Annual Report, No. 475648, December 2022
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

AIMH Research Activities 2022

Nicola Aloia, Giuseppe Amato, Valentina Bartalesi, Filippo Benedetti, Paolo Bolettieri, Donato Cafarelli, Fabio Carrara, Vittore Casarosa, Luca Ciampi, Davide Alessandro Coccomini, Cesare Concordia, Silvia Corbara, Marco Di Benedetto, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, Gabriele Lagani, Emanuele Lenzi, Carlo Meghini, Nicola Messina, Daniele Metilli, Alessio Molinari, Alejandro Moreo, Alessandro Nardi, Andrea Pedrotti, Nicolò Pratelli, Fausto Rabitti, Pasquale Savino, Fabrizio Sebastiani, Gianluca Sperduti, Costantino Thanos, Luca Trupiano, Lucia Vadicamo, Claudio Vairo

Abstract
The Artificial Intelligence for Media and Humanities laboratory (AIMH) has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability. This report summarize the 2022 activities of the research group.

Keywords
Multimedia Information Retrieval – Artificial Intelligence — Computer Vision — Similarity Search – Machine Learning – Text Classification – Deep Learning – Transfer learning – Representation Learning

[1] *AIMH Lab, ISTI-CNR, via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy*
***Corresponding author**: giuseppe.amato@isti.cnr.it

Contents

Introduction **2**
1 **Research Topics** **2**
1.1 Artificial Intelligence 2
1.2 AI and Digital Humanities 2
1.3 AI for Text 3
1.4 Computer Vision 4
1.5 Multimedia Information Retrieval 6
1.6 Medical Imaging 7
1.7 Quantum Machine Learning 8
1.8 Fighting misinformation 8
2 **Projects & Activities** **9**
2.1 EU Projects 9
2.2 CNR National Virtual Lab on AI 11
2.3 National Projects 11
3 **Papers** **12**
3.1 Journals 12
3.2 Proceedings 18
3.3 Magazines 25
3.4 Editorials 26
3.5 Preprints 26
4 **Dissertations** **27**
4.1 PhD Thesis 27
4.2 Master of Science Dissertations 28
5 **Resources** **30**
5.1 Datasets 30
5.2 Code 32
5.3 Shared Tasks 34
6 **Services** **35**
6.1 Services in conferences 35
7 **Awards** **35**
7.1 ERCIM Cor Baayen Young Researcher Award . 35
7.2 Video Browser Showdown 35
7.3 ISTI Young Research Awards 35
References **35**

AIMH
Artificial Intelligence for
Media and Humanities

http://aimh.isti.cnr.it

Introduction

The Artificial Intelligence for Media and Humanities laboratory (AIMH) of the Information Science and Technologies Institute “Alessandro Faedo” (ISTI) of the Italian National Research Council (CNR) located in Pisa, has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability.

The laboratory is composed of four research groups:

AI4Text

The AI4Text group is active in the area at the crossroads of machine learning and text analysis; it investigates novel algorithms and methodologies, and novel applications of these to different realms of text analysis. Topics within the above-mentioned area that are actively researched within the group include representation learning for text classification, transfer learning for cross-lingual and cross-domain text classification, sentiment classification, sequence learning for information extraction, learning to quantify, transductive text classification, cost-sensitive text classification, and applications of the above to domains such as authorship analysis, technology-assisted review, and native language identification. The group consists of Fabrizio Sebastiani (Director of Research), Andrea Esuli (Senior Researcher), Alejandro Moreo (Researcher), Silvia Corbara, Alessio Molinari, Andrea Pedrotti, and Gianluca Sperduti (PhD Students), and is led by Fabrizio Sebastiani.

Humanities

Investigating AI-based solutions to represent, access, archive, and manage tangible and intangible cultural heritage data. This includes solutions based on ontologies, with a special focus on narratives, and solutions based on multimedia content analysis, recognition, and retrieval. The group consists of Carlo Meghini (Director of Research), Valentina Bartalesi, Cesare Concordia (Researchers), Luca Trupiano (Technologist), Daniele Metilli (PhD Student), Emanuele Lenzi, Nicolò Pratelli (Graduate Fellows), and Costantino Thanos, Vittore Casarosa, Nicola Aloia (Research Associates), and is led by Carlo Meghini.

Large-scale IR

Investigating efficient, effective, and scalable AI-based solutions for searching multimedia content in large datasets of non-annotated data. This includes techniques for multimedia content extraction and representation, scalable access methods for similarity search, multimedia database management. The group consists of Claudio Gennaro, Pasquale Savino (Senior Researchers), Lucia Vadicamo (Researcher), Nicola Messina, Claudio Vairo (Researchers), Paolo Bolettieri (Technician), Davide Alessandro Coccomini (PhD Students), Gabriele Lagani (PhD Students), Fausto Rabitti (Research Associate), and is led by Claudio Gennaro.

Vision and Deep Learning

Investigating novel AI-based solutions to image and video content analysis, understanding, and classification. This includes techniques for detection, recognition (object, pedestrian, face, etc), classification, counting, feature extraction (low- and high-level, relational, cross-media, etc), anomaly detection also considering adversarial machine learning threats. We also have specific AI research fields such as hebbian learning and relational learning. The group consists of Giuseppe Amato (Director of Research), Fabrizio Falchi (Senior Researcher), Marco Di Benedetto, Fabio Carrara (Researchers), Alessandro Nardi (Technician), Davide Alessandro Coccomini (PhD Students), Gabriele Lagani (PhD Students), Donato Cafarelli, Luca Ciampi (Graduate Fellow), and is led by Fabrizio Falchi.

The rest of the report is organized as follows. In Section 1, we summarize the research conducted on our main research fields. In Section 2, we describe the projects in which we were involved during the year. We report the complete list of papers we published in 2022, together with their abstract, in Section 3. The list of theses on which we were involved can be found in Section 4. In Section 5.2 we highlight the datasets we created and made publicly available during 2022.

1. Research Topics

In the following, we report a list of active research topics and subtopics at AIMH in 2021.

