Toward a Holistic Approach to the Socio-historical Analysis of Vernacular Photos

LORENZO STACCHIO, Department for Life Quality Studies, University of Bologna, Italy
ALESSIA ANGELI and GIUSEPPE LISANTI, Department of Computer Science and Engineering, University of Bologna, Italy
DANIELA CALANCA and GUSTAVO MARFIA, Department of the Arts, University of Bologna, Italy

Although one of the most popular practices in photography since the end of the 19th century, an increase in scholarly interest in family photo albums dates back to the early 1980s. Such collections of photos may reveal sociological and historical insights regarding specific cultures and times. They are, however, in most cases scattered among private homes and only available on paper or photographic film, thus making their collection and analysis by historians, socio-cultural anthropologists, and cultural theorists very cumbersome. Computer-based methodologies could aid such a process in various ways, speeding up the cataloging step, for example, with the use of modern computer vision techniques. We here investigate such an approach, introducing the design and development of a multimedia application that may automatically catalog vernacular pictures drawn from family photo albums. To this aim, we introduce the IMAGO dataset, which is composed of photos belonging to family albums assembled at the University of Bologna's Rimini campus since 2004. Exploiting the proposed application, IMAGO has offered the opportunity of experimenting with photos taken between the years 1845 and 2009. In particular, it has been possible to estimate their socio-historical content, i.e., the dates and contexts of the images, without resorting to any other sources of information. Exceeding our initial expectations, such an approach has revealed its merit not only in terms of performance but also in terms of the foreseeable implications for the benefit of socio-historical research. To the best of our knowledge, this contribution is among the few that move along this path at the intersection of socio-historical studies, multimedia computing, and artificial intelligence.

CCS Concepts: • **Applied computing** → **Digital libraries and archives**; • **Information systems** → *Multimedia information systems*; • **Computing methodologies** → *Computer vision;*

Additional Key Words and Phrases: Family photo albums, historical images, image dating, socio-historical context classification, multimedia application

ACM Reference format:
Lorenzo Stacchio, Alessia Angeli, Giuseppe Lisanti, Daniela Calanca, and Gustavo Marfia. 2022. Toward a Holistic Approach to the Socio-historical Analysis of Vernacular Photos. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 3s, Article 146 (October 2022), 23 pages.
https://doi.org/10.1145/3507918

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.

Authors' addresses: L. Stacchio, Department for Life Quality Studies, University of Bologna, Rimini, Italy; email: lorenzo.stacchio2@unibo.it; A. Angeli and G. Lisanti, Department of Computer Science and Engineering, University of Bologna, Bologna, Italy; emails: {alessia.angeli2, giuseppe.lisanti}@unibo.it; D. Calanca and G. Marfia, Department of the Arts, University of Bologna, Rimini, Italy; emails: {daniela.calanca, gustavo.marfia}@unibo.it.

1551-6857/2022/10-ART146 $15.00
https://doi.org/10.1145/3507918

1 INTRODUCTION

Following Kodak's invention of the first megapixel sensor in 1986, digital photography has slowly grown to substitute its analog predecessor, playing a key role in the early 21st-century digital revolution and social transformation [39, 53]. As a relevant example, photography has modified the way mobile phones are used, as their integration of digital cameras has at once fostered an exponential growth of the photos that are shot and uploaded to the Internet every year, as well as a paradigm shift in mobile communications, which today rely on high-quality multimedia [6, 15, 43, 63]. These phenomena have proven to be game-changers for both how people communicate and the bloom of new fields of research, as both academia and industry have exploited such plethora of visual data to develop and apply computer vision models to a variety of different problems (e.g., face recognition, autonomous driving) [16, 32, 33, 44, 59, 66]. Now, while a wealth of research is being devoted to the processing and analysis of digital images, much has to be done regarding analog ones, mainly because printed images representing a place (or an environment) at a given time may be (1) scattered in numerous public and private collections, (2) of variable quality, and (3) damaged due to hard or continued use or exposure. In addition, any analysis by means of image processing and computer vision algorithms requires the potentially quality-degrading initial digitization step. However, despite the complications and challenges brought on by analog photographs, they still represent an unparalleled source of information regarding the recent past: in fact, no other visual media has been used as pervasively to capture the world throughout the 20th century, as the availability of consumer-grade photo cameras supported the spread and popularity of vernacular photography practices (e.g., travel photos, family snapshots, photos of friends and classes) [35, 36].

Family photo albums represent an example of vernacular photography that has drawn the attention of researchers and public institutions. A recent work defines family photo albums as *a globally circulating form that not only takes locally specific forms but also "produces localities" that create and negotiate individual stories* [48]. Along the same lines, in another relevant contribution, family albums *represent a reference point for the conservation, transmission, and development of a community Social Heritage* [14]. In essence, scholars from different fields agree in identifying such type of photography collections as capable of capturing salient features regarding the evolution of local communities in space and time. A large-scale analysis of such collections of photos is often impossible, as a manual verification of the characteristics of more than a few hundred of pictures would be exceedingly burdensome, considering also that in many cases no associated descriptions are available. This is why contributions in this field normally base their findings on the study of small corpora of photos [14, 48]. This work addresses such problem, taking as a case study the socio-historical analysis of a collection of family album photographs: we here present the design and implementation of a multimedia application that, resorting to deep learning models, implements their classification for cataloging purposes. To verify the validity of such an approach, the application is exploited to classify a novel dataset, namely IMAGO, collected and maintained at the University of Bologna [14]. In particular, the contributions of this work amount to:

- A deep-learning-based multimedia application to assist socio-historians in their cataloging work, which consists in identifying the socio-historical information of an image, i.e., its shooting year and socio-historical context, according to the definitions provided in [14]. While the dating task has been so far considered in the literature [26, 37, 47], the estimation of the socio-historical context has not been yet investigated.
- The introduction of a family photo album collection, namely IMAGO, comprising over 80,000 analog photos taken between 1845 and 2009, belonging to ca. 1,500 families, primarily from the Emilia-Romagna and immediately neighboring regions in Italy.

- A thorough evaluation of the performance obtained by **Convolutional Neural Network (CNN)** models [29, 31, 55] trained on the IMAGO dataset for both the dating and the estimation of the socio-historical context. In order to assess the validity of the proposed framework, the performance of the proposed approach is also contrasted with the expertise of a socio-historical scholar.
- A comparison between the performance of the adopted CNN-based approach and a Transformer-based one [23, 58].

The rest of the article is organized as follows: In Section 2 we sketch the necessary socio-historical background. In Section 3 we review the state of the art that falls closest to our contribution. Section 4 aims at presenting our multimedia tool designed to assist socio-historians, whereas Section 5 provides a description and the main characteristics of the dataset adopted to verify the proposed approach. Section 6 presents and validates the models trained on the proposed dataset to define an evaluation baseline. In Section 7 we compare the classification performance of our application with the results obtained by a socio-historical scholar. In Section 8 the performance of the adopted CNN models is instead contrasted with the Transformer-based ones. Finally, in Section 9 an overall discussion is carried out and possible directions for future works are provided.

2 THE SOCIO-HISTORICAL BACKGROUND

In this section we sketch the socio-historical background required to set the stage for this work. In fact, no classification problem can be solved without first clarifying what the classification categories are. This review aims at providing the basics necessary to understand how contexts and categories emerge in socio-historical studies. To do this, we begin delineating the main differences between traditional and social history. We then explain how and why family photo albums fall within the areas of interest of such field of study, finally introducing the process that socio-historians implement when cataloging a corpus of data.

Social History amounts to an interdisciplinary field of research that combines sociological and historical methods to understand how societies have developed over time and how the past has and may influence the present [50]. In the words of Cabrera, traditional history and social history differ as follows: *Traditional history, especially classical political history, was based on the concept of subject: the subjectivity of historical agents was rational and autonomous; the subject a preconstituted center; and, therefore, actions were caused, and fully explained, by the intentions that motivated them. Social history, on the other hand, was based on the concept of society. For social historians, subjectivity and culture are not rational creations but representations or expressions of the social context in which the causes of actions were to be found* [9]. Such social contexts, with their own historical logic, represent the ground on which categories are constructed, to grasp the meaning and organize social reality [10]: the categories represent a complex relational network whose nature is neither subjective nor objective but the result of a specific historical phenomenon with its own behavior. Therefore, the categories do not constitute a simple mean for transmitting social reality but are an active part in its definition and are called *socio-historical contexts*.

Now, the starting point of a socio-historical analysis is the space in which the interweaving between individual initiative and social coercion takes place. An attempt is usually made to explain how society works on different theoretical bases resorting to traditional oppositions: public/private, subjective/objective, ideal/material, visible/invisible, body/conscience. Further analyses are then introduced turning to the concept of social imaginary, defined as “The way in which ordinary people imagine their social contexts which, often, does not translate into a theoretical formulation but is conveyed in images, stories and legends” [12]. In essence, any socio-historical context introduced in such analyses should describe the evolution of social history and therefore the change of sociality and of people's behavior in a defined space/time. To this aim, socio-historical categories are first identified studying historical archival documents from different topics (e.g., economics, traditions, wars). Among such documents, now contemporary historians also resort to multimedia sources [19]. Out of the many multimedia sources today available, photography emerges as the one capable to cover the greatest time-span so far, although photographs have risen to the dignity of primary sources of information just in the last few decades [54].

For the purposes of this work, socio-historical categories have been obtained relying on the study of family album photos. This particular kind of pictures originates from and at the same time represents a fundamental component of social structures, also a well known socio-historical abstraction, the Family [8]. The Family is, indeed, a fundamental construct in Social History studies, since it embodies at once the public and the private spheres. In fact, the photos contained in the family albums can be read, on the one hand, as private visual memories of one's own history, destined to remain hidden from society, and, on the other hand, as traces and signs of the collective social imaginary of a given historical period. So, family album photographs (e.g., spontaneous and/or anonymous images otherwise destined to remain hidden) depict the daily existence of their time, not considering them solely as memories but also as a network of signs, traces, and documents that may be used to interpret the past [54].

