectively estimate the value, in terms of prediction performance and, in particular, the comparison between the power of human (e.g., faces and people) vs. non-human features in image dating, we also considered random-patches. To study the possible use of non-human features we created a set of images called RANDOM, comprising eight randomly cropped regions, of 128×128 pixels, from each image belonging to FULL-IMAGES. Other window sizes were also tested but returned a lower performance. On top of this set of images, we fine-tuned an additional ResNet-50 to study its performance against the other models. The evaluation followed the same protocol already described for the *faces* and *people* classifiers in Section 4. The accuracies obtained with the single-input *random* classifier are **11.64** for time-distance equal to 0 (d = 0), **54.26** for d = 5 and **76.12** for d = 10. It is interesting to observe that, as also exhibited by *faces* and *people* classifiers, the *random* one achieved a slightly higher score with respect to the *full-image* classifier, considering a time distance equal to 0. However, it exhibited a lower accuracy than all the other classifiers with greater time distances. Even taking into consideration the averaging effect, this difference in performance between the *random* and the other classifiers may be caused by the different learned visual characteristics of given time-slices. This said, and considering that the time distance normally adopted in historical analyses is ±5 years, as reported in literature [2], we did not consider the RANDOM images and the *random* classifier in the rest of our study.

We also investigated which cues led the trained models to determine the specific year of a picture. We applied the Grad-Cam algorithm [23] to delimit the areas exploited by the deep learning models to perform the classification. In Fig. 4 are reported the grad-cam results for some correctly classified images.

In particular, each row corresponds to a specific decade and includes the grad-cam of an IMAGO full-image, and the two corresponding FACES and PEOPLE images, respectively. It is possible to see that the single-input classifiers focused on different regions. This may support the increased accuracy obtained in the multi-input model: different single-input classifier exploits different features. From a historical perspective, these visual results may be exploited to verify whether the highlighted cues correspond to visual factors which are recognized as representative for a specific period.

6. Cross-dataset experiments: evidences of intercultural influences?

To study the effects of possible intercultural influences (i.e., the adoption of different customs and habits in different epochs and countries) we carried out a cross-dataset study. We considered the datasets reported in [7, 8, 10, 5]. While [10, 5] included vernacular photos in heterogeneous settings and countries, where often no people are portrayed, [7, 8] analyzed American datasets comprising people's faces and torsos. Although such datasets do not include family album photos, they share some common traits with IMAGO: people in pictures are often in pose and