1.1 Artificial Intelligence

1.1.1 Hebbian Learning

Traditional neural networks are trained using gradient descent methods with error backpropagation. Despite the great success of such training algorithms, the neuroscientific community has doubts about the biological plausibility of backpropagation learning schemes, proposing a different learning model known as *Hebbian principle*: "Neurons that fire together wire together". Starting from this simple principle, different Hebbian learning variants have been formulated. These approaches are interesting also from a computer science point of view, because they allow to perform common data analysis operations - such as clustering, Principal Component Analysis (PCA), Independent Component Analysis (ICA), and others - in an online, efficient, and neurally plausible fashion. Taking inspiration from biology, we investigate how Hebbian approaches can be integrated with today’s machine learning techniques [58, 59], in order to improve the training process in terms of speed, generalization capabilities, sample efficiency [57].

1.2 AI and Digital Humanities

The AI & DH group at AIMH employs AI-based methods to research, design and experimentally develop innovative tools to support the work of the scholar humanist. These methods hinge on formal ontologies as powerful tools for the design and the implementation of information systems that exhibit intelligent behavior. Formal ontologies are also regarded as the ideal place where computer scientists and humanists can meet and collaborate to co-create innovative applications that can effectively support the work of the latter. The group pursues in particular the notion of formal narrative as a powerful addition to the information space of digital libraries; an ontology for formal narratives has been developed in the last few years and it is currently being enriched through the research carried out by the members of the group and tested through the validation carried out in the context of the Mingei project. The group is also engaged in the formal representation of literary texts and of the surrounding knowledge, through the IMAGO project which is focused on the geographical medieval manuscripts and the HDN project which continues the seminal work that led to the DanteSources application where an ontology-based approach was firstly employed. Finally, through the participation to the ARIADNEplus and SSHOC projects, the group is actively involved in shaping the European landscape on research Infrastructures in Archaeology and Social Sciences, respectively. And through the participation to the MINGEI project, the group is actively involved in experimenting with narratives for the representation and reservation of intangible craft heritage.

1.3 AI for Text

1.3.1 Learning to quantify

Learning to quantify has to do with training a predictor that estimates the prevalence values of the classes of interest in a sample of unlabelled data. This problem has particular relevance in scenarios characterized by distribution shift (which may itself be caused by either covariate shift or prior probability shift), since standard learning algorithms for training classifiers are based on the IID assumption, which is violated in scenarios characterized by distribution shift. The AI4Text group has carried out active research on learning to quantify since 2010.

One of our recent activities include our study [76] in which we have looked back at past research on sentiment quantification, and found that the different approaches to such a task have been compared inappropriately, due to a faulty experimental protocol. We have thus carried out a complete reassessment of these approaches, this time using a much more robust protocol which involves a much more extensive experimentation; also these results have upturned past conclusions concerning the relative merits of such approaches.

In a different effort [12], we have devised new methods for *ordinal* quantification, i.e., for the multiclass setting in which a total order is defined among the classes of interest. In this study, we first analyze the main drawbacks of currently available datasets for this task and then propose new, more adequate ones. Using these datasets, we experimentaly compare most important existing algorithms for this task, bringing together methods arising from different disciplines, as well as new methods that we propose in our study. Our newly proposed methods, that tackle the ordinal aspect of the problem via a regularization penalty, proved superior in the experimental evaluation we have carried out.

In a different study [74], we looked at a different setting of multiclass problems: the *multi-label* setting. Surprinsingly enough, this multiclass scenario in which the classes of interest are not mutually exclusive, has remained mostly unexplored in the literature. In our study, we first offer a critical overview of the few existing related work. We argued these methods are naive, since they made no attempt to leverage (i.e., to learn from) the stochastic dependencies among the classes. Based on this observation, we propose the first truly multi-label quantification methods, i.e., methods that, in order to predict the relative frequencies of the classes, take into account the correlations amongst the classes as inferred from the training data. Our experiments prove that our methods outperform, by a large margin, previously existing ones.

Three further activities in which we have engaged are:

- the organization of the 2nd International Workshop on Learning to Quantify (LQ 2022)[1] [39], which has taken place in September 2022 as an hybrid event in Grenoble (France);
- the organization of LeQua 2022[2] [44, 45, 46], the first shared task entirely devoted to learning to quantify, which has been organized under the umbrella of the CLEF 2022 conference [3]

1.3.2 Learning to classify text

The supervised approach to text classification (TC) is almost 30 years old; despite this, text classification continues to be an active research topic, due to its central role in a number of text analysis and text management tasks.

One problem we have worked on is *cross-lingual TC*, i.e., the task of leveraging training data for a “source” language in order to perform TC in a different, “target” language for which we have little or no training data. In [75] we have extended a previously proposed method for heterogeneous transfer learning (called “Funnelling”) to leverage correlations in data that are informative for the TC process; while Funnelling exploit class-class correlations, our “Generalized Funnelling" system also exploits word-class correlations, word-word correlations (for which we employ MUSE embeddings), and correlations between words-in-context, obtained via Multilingual BERT.

We are also working on the definition of user interfaces and machine learning algorithms that support the activity of classification of documents by humans. In this context, we have developed the Interactive Classification System [43], a first implementation of the concept of *unobtrusive learning*, i.e., a machine learning paradigm in which the learning algorithm observes the actions of the human annotators, never

[1] https://lq-2022.github.io/

[2] https://lequa2022.github.io/

[3] https://clef2022.clef-initiative.eu/

interrupts them in any way, transparently updates the automatic classification models, which are always available to make predictions on new data.

1.3.3 Technology-assisted review

Technology-assisted review (TAR) is the task of supporting the work of human annotators who need to “review” automatically labelled data items, i.e., check the correctness of the labels assigned to these items by automatic classifiers. Since only a subset of such items can be feasibly reviewed, the goal of these algorithms is to exactly identify the items whose review is expected to be cost-effective. We have been working on this task since 2018, proposing TAR *risk minimization* algorithms that attempt to strike an optimal tradeoff between the contrasting goals of minimizing the cost of human intervention and maximizing the accuracy of the resulting labelled data. An aspect of TAR we have worked on more recently is improving the quality of the posterior probabilities that the risk minimization algorithm receives as input by an automated classifier. To this end, we have continued the study of the SLD algorithm, exploring how it interacts with the Active Learning (AL) protocols that can be used in TAR. Our study [73] found that SLD fail to improve the posteriors when it is applied to classifiers produced by means of AL procedures. We are currently investigating variations of SLD to make work with AL-generated classifiers.