Although socio-historical contexts may emerge from the study of archival documents and family album photographs, the specific context of a specific photo may remain hard to tell. This is due the fact that, without knowing when a picture was taken and what the people there portrayed were doing, it may be impossible to associate any accurate information to the picture. Accurate information in most cases may be obtained only resorting to the knowledge of the subjects represented in the photo. For this reason, social-historians rely on the knowledge of the main source, if available, which may be the owner of the photograph, for example. Indeed, such information could be impossible to find: when studying and cataloging a corpus of photos, no reliable source of information may be available. This problem is common for socio-historical scholars, and in such case they resort to other approaches, which may include classifying data based on a visual inspection, implementing onerous processes to reduce as much as possible the misclassifications of socio-historical features. As a relevant example consider Enns and Martin [24], where the authors collected and visually analyzed and classified 355 photos related to women involved in agriculture learning activity.

3 RELATED WORK

In this section we analyze the works that fall closest to ours in terms of datasets and tasks. Only a few have so far analyzed analog collections of vernacular photographs [26, 37, 47]. For example, Ginosar et al. [26] employed a deep learning approach to analyze and date 37,921 historical frontal-facing American high school yearbook photos taken from 1928 to 2010 [26]. Here, a CNN architecture was trained to analyze people's faces and predict the year in which a photo was taken. Along the same line, Salem et al. [47] presented a dataset containing images from high school yearbooks, covering in this case the 1950-to-2014 time span (considering 1,400 photos per year). They resorted to CNNs to estimate the precise image shooting year. In order to assess the characteristics that allow to correctly classify a picture, they considered both color and grayscaled images containing (1) faces, (2) torsos (i.e., upper bodies including people's faces), and (3) random regions from the images. The best performance was obtained considering color images portraying the torso of people. Their results confirmed that the human appearance is strongly related to time. Müller et al. [37] instead analyzed the dating task through the lenses of vernacular and landscape photos belonging to years 1930 through 1999, amounting to at most 25,000 pictures per year. The authors proposed different baselines relying on deep CNNs, considering the dating as both a regression and

Table 1. Characteristics of Existing Datasets and IMAGO

Original Dataset	**Type(s) of Photography**	**Type(s) of Camera**	**Theme**	**Cardinality**	**Period**
Ginosar et al. [26]	Portrait	Digital and analog	Frontal face from high school yearbook	168,055	1905–2013
Salem et al. [47]	Portrait	Digital and analog	High school yearbook	ca. 600,000	1912–2014
Müller et al. [37]	Vernacular and landscape	Digital and analog	No specific theme	1,029,710	1930–1999
IMAGO collection	**Vernacular**	**Analog**	**Family albums**	**ca. 80,000**	**1845–2009**

a classification task. In Table 1 we summarize the characteristics (image content, number of images, and covered time span) of the archives employed in the works described so far. In most cases only specific subsets of such archives have been analyzed by means of computer vision techniques. To provide a comfortable comparison, the same information regarding the IMAGO collection (i.e., the collection originating the dataset analyzed in this work) is provided in the last row of the same table.

Other works have already investigated the digital cataloging of historical photos [2, 18, 45]. Lincoln et al. [34], for example, developed a prototype to find duplicates and tag photos depicting similar scenes in the Carnegie Mellon University Archives' General Photograph Collection. Tilton and Arnold [56], instead, draw on scholarship from semiotics and visual cultural studies to develop a framework called *distant viewing*, to individuate larger patterns within a corpus that may be difficult to discern by closely studying only a small set of objects (e.g., narrative arcs in American sitcoms). One of the works that falls closest in scope was published by Wevers and Smits [62], where the CHRONIC and the SIAMESET datasets were introduced to study the transition from illustrations to photographs in the history of Dutch newspapers.

Concluding, for the works and datasets cited in this section, no pre-defined socio-historical categories were utilized as means of analysis. In addition, none considered the family album theme: to the best of our knowledge, the present amounts to the first contribution to investigate their classification according to the socio-historical context definitions and background (Sections 2, 5.1, and 5.2).

4 A SOCIO-HISTORICAL CATALOGING TOOL FOR FAMILY PHOTO ALBUMS

Socio-historical analyses include dealing with various sources of information, systematically examining their soundness, exemplarity, and meaning, seeking for inter- and intra-correlations and relationships that may help understanding what really happened in the past [7]. Sources are in general not objective but *shaped by the politics, practices, and events that selectively document protest* [17]. In summary, the procedure of historical inquiry implies the following steps: (1) identification and selection of sources, (2) registration and classification for further investigation, and (3) a critical inquiry of the collection. From here, a socio-historian's work can then proceed in multiple directions. A sound socio-historical study may hence require the inspection and classification of hundreds or even thousands of documents and images [4, 24, 49]. This amounts to burdensome work, which often seeks for the big picture provided by large corpora of data rather than the specific information returned by a single document or image. Such type of process opens to the use of automatic tools, capable of classifying great amounts of data in short amounts of time. This has already been discussed over two decades ago, for example, in [25], where the author illustrated linguistic and statistical tools that could be profitably used by historians and social historians in the study of events. Now, much more can clearly be expected thanks to the development of computing tools,

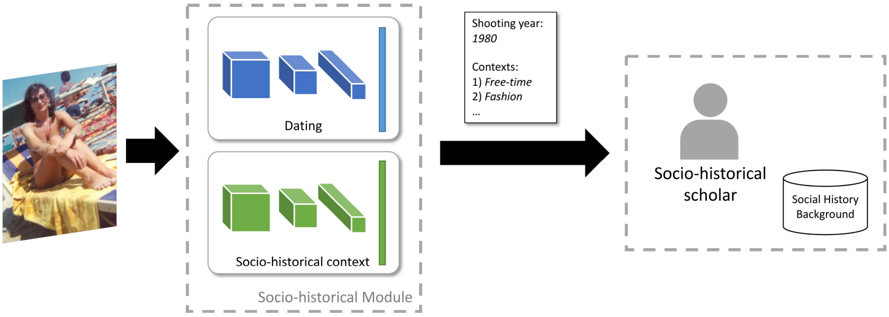


Fig. 1. Schema of the multimedia support application for socio-historians.

capable of handling growing amounts of multimedia data originating from heterogeneous sources. This would require a holistic approach taking care of source(s): (1) digitization, (2) accessibility through standard interfaces, and (3) analysis with models capable of translating socio-historical tasks into computing ones.

Now, a typical socio-historical task amounts to inferring from and subsequently applying categorical models to large corpora of data (Section 2). We apply such idea to the case of family photos, proposing a multimedia tool capable of processing and cataloging such type of pictures. To this aim, in Figure 1 we show the components of the proposed application. The core is the **Socio-Historical Module (SHM)**, which is composed by one or more classifiers, depending on socio-historical tasks of interest. For the purpose of this work, such tasks have been defined on top of family album photos, originating from the IMAGO dataset (details regarding its socio-historical value are discussed in Section 5). Such dataset offered the opportunity of predicting two pieces of socio-historical information: the context and the shooting year. In brief, the SHM amounts to a tool that may automatically label photos with the obtained predictions, giving, in addition, the opportunity of confirming or correcting such estimates, when necessary, during cataloging procedures.

The classifiers that compose the SHM could be defined exploiting different kinds of computer vision techniques. However, in the last decade, **Deep Learning (DL)** approaches have generally provided higher accuracies [51], both for the dating task [26, 37, 47] and for the analysis of historical picture datasets [34, 56]. For such reasons, we also exploited such tools in the development of the SHM. In particular, inspired by the work of Salem et al. [47], we trained several classifiers considering different image regions belonging to the same picture, selected using different criteria. To this aim, we considered the whole image and the crops enclosing the faces and the full figures of the people there portrayed. Such patches are always present since we are dealing with family album photos, which always include at least one person in each photo. To effectively estimate the value provided by such patches in terms of prediction performance, we also considered random ones. Hence, for the whole image and for each of the aforementioned regions, we trained two specific single-input classifiers, one per each of the two socio-historical tasks of interest. Such classifiers are named following the analyzed patches: full image, faces, people, and random patches. The single-input architecture utilizes either a CNN or a Transformer-based backbone and a fully connected layer for the final classification. It is important to notice that the results of such classifiers may not be comparable, as the amount of data utilized to perform a prediction varies depending on the fact that the full image is used during testing, or parts of it (patches). This fact required us to establish a different evaluation method, considering not a single face/person/random patch but introducing a layer that merged all of such activations into a single one per each picture. In

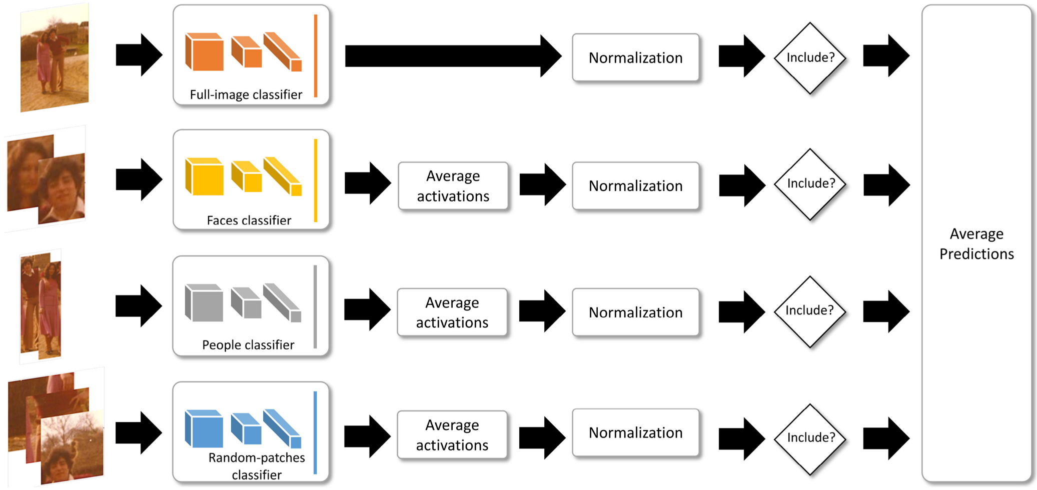


Fig. 2. Ensemble of the different models trained on the proposed datasets. Depending on the information exploited to obtain the final prediction, the activations from a model may be included or not.

practice, the activation vectors returned by a single-input classifier (e.g., the face classifier) for each face region were averaged per each image in order to compute the most probable class. This process was applied also to the people classifier and the random-patches classifier.