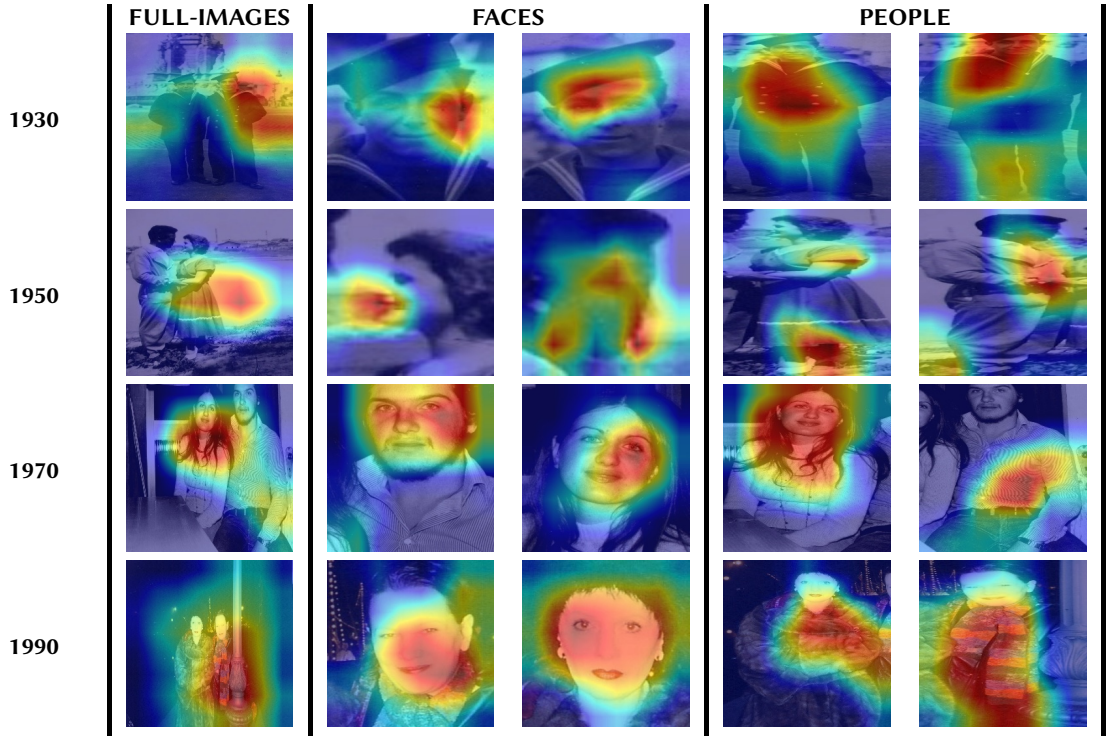
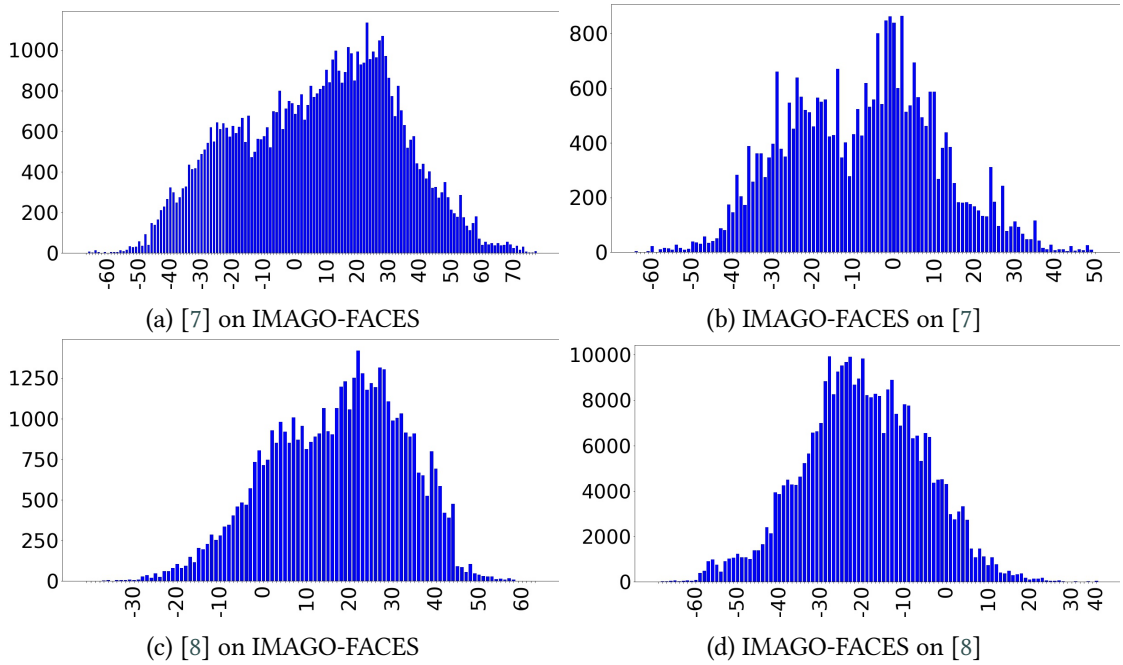


Figure 4: Grad-Cam image samples spread over different decades

dressed for a specific occasion. In particular, it is possible to extract what characterizes all of them: people's faces and torsos. This allowed us to perform a cross-dataset comparison considering the models trained on the IMAGO-FACES and PEOPLE patches and the models trained to exploit the datasets introduced in [7, 8], switching the considered evaluating datasets. To do this, we firstly fine-tuned the architectures used in [7, 8] following the procedures described in their experimental sections. The dataset introduced in [7] considers people's faces, while the one introduced in [8] offers both people's faces and torsos. Then, we evaluated these models on the IMAGO dataset. Vice versa, the *faces* and *people* classifiers, presented in this work, have been evaluated on the corresponding regions offered in the datasets from [7, 8]. For a fair evaluation, the experiments were carried out on the 1930-1999 time-span for the [7] vs. IMAGO comparison, while considering the 1950-1999 for the [8] vs. IMAGO one, respectively. In particular, we collected the error between the predicted and the actual year per each picture. The error distributions are reported in Fig. 5 for the cross-dataset experiments involving faces images.

In particular, in Figs. 5a, 5c the error distributions shifted towards positive values, while, in Figs. 5b, 5d towards negative ones. The models built on top of American datasets [7, 8] applied to IMAGO-FACES tend to overestimate the image shooting year while the opposite phenomenon (underestimation) occurs when the model presented in this work is applied to [7] and [8]. The same phenomenon appeared considering people's torsos. This fact could be due to different reasons. The images contained within the considered datasets have been acquired



(a) [7] on IMAGO-FACES

(b) IMAGO-FACES on [7]

(c) [8] on IMAGO-FACES

(d) IMAGO-FACES on [8]

Figure 5: Dating errors distributions

from different places and locations, using different cameras and scanning devices, leading to what is defined as the problem of dataset shift. However, there is another dimension to consider: the effect of the intercultural influences. Indeed, during the second half of the 1900 people's appearance from USA and Italy were influenced by each other [24, 25]. Finally, the obtained results, even if not confirmatory, provide us clues about possible intercultural influences: the model trained with Italian pictures underestimates the American ones while the model trained with Americans overestimates the Italian ones. These results are not final but certainly motivate further investigations on this topic.