1.3.4 Authorship analysis

Authorship analysis has to do with training predictors that infer characteristics of the author of a document of unknown paternity. We have worked on a sub-problem of authorship analysis called *authorship verification*, which consists of training a binary classifier that decides whether a text of disputed paternity is by a candidate author or not. Specifically, we have concentrated on a renowned case study, the so-called *Epistle to Cangrande*, written in medieval Latin apparently by Dante Alighieri, but whose authenticity has been disputed by scholars in the last century. To this end, we have built and made available to the scientific community two datasets of Medieval Latin texts, which we have used for training two separate predictors, one for the first part of the *Epistle* (which has a dedicatory nature) and one for the second part (which is instead a literary essay). The authorship verifiers that we have built indicate, although with different degrees of certainty, that neither the first nor the second part of the *Epistle* are by Dante. These predictions are corroborated by the fact that, once tested according to a leave-one-out experimental protocol on the two datasets, the two predictors exhibit extremely high accuracy [38].

An additional research we have carried out concerning authorship analysis for prose texts written in the Latin language, is the study of how features derived from “syllabic quantity” can impact the accuracy of predictors [36, 37]. Syllabic quantity is an attribute of words, and it is well-known how different Latin authors used different syllabic quantity patterns in their writings. Our study has determined that extracting syllabic quantity from Latin prose texts is beneficial for authorship attribution.

Based on this evidence, we have further investigated the extent to which rhythmic features can be similarly beneficial for other Latin-derived languages, like Spanish. In particular, we have observed that a combination of rhythmic patterns extracted from the analysis of accentuation, along with pyscholinguistic features [34] and other topic-agnostic features [35], can indeed be of help for analyzing authorial traits of political speeches.

1.3.5 Native language identification

Native language identification (NLI) is the task of training (via supervised machine learning) a classifier that guesses the native language of the author of a text written in a different language. This task has been extensively researched in the last decade, and the performance of NLI systems has steadily improved over the years. While working on the NLI task [9], we focused on analysing the internals of an NLI classifier trained by an *explainable* machine learning algorithm, in order to obtain explanations of its classification decisions, with the ultimate goal of gaining insight into which linguistic phenomena “give a speaker’s native language away”. We used this perspective in order to tackle both NLI and a (much less researched) companion task, i.e., guessing whether a text has been written by a native or a non-native speaker.

1.4 Computer Vision

1.4.1 Visual Counting

The counting task aims to estimate the number of objects instances, like people or vehicles, in still images or videos. Current solutions are formulated as supervised deep learning-based problems belonging to one of two main categories: counting by *detection* and counting by *regression*. Detection-based approaches require prior detection of single instances of objects. On the other hand, regression-based techniques try to establish a direct mapping between the image features and the number of objects in the scene, either directly or via the estimation of a target map, such as a density map (i.e., a continuous-valued), and it is more effective in highly-occluded scenarios.

In [27], we propose a novel solution to improve car counting in parking lots when scaled up with multi-camera setups. We introduce a multi-camera system that combines a CNN-based technique, which can locate and count vehicles present in images belonging to individual cameras, along with a decentralized geometry-based approach that is responsible for aggregating the data gathered from all the devices and estimating the number of cars present in the *entire* parking lot. A remarkable peculiarity of our solution is that it performs the task directly on the *edge* devices, i.e., the smart cameras — vision systems with limited computational capabilities able to capture images, extract information from them, make decisions, and communicate with other devices.

On the other hand, in [3], we proposed a video-based counting technique that estimates the number of people present in clips gathered from surveillance cameras by taking advantage of the temporal correlation between consecutive frames.

Recently, we also tackled the task of counting cells in microscopy images [23, 25]. We describe this activity in more detail in Section 1.6.1.

1.4.2 Human Activity Monitoring

As evidenced during the recent COVID-19 pandemic, there are scenarios in which ensuring compliance with a set of guidelines (such as wearing medical masks and keeping a certain physical distance among people) becomes crucial to secure a safe living environment. However, human supervision could not always guarantee this task, especially in crowded scenes. In [40], we presented an embedded modular Computer Vision-based and AI-assisted system that can carry out several tasks to help monitor individual and collective human safety rules. We strove for a real-time but low-cost system, thus complying with the compute- and storage-limited resources availability typical of off-the-shelves embedded devices, where images are captured and processed directly onboard. Our solution consists of multiple modules relying on well-researched neural network components, each responsible for specific functionalities that the user can easily enable and configure. In particular, by exploiting one of these modules or combining some of them, our framework makes available many capabilities. They range from the ability to estimate the so-called social distance to the estimation of the number of people present in the monitored scene, as well as the possibility to localize and classify Personal Protective Equipment (PPE) worn by people (such as helmets and face masks). To validate our solution, we tested all the functionalities that our framework makes available over two novel datasets that we collected and annotated on purpose. Experiments showed that our system provides a valuable asset to monitor compliance with safety rules automatically.

1.4.3 Object Detection for Man OverBoard Rescue

Modern Unmanned Aerial Vehicles (UAV) equipped with cameras can play an essential role in speeding up the identification and rescue of people who have fallen overboard, i.e., man overboard (MOB). To this end, many AI-based techniques have achieved outstanding results in localizing and recognizing people and objects in images and video frames in recent years. However, evaluating these approaches (or developing new ones) in a MOB scenario is difficult due to the lack of labeled data. To fill this gap, in [14], we collected and publicly released a large-scale dataset of aerial footage of people who, being in the water, simulated the need to be rescued [15]. Our dataset, named *MOBDrone*, contains 66 video clips with $126,170$ frames manually annotated with more than 180K bounding boxes (of which more than 113K belonging to the *person* category). The videos were gathered from one UAV flying at an altitude of 10 to 60 meters above the mean sea level. In the paper, we introduced our dataset, and we described the data collection and annotation processes. Moreover, we presented an in-depth experimental analysis of the performance of several state-of-the-art object detectors on this newly established MOB scenario, serving as baselines.