Finally, we also exploited the ensemble of these models (Figure 2). We resorted to such an approach as it has been successfully applied in the literature [41] and did not require any additional training and tuning of hyper-parameters. This kind of approach was employed not only to exploit the averaging effect [5] but also because it helps identify which type of classifier and data provide a valid contribution at inference time. As represented in Figure 2, such approach is modular, supporting the selection of the single-input classifiers. However, since we are considering activations coming from a single image or obtained averaging across multiple regions, these may contain values at different scales. For this reason, we $l2$-normalized the different inputs of the ensemble, to support the combination of the activation vectors coming from the full-image, faces, people, and random-patches classifiers. In particular, the final prediction is obtained by averaging the outputs described above and then computing the most probable class.

We now move on, in the following section, to present the details regarding the family photo collection considered in this work, the IMAGO collection.

5 IMAGO DATASET

The IMAGO project started in 2004 by socio-historical scholars to study the evolution of Social History through the lenses of family album photographs. This produced a digitized collection, namely IMAGO,[1] of analog family album photos gathered year by year and conserved by the Department of the Arts of the University of Bologna.[2] The collection comprises ca. 80,000 photos, taken between 1845 and 2009, belonging to ca. 1,500 Italian family albums, offering the opportunity of studying the evolution of Italian society during the 20th century. Among these, 16,642 images have been labeled by the bachelor's students in the Fashion Cultures and Practices course, under the supervision of the socio-historical faculty.

[1] The IMAGO resources are available upon request.
[2] imago.unibo.it.

5.1 Annotation Process

The annotation process followed (and keeps following, as new photos are acquired from new incoming bachelor's students in Fashion Cultures and Practices and annotated every year) a simple but strict protocol, involving the following steps:

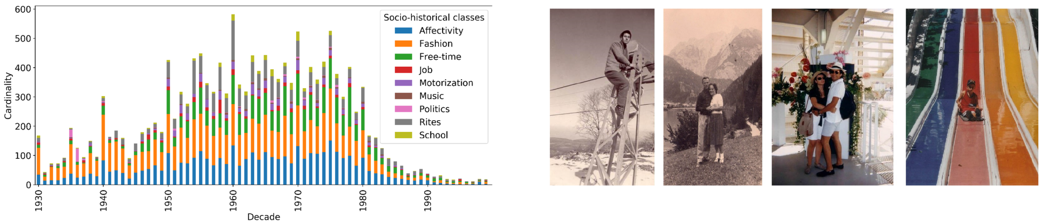
(1) During a first lecture, the socio-historical background, the IMAGO dataset construction project, and the different classification categories are presented and explained.
(2) During a second lecture, the annotation problem is covered in more detail. In particular, the lecture focuses on the importance of the reliability and authenticity of sources of socio-historical materials (including the shooting year). This means explaining that the original owner of the photo should be interviewed whenever possible. In case such person(s) are not available (e.g., the photo is very old), one can find a second-hand informed party (e.g., anyone who might be aware of the context of the given photo). Alternatively, an attempt to infer the socio-historical context and the shooting year (if possible) can be made analyzing any written annotations scripted behind the photo. In case none of such solutions are possible, no annotation is added.

Hence, the information provided by a photograph's owner amounts to the ground truth from a socio-historical point of view. This assumption in the labeling process is what injects the social component along with the historical one in the dataset. Such an approach is not new to the computer vision community either; other works in literature have considered as image metadata the information provided by their owners [3, 38]. These elements highlight the uniqueness of such datasets: since only the owner (or a directly connected party such as a relative or a friend) holds the ground truth, it is not possible to resort to just any standard labeling services (e.g., Amazon SageMaker Ground Truth or the Google AI Platform Data Labeling Service [1, 27]). This annotation process generated two socio-historical metadata per each photo: (1) the socio-historical context and (2) the shooting year [14].

5.2 Socio-historical Context

We here explain how the classes employed to analyze IMAGO have been defined from a socio-historical point of view. To this aim, we here report on the rationale behind the use of two exemplar ones, "Motorization" and "Affectivity," while a more in-depth analysis of all classes may be found in [11–13, 54]. The "Motorization" class is meant to mark an important change in people's lifestyle. We can take as an example the boom of sale for motorcycles. Such phenomena not only changed the production trend and its related economical ecosystem but also changed the social behavior of people in the area in which such boom took place. It affected the society idea of mobility and of how people gathered together. In these terms, the motorization aspect becomes therefore fundamental for the study of Social History. On a completely different plane, instead, the "Affectivity" class regards personal feelings. Such class wants to represent the changes that occurred between the affective and family relationships. For example, in the first decades of the 20th century, the family emotional relationships were considered ones of estrangement. This phenomenon is also reflected in the photographs that depict wife and husband, parents and children, brothers and sisters. Although all members of the same family, they all posed without any affectional gestures (e.g., hugs). After World War II, things change, starting with younger people who changed poses in terms of distances, contacts, hugs, and so forth. In the following we provide the socio-historical categories individuated in the IMAGO dataset [54], along with a brief explanation:

- *Work*: Photos belonging to this class are mostly characterized by people sitting and/or standing in workplaces and wearing work clothes and/or gear.



(a) Classes distribution

(b) Sample images

Fig. 3. IMAGO characteristics.

- *Free-time*: This class includes scenes of leisure time, reconstructing, wherever possible, generational and gender differences. It also includes images representing people visiting far-off landmarks, expanding social relationships and interacting with nature.
- *Motorization*: Although often closely related to the *Free-time* category, this class has been distinguished as it includes symbolic objects such as cars and motorcycles, which represent a social and historical landmark.
- *Music*: Similar to the *Motorization* one, this class may also include scenes from leisure time, characterized in this case by the appearance of musical instruments or events.
- *Fashion*: This class includes clothing, which represents a mirror of the articulated intertwining of socio-economic, political, and cultural phenomena. This class is characterized by the presence of symbolic objects and clothes, such as suits, trousers, skirts, and coats.
- *Affectivity*: This class is characterized by the presence of people (e.g., couples, friends, families, or colleagues) bound by inter-personal relationships.
- *Rites*: These are portraits of sacred and/or celebratory events from family lives.
- *School*: This class includes all the photos that represent schools, often characterized by symbolic objects (e.g., desk, blackboard) or groups of students.
- *Politics*: This class contains photos related to political gatherings, demonstrations, and events.

These aforementioned categories amount to the ones that from now on will be used to implement the socio-historical classification task.

5.3 Exploratory Dataset Analysis

In Figure 3(a) we show the number of labeled images available per year in the 1930-to-1999 time frame; out of such time interval, the number of available images is too little to be visually represented. This figure also exhibits the distribution of the socio-historical information (i.e., shooting year and socio-historical context) over the entire dataset. From such plot, the unbalance that exists in terms of number of photos both per year and socio-historical context is evident. Figure 3(b) shows four exemplar images from the IMAGO dataset, which belong to different decades and represent different socio-historical contexts. These images are representative of the different characteristics that may be found in each photo (e.g., number of people, clothing, colors, and location).

6 EXPERIMENTAL VALIDATION

We first provide details about the dataset pre-processing and the training process and then report the results obtained for both the socio-historical context classification and the dating tasks. The entire IMAGO dataset (the 16,642 labeled photos spanning the 1845–2009 time period) was used during the analysis of the socio-historical context classification task. For what concerns the image

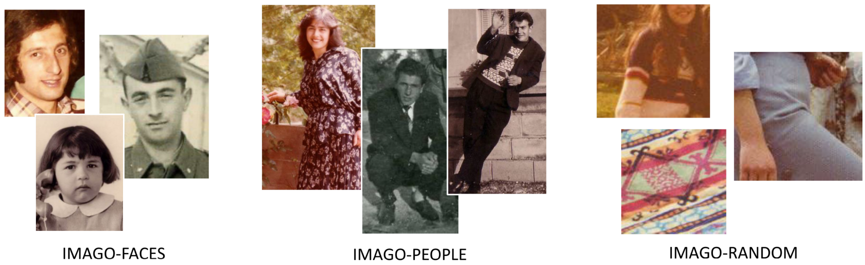


Fig. 4. Sample of different patches: (a) IMAGO-FACES, (b) IMAGO-PEOPLE, and, (c) IMAGO-RANDOM samples.

dating, 15,673 pictures, covering the 1930-to-1999 temporal interval, have been employed to avoid those years with a very limited number of samples, as already shown in Figure 3(a).

6.1 Dataset Pre-processing and Subdivision

The pre-processing phase aimed at (1) isolating the regions of interest from each photo and (2) improving the quality of the images composing the dataset, resorting to different techniques.