7. Conclusions and future works

In this work, we analyzed the problem of image dating exploiting the IMAGO dataset, a collection composed of analog prints belonging to family albums and shot during the 20th century. We trained and tested single and multi-input deep learning models exploiting different regions of a given photo to identify its shooting year. We adopted these models to search for cues of intercultural influences through cross-dataset experiments. We evaluated the models trained on IMAGO-FACES images and the classifiers trained on the datasets exposed in [7, 8], following a cross-dataset configuration. The dating error distributions exhibited an interesting symmetry that motivates further experiments. This work may benefit from the use of larger and more balanced amounts of data and a deeper analysis of the different IMAGO image regions. We could also resort to different sources of historical information (e.g., journals, archival documents) to

multimodally approach the dating problem, mimicking, even more, the process that is usually carried out by historians in their analyses.

Acknowledgments

This work was supported by the University of Bologna with the Alma Attrezzature 2017 grant and by AEFFE S.p.a. and the Golinelli Foundation with the funding of two Ph.D. scholarships.

References

[1] M. Sandbye, Looking at the family photo album: a resumed theoretical discussion of why and how, Journal of Aesthetics & Culture 6 (2014) 25419.

[2] D. Calanca, Italians posing between public and private. theories and practices of social heritage, Almatourism-Journal of Tourism, Culture and Territorial Development 2 (2011) 1–9.

[3] G. Mitman, K. Wilder, Documenting the world: film, photography, and the scientific record, University of Chicago Press, 2016.

[4] MoMA, Vernacular photography, 2020.

[5] A. Molina, P. Riba, L. Gomez, O. Ramos-Terrades, J. Lladós, Date estimation in the wild of scanned historical photos: An image retrieval approach, in: International Conference on Document Analysis and Recognition, Springer, 2021, pp. 306–320.

[6] L. Stacchio, A. Angeli, G. Lisanti, D. Calanca, G. Marfia, Towards a holistic approach to the socio-historical analysis of vernacular photos, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022).

[7] S. Ginosar, K. Rakelly, S. Sachs, B. Yin, A. A. Efros, A century of portraits: A visual historical record of american high school yearbooks, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 1–7.

[8] T. Salem, S. Workman, M. Zhai, N. Jacobs, Analyzing human appearance as a cue for dating images, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–8.

[9] B. Fernando, D. Muselet, R. Khan, T. Tuytelaars, Color features for dating historical color images, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 2589–2593.

[10] E. Müller, M. Springstein, R. Ewerth, "When was this picture taken?"–image date estimation in the wild, in: European Conference on Information Retrieval, Springer, 2017, pp. 619–625.

[11] Thanh Nguyen, Yolo face implementation, https://github.com/sthanhng/yoloface, 2018. Online; accessed 3 August 2020.

[12] Joseph Redmon, YOLO: Real Time Object Detection, https://github.com/pjreddie/darknet/wiki/YOLO:-Real-Time-Object-Detection, 2019. Online; accessed 3 August 2020.

[13] Zhang, Kai, Zuo, Wangmeng, L. Zhang, Ffdnet: Toward a fast and flexible solution for cnn-based image denoising, IEEE Transactions on Image Processing, 2018.

[14] S. Paris, P. Kornprobst, J. Tumblin, F. Durand, A gentle introduction to bilateral filtering and its applications, in: ACM SIGGRAPH 2007 Courses, SIGGRAPH '07, Association for

Computing Machinery, New York, NY, USA, 2007, p. 1–es. URL: https://doi.org/10.1145/1281500.1281602. doi:10.1145/1281500.1281602.
[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, X. Tang, Esrgan: Enhanced super-resolution generative adversarial networks, 2018. arXiv:1809.00219.
[16] K. Zhang, Image restoration toolbox, https://github.com/cszn/KAIR, 2019.
[17] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.
[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, 2015. arXiv:1512.00567.
[20] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, 2018. arXiv:1608.06993.
[21] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.
[22] C. Coleman, D. Kang, D. Narayanan, L. Nardi, T. Zhao, J. Zhang, P. Bailis, K. Olukotun, C. Re, M. Zaharia, Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark, 2019. arXiv:1806.01427.
[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, International Journal of Computer Vision 128 (2019) 336–359. URL: http://dx.doi.org/10.1007/s11263-019-01228-7. doi:10.1007/s11263-019-01228-7.
[24] S. Gundle, M. Guani, L'americanizzazione del quotidiano. televisione e consumismo nell'italia degli anni cinquanta, Quaderni storici (1986) 561–594.
[25] W. post, How america became italian, https://www.washingtonpost.com/opinions/how-america-became-italian/2015/10/09/4c93b1be-6ddd-11e5-9bfe-e59f5e244f92_story.html?utm_term=.5a515dec12c5&noredirect=on, 2022.