1.4.4 Learning from Virtual Worlds

In the new spring of artificial intelligence, particularly in its sub-field known as machine learning, a significant series of important results have shifted the focus of industrial and research communities toward the generation of valuable data from which learning algorithms can be trained. In the era of big data, the availability of real input examples to train machine learning algorithms is not considered an issue for several applications. However, there is not such an abundance of training data for several other applications. Sometimes, even if data is available, it must be manually revised to make it usable as training data (e.g., by adding annotations, class labels, or visual masks), with a considerable cost. Although a series of annotated datasets are available and successfully used to produce noteworthy academic results and commercially profitable products, there is still a considerable amount of scenarios where laborious human intervention is needed to produce high-quality training sets. An appealing solution is to gather synthetic data from virtual environments resembling the real world, where the labels are *automatically* collected by interacting with the graphical engine. In [50] and [51], we presented and publicly released *CrowdSim2*, a new synthetic collection of images gathered from a simulator based on the *Unity* graphical engine suitable for people and vehicle detection and tracking [89]. It consists of thousands of images gathered from various synthetic scenarios resembling the real world, where we varied some factors of interest, such as the weather conditions and the number of objects in the scenes. The labels are automatically collected and consist of bounding boxes that precisely localize objects belonging to the two object classes, leaving out humans from the annotation pipeline. We exploited this new benchmark as a testing ground for some state-of-the-art detectors and trackers, showing that our simulated scenarios can be a valuable tool for measuring their performances in a controlled environment. Following the same approach, in [19], we tackle falling people detection from RGB cameras exploiting synthetic data generated from a video game engine. Falling is one of the most common causes of injury, and early detection can ensure appropriate intervention. A typical limitation in commonly used approaches is the lack of generalization to unseen environments mainly due to the unavailability of large-scale and varied datasets. In this work, we mitigate these limitations with a general-purpose object detector trained using a virtual world dataset in addition to real-world images. Through extensive experimental evaluation, we verified that by training our models on synthetic images as well, we were able to improve their ability to generalize.

1.4.5 Video Violence Detection

Automatic detection of violent actions in videos is a challenging yet important task in many real-world scenarios since it is crucial for investigating the harmful abnormal contents from vast amounts of surveillance video data. One of the potential places in which an automatic violence detection system should be developed is public transport, such as buses, trains, etc. However, evaluating the existing approaches (or creating new ones) in this scenario is difficult due to the lack of labeled data. Although some annotated datasets for video violence detection in general contexts already exist, the same cannot be said for the case of public transport environments. To fill this gap, in [26], we introduced a benchmark specifically designed for this scenario. We collected and publicly released [49] a large-scale dataset gathered from multiple cameras located inside a moving bus where several people simulated violent actions, such as stealing an object from another person, fighting between passengers, etc. Our dataset, named *Bus Violence*, contains 1,400 video clips manually annotated as having or not violent scenes. Furthermore, we presented an in-depth experimental analysis of the performance of several state-of-the-art video violence detectors in this newly established scenario, serving as baselines. Specifically, we employed our *Bus Violence* dataset as a testing ground for evaluating the generalization capabilities of some of the most popular Deep Learning-based architectures suitable for video violence detection, pre-trained over general violence detection databases present in the literature. Indeed, the *Domain Shift* problem, i.e., the domain gap between the train and the test data distributions, is one of the most critical concerns affecting Deep Learning techniques, and it has become paramount to measure the performance of these algorithms against scenarios never seen during the supervised learning phase.

1.5 Multimedia Information Retrieval

1.5.1 Video Browsing

Video data is the fastest growing data type on the Internet, and because of the proliferation of high-definition video cameras, the volume of video data is exploding. This data explosion in the video area has led to push research on large-scale video retrieval systems that are effective, fast, and easy to use for content search scenarios.

Within this framework, we developed a content-based video retrieval system VISIONE[4], which also competed at the Video Browser Showdown (VBS), an international video search competition that evaluates the performance of interactive video retrievals systems. The tasks evaluated during the competition are: *Known-Item-Search (KIS)*, *textual KIS* and *Ad-hoc Video Search (AVS)*. The visual KIS task models the situation in which someone wants to find a particular video clip that he has already seen, assuming that it is contained in a specific collection of data. In the textual KIS, the target video clip is no longer visually presented to the participants of the challenge but it is rather described in details by text. This task simulates situations in which a user wants to find a particular video clip, without having seen it before, but knowing the content of the video exactly. For the AVS task, instead, a textual description is provided (e.g. “A person playing guitar outdoors") and participants need to find as many correct examples as possible, i.e. video shots that fit the given description.

VISIONE partecipated to VBS in 2019, 2021, and 2022. In [55], we described the 2021 competition settings, tasks and results and give an overview of state-of-the-art methods used by the competing systems. In [1] we presented the last release of the system that in the 2022 competition ranked first in the KIS visual task, and third in the entire competition.

VISIONE integrates several search functionalities that allow a user to search for a target video segment by formulating textual and visual queries, which can be also combined with a temporal search. In particular it supports *free text search*, *spatial color and object search*, *visual similarity search*, and *semantic similarity search*.

VISIONE is based on state-of-the-art deep learning approaches for the visual content analysis and exploits highly efficient indexing techniques to ensure scalability. In particular, it uses specifically designed textual encodings for indexing and searching video content. This aspect of our system is crucial: we can exploit the latest text search engine technologies, which nowadays are characterized by high efficiency and scalability, without the need to define a dedicated data structure or even worry about implementation issues.

In the last decade, user-centric video search competitions, such as VBS, have facilitated the evolution of interactive video search systems. So far, these competitions focused on a small number of search task categories, with few attempts to change task category configurations. In [63] we provided a structured description of a task category space for user-centric video search competitions. Using this concept of category space, new user-centric video search competitions can be designed to benchmark video search systems from different perspectives. We further analysed the three task categories considered so far at the Video Browser Showdown and discussed possible (but sometimes challenging) shifts within the task category space.