As reported in Section 4, both faces and people represent regions of interest to be exploited for the dating analysis [26, 47]. Following such insight, we created the IMAGO-FACES and the IMAGO-PEOPLE datasets, comprising over 60,000 samples each: the first composed of individual faces, the second of a single person's full-figure images. These have been obtained by processing each image of the IMAGO dataset using the open source implementations of YOLO-FACE and YOLO available at [30, 57], respectively. The IMAGO-FACES dataset has been constructed accounting for the number of people portrayed in a photo. In fact, adopting a fixed-size bounding box, it may be possible to lose relevant details (e.g., hairstyle) or to include pixels related to the faces of other people. To avoid such problem, an adaptive strategy has been adopted: the size of the bounding box used to crop a face depends on the number of people portrayed in a photo—the greater the number of people, the smaller the bounding box. In this way, it was possible to extract the shoulders and the full head of a single person even when a picture portrayed tens of people. Figure 4(a) shows some sample images taken from the IMAGO-FACES dataset considering different decades and different socio-historical contexts. The construction of the IMAGO-PEOPLE dataset follows the same criteria employed for IMAGO-FACES, though images can present different aspect ratios (i.e., people may be standing or sitting in photos). Figure 4(b) shows exemplar images from IMAGO-PEOPLE. It is possible to appreciate that IMAGO-PEOPLE includes details that are not present in IMAGO-FACES (e.g., the clothing of a person).

We then verified the utility of performing denoising and super-resolution operations, as all the images considered in this work derive from scans of the analog prints. For denoising we tested the neural network model from [65] and the Bilateral Filter [46]. For super-resolution, we used an open source implementation of the ESRGAN model [61] within the Image Restoration Toolbox [64]. The overall improvement obtained adopting such strategies was revealed to be negligible, so we hence opted for an analysis based on the original scans of analog photos.

The IMAGO-FACES and IMAGO-PEOPLE were defined only to fine-tune the deep learning models for the socio-historical tasks introduced with the IMAGO dataset. So, we will not release such datasets, since their creation is technology dependent. Indeed, in the future, algorithms or models providing more accurate bounding boxes for faces and people regions could be introduced.

Finally, to study the possible usefulness of non-human features within a family album photo dataset, we also created a dataset called IMAGO-RANDOM, comprising eight randomly cropped

Table 2. Accuracy for the Socio-historical Single-input Classifiers Considering the Top-k Predicted Classes (k Ranging from 1 to 5)

	Single-input Classifiers			
Top-k	**Full-image**	**Faces**	**People**	**Random-patches**
Top-1	**64.35**	41.30	56.54	37.35
Top-2	**85.00**	65.55	78.48	62.40
Top-3	**92.85**	82.75	89.90	80.31
Top-4	**96.66**	90.86	94.74	90.42
Top-5	**98.35**	94.98	97.42	95.35

regions, of 128 × 128 pixels, from each image in the IMAGO dataset (some samples are reported in Figure 4(c)). Other window sizes were also tested but returned a lower performance.

All these datasets have been partitioned as follows: 80% for training and 20% for testing; in addition, 10% of the training images are used as the validation set for hyper-parameter tuning. For each image in the train set of IMAGO, the faces and the people there portrayed and the random patches are extracted and added to the corresponding dataset subset. This process is repeated also for the validation and test sets, as it guarantees that none of the training samples may end up in the validation and test sets.

6.2 Model Architecture and Training Settings

All our CNN-based single-input classifiers adopt a well-known architecture pre-trained on ImageNet [21]: the ResNet50 [31]. This architecture was modified replacing the top-level classifier with a new classification layer, whose structure depends on the socio-historical task (i.e., the number of output classes) and whose weights have been randomly initialized. The pre-trained convolutional layers have been specifically fine-tuned for the given input data and task. In order to verify the independence of our dataset from the specific architecture, we have also considered two other well-known ones: InceptionV3 [55] and DenseNet121 [29]. However, the results were very similar and we decided to choose the ResNet50 as the main backbone for our analysis since it represents a good tradeoff between performance and number of parameters [20].

During the training phase we applied data augmentation (e.g., random crop and horizontal flip) in order to make the model less prone to overfitting. Each model has been fine-tuned using a weighted cross-entropy loss to counter the unbalance in our dataset [40]. The Adam optimizer has been employed with a learning rate of 1e-4 and a weight decay of 5e-4. We set the batch size to 32 for the training of the full-image classifier and to 64 for the faces, people, and random-patches models.

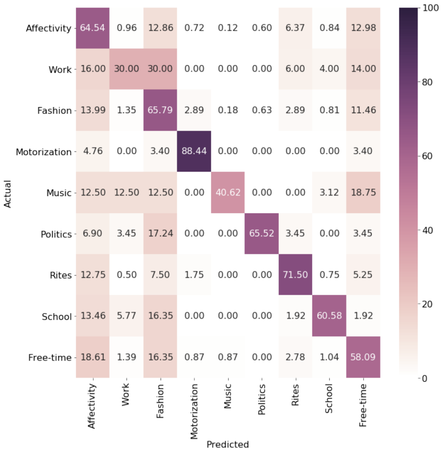
6.3 Socio-historical Context Classification Task Results

In the following sections, we proceed to report on the performance obtained with single-input classifiers and with the ensemble model. We finally provide a qualitative grad-cam-based analysis on the behavior of the models.

6.3.1 Single-input and Ensemble Classifiers. The results are reported in Table 2 and expressed in terms of top-k metric accuracy: if the correct class is not the one with the highest predicted

Table 3. Single-class Accuracy for Each Socio-historical Context Classifier

	Single-input Classifiers			
Socio-historical Context	**Full-image**	**Faces**	**People**	**Random-patches**
Affectivity	**64.54**	28.25	43.15	29.58
Work	**30.00**	24.00	22.00	29.00
Fashion	65.79	55.87	**67.60**	38.80
Motorization	**88.44**	17.01	51.02	29.66
Music	**40.62**	15.62	25.00	12.50
Politics	65.52	24.14	48.28	**66.67**
Rites	**71.50**	42.50	66.50	39.59
School	**60.58**	22.12	48.08	14.42
Free-time	58.09	46.09	**58.78**	51.94



Actual \ Predicted	Affectivity	Work	Fashion	Motorization	Music	Politics	Rites	School	Free-time
Affectivity	64.54	0.96	12.86	0.72	0.12	0.60	6.37	0.84	12.98
Work	16.00	30.00	30.00	0.00	0.00	0.00	6.00	4.00	14.00
Fashion	13.99	1.35	65.79	2.89	0.18	0.63	2.89	0.81	11.46
Motorization	4.76	0.00	3.40	88.44	0.00	0.00	0.00	0.00	3.40
Music	12.50	12.50	12.50	0.00	40.62	0.00	0.00	3.12	18.75
Politics	6.90	3.45	17.24	0.00	0.00	65.52	3.45	0.00	3.45
Rites	12.75	0.50	7.50	1.75	0.00	0.00	71.50	0.75	5.25
School	13.46	5.77	16.35	0.00	0.00	0.00	1.92	60.58	1.92
Free-time	18.61	1.39	16.35	0.87	0.87	0.00	2.78	1.04	58.09

Fig. 5. Confusion matrix for the full-image classifier.

probability but falls among the *k* with the highest predicted probabilities, it will be counted as correct. It is possible to appreciate that the full-image classifier exhibits a higher accuracy compared to the other single-input classifiers. To further investigate the reasons behind such result we report in Table 3 a comparison between the accuracy of each class considering the different single-input classifiers. As it is possible to observe, the model trained on IMAGO provides the best performance for the *Motorization*, *Rites*, *Music*, *School*, *Affectivity,* and *Work* classes. This may be due to the presence of specific objects that drive the performance of the model, also considering that the model was initialized with the ImageNet pre-trained weights [21], which contains classes such as race car and car wheel. Indeed, from a socio-historical point of view, images from the classes *Rites* and *Music* could contain physical objects and/or symbols that are representative for that class (e.g., formal attires, musical instruments). Nevertheless, such objects only acquire a meaning when people deal with them. However, the fact that the full-image classifier reached the highest accuracy for the *School*, *Affectivity,* and *Work* classes means that the network has also learned to recognize the presence of groups of people (e.g., school classes, friends standing in front of a monument, mother hugging her child) and specific clothing. Despite this classifier performing best, there are some peculiar results that have to be discussed. For example, the people classifier performs slightly better for the *Fashion* and *Free-Time* socio-historical contexts. This is probably due to the fact that the network may be focusing on people's clothing details and poses instead of exploiting specific objects and/or backgrounds that are not present in the people's crops. Exemplar areas on which the models focus in order to classify its images are reported in Section 6.3.2. Finally, the *Politics* class amounts to the only one for which, in terms of performance, the random-patches classifier is comparable to the full-image one.

We also evaluated different ensemble classifiers obtained from the combinations of the single-input classifiers. However, such combinations did not provide any significant improvement with respect to just considering the full-image model. For this reason, from now on, we consider the full-image classifier for the analysis that will follow and as the socio-historical context classifier in our application (check Figure 1).