1.5.2 Similarity Search

Searching a data set for the most similar objects to a given query is a fundamental task in many branches of computer science, including pattern recognition, computational biology, and multimedia information retrieval, to name but a few. This search paradigm, referred to as *similarity search*, overcomes limitations of traditional *exact-match search* that is neither feasible nor meaningful for complex data (e.g., multimedia data, vectorial data, time-series, etc.). In our research, we mainly focus on *metric search* methods, which are based on the assumption that data objects are represented as elements of a space (D,d) where the metric function d provides a measure of the closeness (i.e. dissimilarity) of the data objects. A proximity query is defined by a query object $q \in D$ and a proximity condition, such as “find all the objects within a threshold distance of q” (*range query*) or “finding the k closest objects to q” (*k-nearest neighbour query*). The exact response to a query is the set of all the data objects that satisfy the considered proximity condition. Providing an exact response to a proximity query is not feasible if the search space is

[4] http://visione.isti.cnr.it/

very large or it has a high intrinsic dimensionality since in such cases, the exact search rarely outperforms a sequential scan (phenomenon known as the *curse of dimensionality*). To overcome this issue, the research community has developed a wide spectrum of techniques for *approximate search*, which have higher efficiency though at the price of some imprecision in the results (e.g. some relevant results might be missing or some ranking errors might occur).

All attempts to search a finite space efficiently, whether exactly or approximately, rely on using a data partitioning. A partition is an equivalence relation defined over the space so that each element of the space is contained within precisely one of the equivalence classes of the partition. At query time some set of principles are employed to include or exclude some partitions from the search, i.e. if the query is within one equivalence class, then one or more other classes either cannot, or probably do not, contain any of its solutions.

In [33, 91] we presented a generalised treatment of exclusion power for binary partitions. Our model abstracts over the partition type and we have shown its application to ball partitions, generalised hyperplane partitions and 4-point partitions. Exclusion power explains the well known differences in the number of exclusions that are possible with respect to both the dimensionality of the data and partition balance ratio. Understanding the probability of exclusion power determines if a dataset can be usefully queried at all using an exact metric search, i.e. if the size of candidate set is a small fraction of the size of the total dataset. This is useful in its own right since it may be applied independently of any particular algorithm to establish the amount of exclusion that is potentially possible.

1.5.3 Relational Cross-Modal Visual-Textual Retrieval

In the growing area of computer vision, modern deep-learning architectures are quite good at tasks such as classifying or recognizing objects in images. Recent studies, however, demonstrated the difficulties of such architectures to intrinsically understand a complex scene to catch spatial, temporal and abstract relationships among objects. Motivated by these limitations of the content-based information retrieval methods, we tried to explicitly handle relationships in multi-modal data, using attentive models. Specifically, we addressed the problem of cross-modal visual-textual retrieval, which consists in finding pictures given a natural language description as a query (sentence-to-image retrieval) or vice-versa (image-to-sentence retrieval). This task requires a deep understanding of both intra and inter-modal relationships to be effectively solved. This research direction has been deeply investigated in the PhD thesis by Messina (Section 4.1.2). Furthermore, we built on two earlier-presented architectures, TERN and TERAN, to develop ALADIN [70], a cross-modal architecture able to effectively and efficiently search large databases of images given a natural language description. Specifically, we employed as a backbone a large pre-trained vision-language transformer in a two-stream configuration with two different heads on top. The *alignment* head is able to obtain a fine-grained region-word alignment keeping the context into consideration. Despite its effectiveness, the alignment head is heavier to compute at inference time. The *matching* head, on the other hand, implements the matching of global features extracted from the visual-textual backbones in a common embedding space, where k-NN search is really fast. This space is optimized by distilling the scores coming from the alignment head, through a learning-to-rank training mechanism. We showed that this simple pipeline is better than previously proposed architectures for creating compact relational cross-modal descriptions that can be used for efficient similarity search.

This novel technique was integrated into VISIONE, a large-scale video retrieval tool developed by our group (see Section 1.5.1).

1.6 Medical Imaging

During this year, we applied our expertise on vision-based AI systems to research and develop healthcare and life science applications in collaboration with the Institute of Neuroscience of the CNR of Pisa. Our activity concerned the automatic analysis of medical images such as cell counting in fluorescence microscopy images, real-time pupillometry in IR images on mice and humans subjects, and detection of dementia disease from 3D MRI data. We describe the activities in detail in the following sections.

1.6.1 Cell Detection and Counting

Detection and counting of biological structures are among the earliest fields revolutionized by artificial neural networks now dominating state-of-the-art. Indeed, this task is crucial for the diagnosing of many diseases. To this end, several vision models (mostly convolutional networks) have been successfully adopted to localize, segment, and count cells or other structures from microscopy images and even provide counting-density estimation particularly effective in “crowded” scenarios. In [23], we investigated three baseline solutions belonging to the three main counting methodologies — a *segmentation-based* approach, a *localization-based* approach, and a *count-density estimation* approach — that have been successfully exploited for counting several different categories of objects, such as people and vehicles, and that represent the conceptual basis also for the cell counting techniques. Specifically, in addition to comparing the performance of these considered methods against state-of-the-art cell counters using established counting evaluation metrics, we also measured the ability of the models to localize the counted cells correctly. We showed that commonly adopted *counting* metrics (like mean absolute error) do not always agree with the *localization* performance of the tested models, and thus we suggested measuring both whenever possible to facilitate the practitioner in picking the most suitable solution.