Figure 5 shows the confusion matrix obtained with the full-image classifier. It is possible to observe that the classes responsible for the largest share of misclassifications are *Fashion*, *Affectivity,* and *Free-time*. This may be due to different causes. First, some classes share visual elements. For

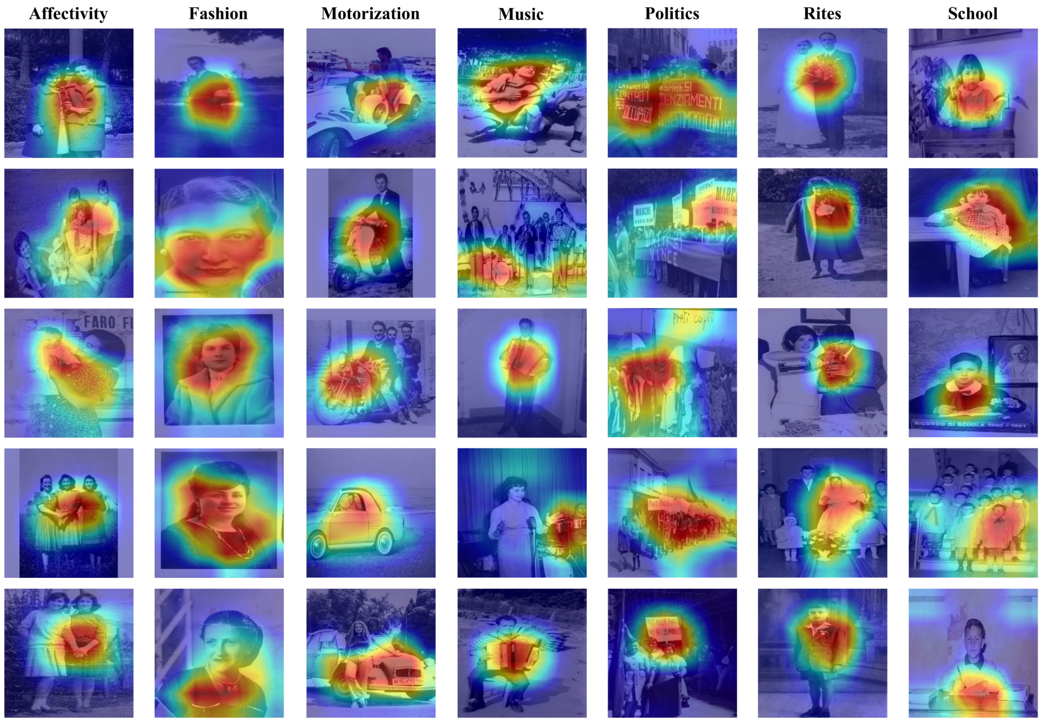


Fig. 6. Grad-Cam analysis of socio-historical contexts of pictures within IMAGO.

example, pictures labeled with *Work* class often depict people in uniform in workspaces. These could mistakenly be classified as belonging to the *Fashion* class, as pictures in this class are characterized by people in poses wearing some particular clothing items. Another example involves the *Music* and *Free-time* classes. Indeed, the *Music* category is characterized by photos portraying people playing some instruments or taking part in some musical event. The latter, however, could be easily associated to *Free-time* photos, since they also often portray groups of people in similar environments and poses. Second, the IMAGO dataset is unbalanced, as reported in Figure 3(a) (Section 5). Indeed, the most misclassified classes are also those that contain fewer samples.

6.3.2 Grad-Cam Analysis. We here report a qualitative analysis that aims at highlighting which visual cues led the classifier to associate a specific socio-historical context to a picture. To do so, we exploited the Grad-Cam algorithm [52], which delimits the areas driving the predictions performed by a deep learning model.

Figure 6 depicts samples of correctly classified IMAGO images processed by the Grad-Cam algorithm. Each column, starting from the left, shows five exemplary images belonging to the *Affectivity*, *Fashion*, *Motorization*, *Music*, *Politics*, *Rites,* and *School* classes, respectively. Such images are representative of the regions exploited by the full-image classifier. More in detail, people in certain poses close to each other (e.g., hugs, holding a baby, handshakes), as shown in the first column of Figure 6, are characteristic of the *Affectivity* class. Specific objects like earrings, necklaces, and lapels and also particular hairstyles are used to classify a picture as belonging to the *Fashion* class (second column of the figure). All kinds of vehicles, as well as musical instruments, are used to recognize a given picture as a member of the *Motorization* or the *Music* classes, shown in the third and fourth columns, respectively. The presence of a political banner is typical of pictures in the

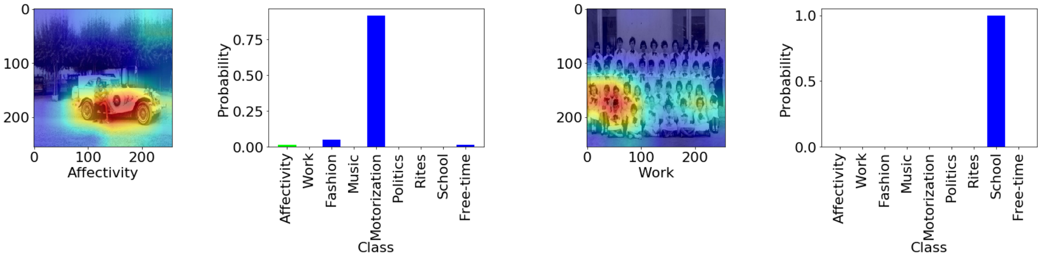


Fig. 7. Grad-Cam examples of failure cases: *Affectivity* recognized as *Motorization* and *Work* recognized as *School*.

Politics class (fifth column). The model also appears to individuate the objects that characterize the *Rites* class (e.g., white dress, flowers, pouring a drink, cheers), as shown in the sixth column of Figure 6. Finally, children wearing school uniforms, as well as school gear (e.g., books, pens, desks), are used to recognize pictures in the *School* class (last column).

It is not surprising that the model was able to correctly classify pictures belonging to the *Motorization* and *Music* classes, as these are clearly characterized by specific objects but, more importantly, already part of the model pre-trained on ImageNet [21]. However, also for the majority of the other classes (not studied so far in literature, to the best of our knowledge), the model seems to be able to isolate and focus on the details that distinguish them.

Figure 7 shows instead some failure cases for the full-image classifier. From the leftmost picture and its probability histogram it is possible to see that a photo containing a car was classified as belonging to the *Motorization* class but the ground-truth label assigned to the picture was *Affectivity* (two people standing close to each other in a specific pose). Instead, the rightmost picture and its corresponding probability histogram show that a picture depicting a school class was classified as belonging to *School*, while the actual one was *Work* (a teacher is standing in the rightmost part of the picture). Such misclassifications may be traced back to the fact that the IMAGO dataset was labeled by the owners of the pictures. The pictures thus convey such specific points of view, which may not be correctly predicted by the network. On the other hand, however, the point of view of the photo owner amounts to the ground truth, according to the methods adopted in socio-historical studies. In fact, the leftmost picture presented in Figure 7 was classified as *Affectivity* since the owner of the photograph was the child of the couple there portrayed. The same phenomenon happens in the rightmost one, since the one who labeled the photo was a teacher of those students. This proves the intrinsic challenge that the socio-historical classification task poses, since any classifier, including an expert socio-historian, may be subject to such kind of errors. For such reason, we further investigate such phenomenon in Section 7, analyzing the differences between the predictions obtained with the deep learning model and the choices made by a socio-historian.

6.4 Dating Task Results

The evaluation of performance for the dating is computed exploiting time distances, as also reported in [26, 47]. The time distance defines the tolerance accepted in prediction with respect to the actual year. For example, if a photo was labeled with the year 1932 and the model returned 1927 (or even 1937), this would be considered correct for those cases where the time distance was set to values equal to or smaller than 5 and wrong otherwise. In this work, model accuracies were computed considering temporal distances of 0, 5, and 10 years and have been assessed for both single-input and ensemble classifiers.

Table 4. Comparison of Single-input Classifiers Dating Performance

	Single-input Classifiers			
Time Distance	**Full-image**	**Faces**	**People**	**Random-patches**
d = 0	11.31	15.01	15.77	11.64
d = 5	62.56	58.09	62.40	54.26
d = 10	82.54	78.39	82.47	76.12

The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

Table 5. Single-input Classifiers Averaging Accuracies, along with Their Standard Deviation, Considering an Increasing Number of Patches and a Time Distance d = 0

	Single-input Classifiers		
# of Crops	**Faces**	**People**	**Random-patches**
1	11.70 (1.47)	11.74 (1.56)	6.35 (1.27)
2	12.88 (1.39)	14.32 (1.46)	6.97 (1.23)
3	13.46 (1.27)	15.09 (1.44)	8.01 (1.20)
4	13.87 (1.25)	15.47 (1.26)	8.15 (1.14)
5	14.19 (1.19)	15.71 (1.14)	8.16 (1.07)
6	14.40 (1.10)	15.89 (1.06)	8.42 (0.95)
7	14.58 (1.06)	16.07 (1.04)	8.47 (0.86)
8	14.82 (0.95)	15.93 (0.91)	9.00 (0.00)

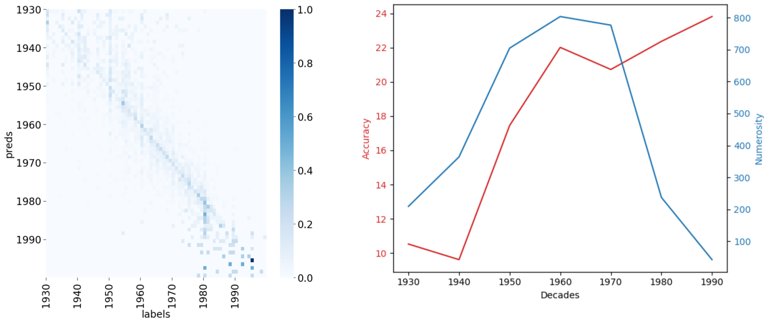
Table 6. Ensemble Model Considering Different Combinations of Full-image (T), Faces (F), and People (P) Classifiers

	Ensemble Classifiers			
Time Distance	**T + F**	**T + P**	**F + P**	**T + F + P**
d = 0	17.14	16.79	17.91	**18.51**
d = 5	66.51	66.44	64.02	**67.53**
d = 10	85.66	84.80	83.75	**86.17**

The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

The evaluation of the single-input classifiers is reported in Table 4. The models fine-tuned on faces and people regions achieved a higher accuracy compared to the full-image classifier, when considering a time distance equal to 0. This is also true for the random-patches classifier, which performed even worse with larger time distances. These results could be explained by model averaging, as the use of more data allows controlling uncertainty and reducing the prediction error rate [5]. Nevertheless, this increase in performance may also be due to the faces and people classifiers learning specific visual features characteristic of given time-slices. To verify whether such improvement was due to the averaging effect, we designed a specific experiment. We considered a test subset composed by all those images containing at least $n = 8$ faces or people crops (as in the case of random crops; see Section 6.1). To weigh the role of the number of faces/people, the accuracy values were computed considering k faces/people, with k growing from 1 to n. To ensure the completeness and fairness of this experiment, 1,000 random trials per each k faces/people/random patches were considered. Results have been grouped by k and reported in Table 5. From these results, we can observe that averaging across multiple inputs, in general, results in a higher performance, which increases when considering the faces and people regions.