It is worth noting that the success of these methods assumes the availability of a representative set of images with well-labeled biological structures. However, when trying to detect and count cells with non-trivial patterns on a large scale, several factors can produce weak labels: raters can incur errors due to fatigue or inexperience (common when hiring less-experienced raters to reduce labeling time) or have different judgments that can span from conservative to liberal when assigning labels. In [25], we investigated cell counting under the assumption of weak multi-rater labels, that is, in the presence of non-negligible disagreement between multiple raters. Specifically, we proposed a two-stage counting methodology that we tested against a novel weakly-labeled dot-annotated dataset that we publicly released [24]. It consists of a collection of fluorescence microscopy images of mice brain slices containing Perineuronal Nets (PNNs), extracellular matrix aggregates surrounding the cell body of a large number of neurons throughout the nervous system. Multiple expert raters have labeled a small part of the dataset; nonetheless, the maximum agreement between raters is roughly 70%, highlighting the need for an automated counting technique that accounts for uncertain patterns. We showed through experimental evaluation that our proposed two-stage pipeline, independently from the specific implementation of each stage, can improve the performance of several state-of-the-art counting methods on multiple ground-truth settings, from liberal to conservative ones.

1.6.2 Dementia Detection from 3D Imaging

Behavioral variant frontotemporal dementia (bvFTD) is a neurodegenerative syndrome whose clinical diagnosis remains a challenging task especially in the early stage of the disease. Currently, the presence of frontal and anterior temporal lobe atrophies on magnetic resonance imaging (MRI) is part of the diagnostic criteria for bvFTD. However, MRI data processing is usually dependent on the acquisition device and mostly require human-assisted crafting of feature extraction. Following the impressive improvements of deep architectures, in [41] we report on bvFTD identification using various classes of artificial neural networks, and present the results we achieved on classification accuracy and obliviousness on acquisition devices using extensive hyperparameter search. In particular, we will demonstrate the stability and generalization of different deep networks based on the attention mechanism, where data intra-mixing confers models the ability to identify the disorder even on MRI data in inter-device settings, i.e., on data produced by different acquisition devices and without model fine tuning, as shown from the very encouraging performance evaluations that dramatically reach and overcome the 90% value on the AuROC and balanced accuracy metrics.

1.7 Quantum Machine Learning

In recent years, Quantum Computing witnessed massive improvements both in terms of resources availability and algorithms development. The ability to harness quantum phenomena to solve computational problems is a long-standing dream that has drawn the scientific community's interest since the late '80s. In this regard, quantum computers might offer new solutions that exploit quantum phenomena such as interference, superposition, and entanglement. Such a characteristic is expected to speed up the computational time and to reduce the requirements for extensive resources, yielding the concepts of *quantum advantage* and *quantum supremacy*. In the last two decades, there has been a strong interest and commitment in the scientific community to develop quantum algorithms to solve Machine Learning problems, giving life to the field of *Quantum Machine Learning*.

In such a context, we posed our contribution [66]. First, we provided a gentle introduction to several basic notions about quantum mechanics, quantum information, and quantum computational models. Finally, we gathered, compared and analyzed the current state-of-the-art concerning Quantum Perceptrons and Quantum Neural Networks by discerning among theoretical formulations, simulations, and implementations on real quantum devices. The result is a thorough survey on the field of Quantum Neural Networks, which can be used as a guide by beginners, as well as a reference for more experienced practitioners. Moreover, we collected and organized the most relevant papers on this field on a GitHub page[5] allowing the interested readers to easily and quickly browse through the research literature.

1.8 Fighting misinformation

1.8.1 Deep Fake Detection

Detecting deep fakes has become a crucial task in modern society as sophisticated generative methods are increasingly being used to create fake images, videos, or news through social bots. Such ad-hoc-generated content is spread on the web, mostly via social networks, and is used to spread misinformation and fake news to contaminate public debate. Deep fake images and videos can harm public figures and spread fake news, so prompt detection is essential to prevent their spread. In this context, we have done a lot of work both investigating the problems of deepfake detection and proposing other solutions, in particular:

- Development of a new hybrid deepfake detection method, namely Convolutional Cross Vision Transformer [28], consisting of a Cross Vision Transformer and an EfficientNet-B0 that was able to achieve state-of-the-art results on DFDC [42] and FaceForensics++ [85] datasets while maintaining a low number of parameters;

- Participation in the ICIAP 2021 Face Deepfake Detection Challenge, which resulted in a paper summarising the various solutions presented by participants[53];

- Investigated the generalisation capability of various deep learning architectures on the deepfake detection task with the works [29];

- Development of a deepfake detector, called MINTIME [31], capable of effectively handling real-world situations such as multi-identity videos or variations in face-frame area ratio

[5] `https://github.com/fvmassoli/survey-quantum-computation`

Research activity reports of the previous years can be found in [78, 77].

2. Projects & Activities

2.1 EU Projects

AI4EU

In January 2019, the AI4EU consortium was established to build the first European Artificial Intelligence On-Demand Platform and Ecosystem with the support of the European Commission under the H2020 programme. The activities of the AI4EU project include:

- The creation and support of a large European ecosystem spanning the 28 countries to facilitate collaboration between all Europeans actors in AI (scientists, entrepreneurs, SMEs, Industries, funding organizations, citizens. . .);
- The design of a European AI on-Demand Platform to support this ecosystem and share AI resources produced in European projects, including high-level services, expertise in AI research and innovation, AI components and datasets, high-powered computing resources and access to seed funding for innovative projects using the platform;
- The implementation of industry-led pilots through the AI4EU platform, which demonstrates the capabilities of the platform to enable real applications and foster innovation; Research activities in five key interconnected AI scientific areas (Explainable AI, Physical AI ,Verifiable AI, Collaborative AI, Integrative AI), which arise from the application of AI in real-world scenarios;
- The funding of SMEs and start-ups benefitting from AI resources available on the platform (cascade funding plan of €3M) to solve AI challenges and promote new solutions with AI; The creation of a European Ethical Observatory to ensure that European AI projects adhere to high ethical, legal, and socio-economical standards;
- The production of a comprehensive Strategic Research Innovation Agenda for Europe; The establishment of an AI4EU Foundation that will ensure a handover of the platform in a sustainable structure that supports the European AI community in the long run.

The leader of the AIMH team participating in AI4EU is Giuseppe Amato.

AI4media

Artificial Intelligence for the Society and the Media Industry (AI4Media) is a network of research excellence centres delivering advances in AI technology in the media sector. Funded under H2020-EU.2.1.1., AI4Media started in September 2020 and will end in August 2024.