Differently from the socio-historical context classification, an ensemble of different classifiers provides positive results for the dating task. Following the flow described in Figure 2, we proceeded to evaluate different ensemble combinations, exploiting the full-image, faces, people and random-patches classifiers. Since no significative improvements were observed employing the random-patches classifier, for the sake of clarity, Table 6 only includes the results that involve the full-image (T), faces (F), and people (P) classifiers. It is possible to observe that the best overall



(a) Confusion matrix for the dating task considering a time distance d = 0

(b) Model accuracy (red line) and number of samples (blue line) by decade for a time distance d = 0

Fig. 8. Dating task measures for the ensemble model.

performance is obtained with the ensemble combination of all these classifiers. This shows that the model may benefit from averaging across different classifiers, as well as across multiple regions [5]. This model performs better than any single-input classifier. From now on, we consider, for all the following experiments, the model that reached the best performance that is the ensemble of the full-image, faces, and people classifiers. Moreover, this ensemble has been adopted in our application to estimate the shooting year of a picture.

Figure 8(a) shows the confusion matrix considering a time distance equal to 0. The diagonal structure demonstrates that the confusion mostly occurs between neighboring years, except for the initial and the final decades (this has been observed also in other works, as in [26]). The confusion created within the first 20 years may be caused by the low quality of the images and the limited number of samples representing those years. The confusion created within the last 20 years, instead, may be related to the fact that the number of images for these years is very limited (as shown in Figure 3(a)). Nevertheless, it is interesting to observe the information provided in Figure 8(b), where the model accuracy and the number of samples per decade are reported. This figure confirms the finding exhibited by the confusion matrix; that is, the model accuracy improves after the 1950s. Figure 8(b) also shows that, despite a reduction in terms of available samples per decade after the 1980s, the performance of the model does not decrease. The accuracy generally improves after the 1950s (also when the number of samples drops); again, this could be related to the fact that the images are of better quality with respect to the previous decades.

Differently from the socio-historical task, we did not carry out a qualitative analysis for the dating task, as such type of analysis may already be found in the literature [26, 47].

7 HUMAN VS. MACHINE ASSESSMENT

To this point, we exploited the IMAGO dataset to train the models that compose the SHM, amounting to the core of the application designed to help socio-historians in cataloging family album pictures. To assess the performance the application could attain in terms of accuracy, with respect to a human expert, we designed a specific experiment where a socio-historian was asked to categorize all the pictures in the IMAGO test set (amounting to 3,327 pictures), providing both the socio-historical context and the date. In particular, on one hand, the SHM models can provide a ranking for the classes predicted for a specific photo (i.e., top-k for the socio-historical context classification and a time interval confidence for dating). On the other hand, the socio-historian deals

Table 7. Human vs. Machine: Accuracy Comparison for Increasing Values of k (k Indicates the Number of Selections Made by the Socio-historical Scholar and the Most Probable Classes Returned by the Model)

		Top-k Accuracy	
Cumulative k	**Cardinality**	**Socio-historical Context Module**	**Socio-historian**
1	2,147	64.88	54.82
1-2	3,278	72.02	66.53
1-2-3	3,327	72.24	66.93

Table 8. Human vs. Machine: Accuracy Reported for Different Time Distances (d = 0, d = 5, d = 10)

	Accuracy	
Time Distance	**Dating Module**	**Socio-historian**
d = 0	18.51	5.93
d = 5	67.53	56.36
d = 10	86.17	82.53

with the corpus of images, labeling them based on past archival and cataloging work experiences. In the following we provide the details of the comparison for the two tasks considered in this work.

Socio-historical context classification assessment. For this experiment, the socio-historical expert was given the opportunity of selecting multiple categories per each photo. As a result of this possibility, one class was chosen for 2,147 photos, two classes for 1,131 photos, and three classes for 49 pictures. It is interesting to point out that, although free to use as many labels as desired, no more than three have been considered at once. To make a fair comparison, we considered the k most probable classes chosen by the SHM model and compared them with the k classes selected by the socio-historian. We then proceeded to compute the accuracy of the socio-historian and of the model, as follows. For example, if the ground truth for a photo was "*Affectivity,*" the predictions provided by the application and the selections made by the socio-historian would be considered positive if both contained "*Affectivity.*" Since the scholar could choose the number of categories to assign, we computed such scores cumulatively. In particular, in correspondence of **Cumulative k**, in Table 7, with $k = 1$, a prediction is counted as positive in case it matches the ground truth. It follows that if $k > 1$, a positive match is recorded in case one of the k predictions matches the ground truth. The results are reported in Table 7. The first, simple observation is that the proposed application obtained accuracy levels that surpassed those obtained by the socio-historical scholar. For example, when we consider those pictures that were tagged with only one category by the socio-historical scholar, an accuracy of 54.82% was obtained vs. an accuracy of 64.89% for the application. This occurred also when considering those pictures for which the socio-historian chose one or more classes; the application was still able to obtain a higher performance. In Figure 9 we show a representative example of a case where the model predicts the correct label, unlike the socio-historian. In fact, the socio-historian fails at recognizing a particular detail that only the owner could have known (the subject of the photo is posing wearing a particular outfit); on the contrary, the model correctly classified this image.

Dating classification assessment. The socio-historian labeled all the pictures belonging to the test set assigning a year in the [1930,1999] range. The results are reported in Table 8. The dating module performed better than the socio-historian considering the specific picture shooting year (+12.58%). The difference in performance decreases when a higher time distance is considered, arriving to 3.64% when the time distance equals 10.

8 CNN VS TRANSFORMER PERFOMANCE

The Transformer is a deep learning architecture that relies entirely on the self-attention mechanism to draw global dependencies between input and output [60]. Recent works have shown that such an approach can achieve comparable or even superior performance to CNNs [23, 28, 58]. In particular, the **Vision Transformer (ViT)** architecture, proposed by Dosovitskiy et al. [23], has achieved state-of-the-art performance on several computer vision benchmarks. For these reasons,

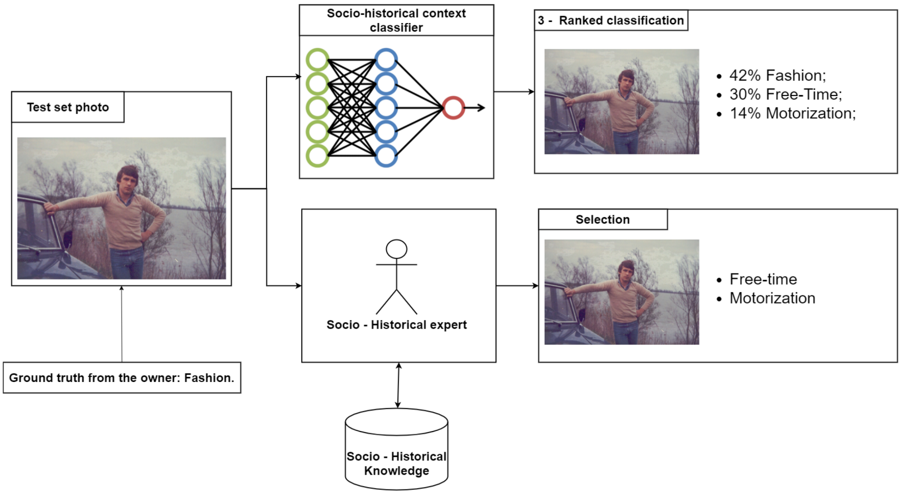

Socio-historical context classifier
3 - Ranked classification
• 42% Fashion;
• 30% Free-Time;
• 14% Motorization;
Test set photo
Selection
• Free-time
• Motorization
Socio - Historical expert
Ground truth from the owner: Fashion.
Socio - Historical Knowledge

Fig. 9. Human vs. machine: experiment diagram.

we decided to exploit the ViT architecture in this work (IMAGO dataset socio-historical context classification and dating). To this aim, we proceeded to fine-tune different ViT configurations (Tiny, Small, Base, and Large), varying the size of the input images (i.e., 224 × 224 or 384 × 384) and considering patches of 16 × 16 pixels. For the training, we followed the procedure reported in [23], while adopting a weighted cross-entropy loss to counter the dataset unbalance [40] and preserving the subdivision in training, validation, and test sets used in our previous experiments. This process was adopted for both the socio-historical context classification and dating for all of the proposed datasets (i.e., IMAGO, IMAGO-FACES, IMAGO-PEOPLE, and IMAGO-RANDOM). The results obtained with ViT, available in Tables 9 and 10, are there contrasted with those previously presented in Sections 6.3 and 6.4.

On one hand, from the results reported in Table 9, it is possible to observe that in most cases either ViT-Base or ViT-Large outperforms the ResNet50 while requiring a much higher number of parameters and thus increasing the complexity of the model. When instead a similar number of parameters is used (e.g., ViT-Small with input size 224 × 224), ViTs exhibit a slightly lower performance. Nevertheless, comparing the results shown in Table 11 and Figure 10 with those reported in Table 3 and Figure 5, it is worth noticing that ViT-Small obtains a more balanced per-class accuracy.