Motivated by the challenges, risks and opportunities that the wide use of AI brings to media, society and politics, AI4Media aspires to become a centre of excellence and a wide network of researchers across Europe and beyond, with a focus on delivering the next generation of core AI advances to serve the key sector of Media, to make sure that the European values of ethical and trustworthy AI are embedded in future AI deployments, and to reimagine AI as a crucial beneficial enabling technology in the service of Society and Media.

The leader of the AIMH team participating in AI4Media is Fabrizio Sebastiani.

ARIADNEplus

The ARIADNEplus project is the extension of the previous ARIADNE Integrating Activity, which successfully integrated archaeological data infrastructures in Europe, indexing in its registry about 2.000.000 datasets. ARIADNEplus will build on the ARIADNE results, extending and supporting the research community that the previous project created and further developing the relationships with key stakeholders such as the most important European archaeological associations, researchers, heritage professionals, national heritage agencies and so on. The new enlarged partnership of ARIADNEplus covers all of Europe. It now includes leaders in different archaeological domains like palaeoanthropology, bioarchaeology and environmental archaeology as well as other sectors of archaeological sciences, including all periods of human presence from the appearance of hominids to present times. Transnational Activities together with the planned training will further reinforce the presence of ARIADNEplus as a key actor. The technology underlying the project is state-of-art. The ARIADNEplus data infrastructure will be embedded in a cloud that will offer the availability of Virtual Research Environments where data-based archaeological research may be carried out. The project will furthermore develop a Linked Data approach to data discovery. Innovative services will be made available to users, such as visualization, annotation, text mining and geo-temporal data management. Innovative pilots will be developed to test and demonstrate the innovation potential of the ARIADNEplus approach. Fostering innovation will be a key aspect of the project, with dedicated activities led by the project Innovation Manager.

Mingei

The Mingei Project explores the possibilities of representing and making accessible both tangible and intangible aspects of craft as cultural heritage (CH). Heritage Crafts (HCs) involve craft artefacts, materials, and tools and encompass craftsmanship as a form of Intangible Cultural Heritage.

Intangible HC dimensions include dexterity, know-how, and skilled use of tools, as well as, tradition, and identity of the communities in which they are, or were, practiced. HCs are part of the history and have impact upon the economy of the areas in which they flourish. The significance and urgency to the preservation of HCs is underscored, as several are threatened with extinction. Despite their cultural significance efforts for HC representation and preservation are scattered geographically and thematically. Mingei provides means to establish HC representations based on digital assets, semantics, existing literature and repositories, as well as, mature digitisation and representation technologies. These representations will capture and preserve tangible and intangible dimensions of HCs. Central to craftsmanship is skill and its transmission from master to apprentice. Mingei captures the motion and tool usage of HC practitioners, from Living Human Treasures and archive documentaries, in order to preserve and illustrate skill and tool manipulation. The represented knowledge will be availed through experiential presentations, using storytelling and educational applications and based on Advanced Reality, Mixed Reality and the Internet. The project has started on December 1, 2019 and will last 3 years.

The main objective of this Action, entitled MULTI-modal Imaging of FOREnsic SciEnce Evidence (MULTI-FORESEE)-tools for Forensic Science[6], is to promote innovative, multi-informative, operationally deployable and commercially exploitable imaging solutions/technology to analyse forensic evidence.

Forensic evidence includes, but not limited to, fingermarks, hair, paint, biofluids, digital evidence, fibers, documents and living individuals. Imaging technologies include optical, mass spectrometric, spectroscopic, chemical, physical and digital forensic techniques complemented by expertise in IT solutions and computational modelling.

Imaging technologies enable multiple physical and chemical information to be captured in one analysis, from one specimen, with information being more easily conveyed and understood for a more rapid exploitation. The enhanced value of the evidence gathered will be conducive to much more informed investigations and judicial decisions thus contributing to both savings to the public purse and to a speedier and stronger criminal justice system.

The Action will use the unique networking and capacity-building capabilities provided by the COST framework to bring together the knowledge and expertise of Academia, Industry and End Users. This synergy is paramount to boost imaging technological developments which are operationally deployable.

The leader of the AIMH team participating in MultiForesee is Giuseppe Amato.

SoBigData++ is a project funded by the European Commission under the H2020 Programme INFRAIA-2019-1, started Jan 1 2020 and ending Dec 31, 2023. SoBigData++ proposes to create the Social Mining and Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”. SoBigData plans to open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research. It plans to not only strengthen the existing clusters of excellence in social data mining research, but also create a pan-European, inter-disciplinary community of social data scientists, fostered by extensive training, networking, and innovation activities.

The leader of the AIMH team participating in SoBigData++ is Alejandro Moreo.

Social and hUman ceNtered XR (SUN) is a project funded by the European Commission under the H2020 Programme HORIZON-CL4-2022-HUMAN-01-14, started Dec 1 2022 and ending Nov 30 2025. SUN aims at investigating and developing extended reality (XR) solutions that integrate the physical and the virtual world in a convincing way, from a human and social perspective. The virtual world will be a means to augment the physical world with new opportunities for social and human interaction.

Our institute is the leading partner of the project and the coordinator is Giuseppe Amato.

Social Sciences & Humanities Open Cloud (SSHOC) is a project funded by the EU framework programme Horizon 2020 and unites 20 partner organisations and their 27 associates in developing the social sciences and humanities area of the European Open Science Cloud (EOSC). SSHOC partners include both developing and fully established European Research Infrastructures from the social sciences and humanities, and the association of European research libraries (LIBER). The goal of the project is to transform the social sciences & humanities data landscape with its disciplinary silos and separate facilities into an integrated, cloud-based network of interconnected data infrastructures. To promote synergies and

[6] https://multiforesee.com/

open science initiatives between disciplines, and accelerate interdisciplinary research and collaboration, these data infrastructures will be supported by the tools and training which allow scholars and researchers to access, process, analyse, enrich and compare data across the boundaries of individual repositories or institutions. SSHOC will continuously monitor ongoing developments in the EOSC so as to conform to the necessary technical and other requirements for making the SSHOC services sustainable beyond the duration of the project. Some of the results obtained by the AIMH team involved in SSHOC have been presented in [NN] The leader of the AIMH team participating in SSHOC is Cesare Concordia. https://sshopencloud.eu

2.2 CNR National Virtual Lab on AI

Fabrizio Falchi has coordinated, together with Sara Colantonio, the activities of the National Virtual Lab of CNR on Artificial Intelligence. This initiative connects about 90 groups in 22 research institutes of 6 departments of the whole CNR. The Nationtal Virtual Lab on AI aims at proposing a strategic vision and big and long-term projects.