On the other hand, the results in Table 10 show that the ResNet50 outperforms all single-input ViT configurations for dating. We also considered different ensemble combinations but no relevant improvements were detected and for this reason are not here reported. Concluding, the ViT approach exhibits divergent behaviors when applied to dating and socio-historical context classification. Why this occurred may be explained by resorting to [42], where the authors highlighted how ViT (1) incorporates more global information than ResNet at lower layers, leading to different features, and (2) strongly preserves spatial information adopting class tokens. Indeed, the inclusion of more global information at lower layers and the strong preservation of spatial information could be the reason socio-historical context classification obtained a better accuracy than dating. This is qualitatively represented by a few GradCam examples reported in Figure 11: more accurate activations are obtained when compared to the corresponding examples for ResNet50, reported in Figure 6. On the contrary, dating often requires focusing on specific local visual cues rather than on global ones, as also highlighted by Ginosaur et al. [26].

Table 9. Comparison of Single-input Classifiers for Socio-historical Context Classification, Considering Both ResNet50 and ViT Models

Single-input Classifiers

	CNN	**Vision Tranformer**							
Architecture	**ResNet50**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**
#Params (K)	23,526	5,526	22,669	85,806	303,311	5,599	21,815	86,097	303,700
Input Dim	256	224				384			
Full-image									
Top-1	64.35	53.62	60.96	**66.24**	**67.87**	57.43	**65.13**	**68.53**	**69.19**
Top-5	98.35	96.63	97.72	**98.71**	**98.74**	97.14	97.84	**99.01**	**99.10**
Faces									
Top-1	41.30	35.58	41.23	**42.98**	**43.13**	35.61	37.21	40.64	39.43
Top-5	**94.98**	89.87	93.54	92.03	93.84	89.84	91.67	93.90	93.21
People									
Top-1	56.54	48.42	53.23	56.08	**59.21**	46.58	51.99	**60.35**	**62.51**
Top-5	97.42	93.78	96.15	97.32	**97.45**	93.36	95.85	**98.02**	**97.69**
random-patches									
Top-1	37.35	33.56	**40.29**	**44.09**	**43.78**	**39.06**	**38.08**	**43.72**	**44.34**
Top-5	**95.35**	86.22	93.20	93.05	95.28	92.74	91.49	92.74	93.57

The accuracy is reported considering the Top-1 and Top-5 predicted classes.

9 CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

In this work we proposed a multimedia application to assist socio-historians in cataloging family album photos. We then presented the IMAGO dataset composed by photos belonging to family albums, representing a source of socio-historical knowledge. The dataset amounts to 16,642 pictures, each of which was labeled with its socio-historical metadata: shooting year and context. We then trained and tested single-input and ensemble deep learning models to carry out those tasks, considering the Convolutional Neural Network. To the best of our knowledge, this is the first work addressing the socio-historical context classification. This consists in identifying the sociological and historical context of a picture, according to the definitions provided by socio-historical scholars [14]. We then proceeded to compare the performance of our application with the performance of a socio-historical scholar. The results of such assessment proved that our application could speed up cataloging processes, with no loss of accuracy when compared to the performance of a human expert, thus providing an important support to socio-historians. Finally, we carried out a comparative analysis considering Transformer-based deep learning models. The results showed that this approach could be promising also for a socio-historical analysis. This only represents a step in the direction of creating a holistic approach to the socio-historic cataloging problem, as many are the involved processes and sources of information. For example, in our specific case, our models were trained utilizing an unbalanced dataset and considering image regions that often included non-relevant information for classification purposes (e.g., background). In addition, when focusing on the socio-historical classification or dating, scholars perform analyses that resort at once to different sources of information (e.g., newspapers, magazines, archival documents), as well as to traces belonging to the same historical period. These represent three of the most relevant limits for this work.

Further investigations in this domain, hence, may consider (1) larger amounts and more balanced sets of data, (2) a better segmentation of the relevant areas of the images, and (3) the implementation of a multi-modal approach, capable of including also other sources of information and data formats. For what concerns the first point, the availability of larger datasets could surely

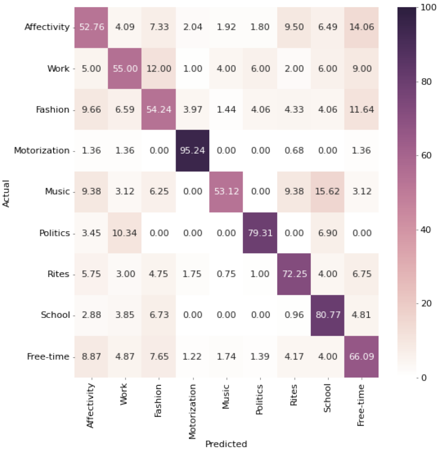
Table 10. Comparison of Single-input Classifiers for the Dating, Considering Both ResNet50 and ViT Models

Single-input Classifiers									
	CNN	**Vision Transformer**							
Architecture	**ResNet50**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**
#Params (K)	23,651	5,538	21,693	85,852	303,373	5,611	21,839	86,144	303,763
Input Dim	256	224				384			
Full-image									
d = 0	**11.31**	5.16	7.27	9.47	10.26	4.62	7.11	10.17	9.97
d = 5	**62.56**	38.85	43.40	50.72	53.44	35.21	46.37	54.11	55.74
d = 10	**82.54**	58.38	62.84	71.92	73.68	54.46	66.19	74.98	75.21
Faces									
d = 0	**15.01**	3.47	5.10	6.66	7.43	4.11	4.78	6.72	7.46
d = 5	**58.09**	31.39	39.42	46.46	46.59	34.58	38.34	45.73	49.24
d = 10	**78.39**	51.21	60.77	68.51	68.58	55.32	59.85	68.01	71.70
People									
d = 0	**15.77**	4.05	4.40	7.11	7.87	4.14	4.68	7.33	7.97
d = 5	**62.40**	33.65	34.51	46.88	48.69	32.22	38.91	46.18	49.17
d = 10	**82.47**	54.21	51.56	68.90	70.24	52.42	59.40	67.81	70.04
Random-patches									
d = 0	**11.64**	4.08	5.00	7.08	7.43	4.21	5.00	7.20	7.49
d = 5	**54.26**	34.08	36.05	41.82	43.39	32.83	34.11	42.17	41.89
d = 10	**76.12**	56.81	57.09	64.81	66.59	51.51	54.22	64.81	64.74

The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

Table 11. Single-class Accuracy for Each Socio-historical Context Classifier Based on ViT-Small

	Single-input Classifiers			
Socio-historical Context	**Full-image**	**Faces**	**People**	**Random-patches**
Affectivity	**52.76**	30.41	32.21	11.00
Work	**55.00**	4.00	26.00	8.00
Fashion	54.24	55.32	**58.75**	51.85
Motorization	**95.24**	37.41	93.20	61.38
Music	53.12	12.50	**56.25**	9.38
Politics	**79.31**	51.72	58.62	59.26
Rites	**72.25**	44.75	62.75	40.61
School	**80.77**	49.04	63.46	58.65
Free-time	**66.09**	34.43	58.61	58.13



Actual \ Predicted	Affectivity	Work	Fashion	Motorization	Music	Politics	Rites	School	Free-time
Affectivity	52.76	4.09	7.33	2.04	1.92	1.80	9.50	6.49	14.06
Work	5.00	55.00	12.00	1.00	4.00	6.00	2.00	6.00	9.00
Fashion	9.66	6.59	54.24	3.97	1.44	4.06	4.33	4.06	11.64
Motorization	1.36	1.36	0.00	95.24	0.00	0.00	0.68	0.00	1.36
Music	9.38	3.12	6.25	0.00	53.12	0.00	9.38	15.62	3.12
Politics	3.45	10.34	0.00	0.00	0.00	79.31	0.00	6.90	0.00
Rites	5.75	3.00	4.75	1.75	0.75	1.00	72.25	4.00	6.75
School	2.88	3.85	6.73	0.00	0.00	0.00	0.96	80.77	4.81
Free-time	8.87	4.87	7.65	1.22	1.74	1.39	4.17	4.00	66.09

Fig. 10. Confusion matrix for the ViT-Small full-image classifier.

improve the models' discriminative power, also reducing possible unbalance problems. Regarding the second, the use of segmentation models may benefit the individuation of more relevant regions. Finally, multi-modal learning appears as the approach that may best replicate the comprehensive approach normally adopted by socio-historians during cataloging processes. Indeed, exploiting knowledge from historical archival documents (and other sources) could improve the general cataloging and analysis effort. For example, knowing how people dressed during a specific

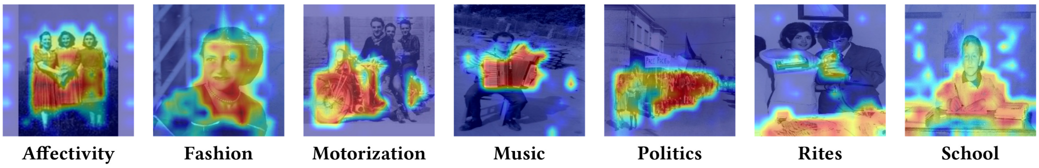


Fig. 11. Grad-Cam analysis of socio-historical contexts using ViT-Small.

period might improve the classification for both the socio-historical context and dating. Such path, although complex, may not be impossible to follow. Indeed, recent natural language processing solutions are able to provide discriminative features that could be exploited in our models to improve the overall performance [22].

REFERENCES

[1] Amazon. 2021. Amazon SageMaker Ground Truth. https://aws.amazon.com/it/sagemaker/groundtruth/.

[2] S. Barba, F. Fiorillo, P. Ortiz Coder, S. D'auria, and E. De Feo. 2011. An application for cultural heritage in Erasmus Placement. Surveys and 3D cataloguing archaeological finds in Merida (Spain). (2011).

[3] B. Fernando, D. Muselet, R. Khan, and T. Tuytelaars. 2014. Color features for dating historical color images. In *IEEE International Conference on Image Processing (ICIP'14)*. 2589–2593. DOI: 10.1109/ICIP.2014.7025524

[4] K. Bentein. 2015. Minor complementation patterns in Post-classical Greek (I–VI AD): A socio-historical analysis of a corpus of documentary papyri. *Symbolae Osloenses* 89, 1 (2015), 104–147.

[5] C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

[6] E. Borcoci, D. Negru, and C. Timmerer. 2010. A novel architecture for multimedia distribution based on content-aware networking. In *2010 3rd International Conference on Communication Theory, Reliability, and Quality of Service*. IEEE, 162–168.

[7] L. Bosi and H. Reiter. 2014. Historical methodologies. *Methodological Practices in Social Movement Research* (2014), 117–143.

[8] P. Bourdieu. 1996. On the family as a realized category. *Theory, Culture & Society* 13, 3 (1996), 19–26. DOI: 10.1177/026327696013003002

[9] M. A. Cabrera. 2001. On language, culture, and social action. *History and Theory* 40, 4 (2001), 82–100.

[10] M. Á. Cabrera. 2004. *Postsocial History: An Introduction*. Lexington Books.

[11] D. Calanca. 2004. Percorsi di storia della famiglia. *Rivista Di Storia E Storiografia* 5, 5 (Nov. 2004), 203–210.

[12] D. Calanca. 2005. Album di famiglia. Autorappresentazioni tra pubblico e privato (1870-1950). *Storia e Futuro* 8–9 (2005).

[13] D. Calanca. 2006. Fotografie amatoriali e fotografie professionali nell'Italia del boom economico. *Storia e Futuro* 12 (2006), 134–144.

[14] D. Calanca. 2011. Italians posing between public and private. Theories and practices of Social Heritage. *Almatourism-Journal of Tourism, Culture and Territorial Development* 2, 3 (2011), 1–9.

[15] J. Chen and H. Wang. 2018. Guest editorial: Big data infrastructure I. *IEEE Transactions on Big Data* 4, 2 (2018), 148–149. https://doi.org/10.1109/TBDATA.2018.2839919

[16] X. Chen, D. Liu, Z. Xiong, and Z-J. Zha. 2020. Learning and fusing multiple user interest representations for micro-video and movie recommendations. *IEEE Transactions on Multimedia* (2020).

[17] E. S. Clemens and M. D. Hughes. 2002. Recovering past protest: Historical research on social movements. In *Methods of Social Movement Research. Minneapolis*, B. Klandermans and S. Staggenborg (Eds.). University of Minnesota Press, 201–230.

[18] E. Coburn, E. Lanzi, E. O'Keefe, R. Stein, and A. Whiteside. 2010. The cataloging cultural objects experience: Codifying practice for the cultural heritage community. *IFLA Journal* 36, 1 (2010), 16–29.

[19] L. Criscenti, G. D'autilia, and G. De Luna. 2005. *L'Italia Del Novecento: Le Fotografie e la Storia*. Giulio Einaudi editore.

[20] E. Culurciello. 2021. Neural Network Architectures. https://towardsdatascience.com/neural-network-architectures-156e5bad51ba.

[21] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly J. Uszkoreit, and N. Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[24] K. J. Enns and M. J. Martin. 2015. Gendering agricultural education: A study of historical pictures of women in the agricultural education magazine. *Journal of Agricultural Education* 56, 3 (2015), 69–89.

[25] R. Franzosi. 1998. Narrative as data: Linguistic and statistical tools for the quantitative study of historical events. *International Review of Social History* 43, S6 (1998), 81–104.

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

[27] Google. 2021. AI Platform Data Labeling Service. https://cloud.google.com/ai-platform/data-labeling/docs.

[28] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao. 2022. A Survey on Vision Transformer. In *IEEE Transactions on Pattern Analysis and Machine Intelligence.* DOI: 10.1109/TPAMI.2022.3152247

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV]

[30] J. Redmon. 2019. YOLO: Real Time Object Detection. Retrieved August 3, 2020, from https://github.com/pjreddie/darknet/wiki/YOLO:-Real-Time-Object-Detection.

[31] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[32] J. Lemley, S. Bazrafkan, and P. Corcoran. 2017. Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision. *IEEE Consumer Electronics Magazine* 6, 2 (2017), 48–56.

[33] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei. 2019. Deep metric learning with density adaptivity. *IEEE Transactions on Multimedia* 22, 5 (2019), 1285–1297.

[34] M. Lincoln, J. Corrin, E. Davis, and S. B. Weingart. 2020. CAMPI: Computer-aided metadata generation for photo archives initiative. Carnegie Mellon University. Preprint. https://doi.org/10.1184/R1/12791807.v2.

[35] G. Mitman and K. Wilder. 2016. *Documenting the World: Film, Photography, and the Scientific Record.* University of Chicago Press Chicago, IL.

[36] MoMA. 2020. Vernacular Photography. https://www.moma.org/collection/terms/vernacular-photography.

[37] E. Müller, M. Springstein, and R. Ewerth. 2017. "When was this picture taken?"—Image date estimation in the wild. In *European Conference on Information Retrieval.* Springer, 619–625.

[38] F. Palermo, J. Hays, and A. A. Efros. 2012. Dating historical color images. In *European Conference on Computer Vision.* Springer, 499–512.

[39] M. R. Peres. 2014. *The Concise Focal Encyclopedia of Photography: From the First Photo on Paper to the Digital Revolution.* CRC Press.

[40] T. H. Phan and K. Yamamoto. 2020. Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses. arXiv:2006.01413 [cs.CV]

[41] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga. 2014. Ensemble deep learning for regression and time series forecasting. In *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL'14).* 1–6. https://doi.org/10.1109/CIEL.2014.7015739

[42] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? In *35th Conference on Neural Information Processing Systems.*

[43] B. Rainer, S. Petscharnig, C. Timmerer, and H. Hellwagner. 2016. Statistically indifferent quality variation: An approach for reducing multimedia distribution cost for adaptive video streaming services. *IEEE Transactions on Multimedia* 19, 4 (2016), 849–860.

[44] M. Roccetti, L. Casini, G. Delnevo, V. Orrù, N. Marchetti, and Nicolò. 2020. Potential and limitations of designing a deep learning model for discovering new archaeological sites: A case with the mesopotamian floodplain. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good.* 216–221.

[45] D. Rosner, M. Roccetti, and G. Marfia. 2014. The digitization of cultural practices. *Communications of the ACM* 57, 6 (June 2014), 82–87. https://doi.org/10.1145/2602695.2602701

[46] J. Tumblin, S. Paris, P. Kornprobst, and F. Durand. 2007. A gentle introduction to bilateral filtering and its applications. In *ACM SIGGRAPH 2007 Courses (SIGGRAPH'07).* Association for Computing Machinery, New York, NY, 1–es. https://doi.org/10.1145/1281500.1281602

[47] T. Salem, S. Workman, M. Zhai, and N. Jacobs. 2016. Analyzing human appearance as a cue for dating images. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV'16).* IEEE, 1–8.

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

[49] C. Schreiber. 2014. The construction of “female citizens”: a socio-historical analysis of girls’ education in Luxembourg. *Educational Research* 56, 2 (2014), 137–154.

[50] J. Scott and G. Marshall. 2009. *A Dictionary of Sociology*. Oxford University Press.

[51] T. J. Sejnowski. 2018. *The Deep Learning Revolution*. MIT Press.

[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[53] E. Serafinelli. 2018. *Digital Life on Instagram: New Social Communication of Photography*. Emerald Group Publishing.

[54] P. Sorcinelli. 2004. Imago. Laboratorio di ricerca storica e di documentazione iconografica sulla condizione giovanile nel XX secolo. *Rivista Di Storia E Storiografia* 5, 5 (Nov. 2004), 200–202.

[55] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2015. Rethinking the Inception architecture for computer vision. arXiv:1512.00567 [cs.CV]

[56] L. Tilton and T. Arnold. 2019. Distant viewing: Analyzing large visual corpora. *Digital Scholarship in the Humanities* 34, Supplement 1 (2019), i3–i16.

[57] T. Nguyen. 2018. Yolo Face Implementation. Retrieved August 3, 2020, from https://github.com/sthanhng/yoloface.

[58] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[59] F. Vaccaro, M. Bertini, T. Uricchio, and A. Del Bimbo. 2020. Image retrieval using multi-scale CNN features pooling. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 311–315.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, and Ł. Kaiser. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[61] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. arXiv:1809.00219 [cs.CV]

[62] M. Wevers and T. Smits. 2020. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities* 35, 1 (2020), 194–207.

[63] W. Yin, T. Mei, C. W. Chen, and S. Li. 2013. Socialized mobile photography: Learning to photograph with social context via mobile devices. *IEEE Transactions on Multimedia* 16, 1 (2013), 184–200.

[64] K. Zhang. 2019. Image Restoration Toolbox. https://github.com/cszn/KAIR.

[65] K. Zhang, W. Zuo, and L. Zhang. 2018. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing* 27, 9 (2018), 4608–4622. DOI: 10.1109/TIP.2018.2839891

[66] W. Zhang, T. Yao, S. Zhu, and A. E. Saddik. 2019. Deep learning–based multimedia analytics: A review. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1s, Article 2 (Jan. 2019), 26 pages. https://doi.org/10.1145/3279952

Received 3 September 2021; revised 20 December 2021; accepted 21 December 2021