2.3 National Projects

AI-MAP

AI-MAP is a project funded by Regione Toscana that aims at analyzing digitized historical geographical regional maps using deep learning methods to increase the availability and searchability of the digitized documents. The main objectives of the project is to develop automatic or semi-automatic pipelines for denoising/repairing of the digitized documents, handwritten toponym localization and transcription. The activities are mainly conducted in the context of this project by Fabio Carrara under the scientific coordination of Giuseppe Amato.

AI4CHSites

AI4CHSites is a project funded by Regione Toscana that aims at analyzing visual content from surveillance camera in a touristic scenario. Partners of the project are: Opera della Primaziale Pisana and INERA srl. The activities in the context of this project are mainly conducted by Nicola Messina under the scientific coordination of Fabrizio Falchi.

HDN

Hypermedia Dante Network (HDN) is a three year (2020-2023) Italian National Research Project (PRIN) which aims to extend the ontology and tools developed by AIMH team to represent the sources of Dante Alighieri's minor works to the more complex world of the Divine Comedy. In particular, HDN aims to enrich the functionalities of the DanteSources Web application (https://dantesources.dantenetwork.it/) in order to efficiently recover knowledge about the Divine Comedy. Relying on some of the most important scientific institutions for Dante studies, such as the Italian Dante Society of Florence, HDN makes use of specialized skills, essential for the population of ontology and the consequent creation of a complete and reliable knowledge base. Knowledge will be published on the Web as Linked Open Data and will be access through a user-friendly Web application.

IMAGO

The IMAGO (Index Medii Aevi Geographiae Operum) is a three year (2020-2023) Italian National Research Project (PRIN) that aims at creating a knowledge base of the critical editions of Medieval and Humanistic Latin geographical works (VI-XV centuries). Up to now, this knowledge has been collected in many paper books or several databases, making it difficult for scholars to retrieve it easily and to produce a complete overview of these data. The goal of the project is to develop new tools that satisfy the needs of the academic research community, especially for scholars interested in Medieval and Renaissance Humanism geography. Using Semantic Web technologies, AIMH team will develop an ontology providing the terms to represent this knowledge in a machine-readable form. A semi-automatic tool will help the scholars to populate the ontology with the data included in authoritative critical editions. Afterwards, the tool will automatically save the resulting graph into a triple store. On top of this graph, a Web application will be developed, which will allow users to extract and display the information stored in the knowledge base in the form of maps, charts, and tables.

INAROS

INtelligenza ARtificiale per il mOnitoraggio e Supporto agli anziani (INAROS) is a 2-year project funded by Regione Toscana, Istituto di Scienza e Tecnologie dell'Informazione “A.Faedo" (ISTI) del CNR, Visual Engines srl. The main goal of the INAROS project is to build solutions for monitoring and surveillance of the elderly based on the use of autonomous smart cameras. Computer vision algorithms will be developed by leveraging artificial intelligence, in particular deep learning to automatically and in real time analyze video streams from smart cameras positioned in the home environment. To achieve these results, techniques will be developed for the tracking and detection of the elderly person's activity in the home environment and for the discovery of new activities and abnormalities of the elderly through off-line analysis of temporal patterns of learned events. Claudio Gennaro is the scientific coordinator of the project.

WAC@Lucca

WeAreClouds@Lucca carries out research and development activities in the field of monitoring public places, such as squares and streets, through cameras and microphones with artificial intelligence technologies, in order to collect useful information both for the evaluation of tourist flows and their impact. on the city, both for purposes of automatic identification of particular events of interest for statistical purposes or for security. The project is funded by Fondazione Cassa di Risparmio di Lucca and Comune di Lucca is a partner. Fabrizio Falchi is the scientific coordinator of the project.

NAUSICAA

NAUSICAA - “NAUtical Safety by means of Integrated Computer-Assistance Appliances 4.0" is a project funded by the Tuscany region (CUP D44E20003410009). The project aims at creating a system for medium and large boats in which the conventional control, propulsion, and thrust systems are integrated with a series of latest generation sensors (e.g., lidar systems, cameras, radar, drones) for assistance during the navigation and mooring phases. In the project, the AIMH researchers are mainly involved in developing techniques for the automatic analysis of video streams from cameras on boats and aerial drones based on artificial intelligence methods. Models and methods will be developed in particular for the recognition and tracking of people and objects in the water (e.g. for rescuing people at sea).

3. Papers

In this section, we report the complete list of paper we published in 2022 organized in four categories: journals, proceedings, magazines, others, and pre-prints.

3.1 Journals

In this section, we report the paper we published (or accepted for publication) in journals during 2022, in alphabetic order of the first author. Our works were published in the following journals (ordered by Impact Factor):

- **ACM Computing Surveys**
  ACM, IF 14.324: [66]
- **Medical Image Analysis**
  Elsevier, IF 13.828: [25]
- **Expert Systems with Applications**
  Elsevier, IF 8.665: [40, 27]
- **Computers in Biology and Medicine**
  Elsevier, IF 6.698: [41]
- **Human Molecular Genetics**
  Oxford University Press, IF 5.121: [93]
- **Neural Computing and Applications**
  Springer, IF 5.102: [58]
- **ACM Transactions on Information Systems**
  ACM, IF 4.657: [75]
- **Sensors**
  MDPI, IF 3.847: [26]
- **PLOS ONE**
  Public Library of Science, IF 3.752: [76]
- **IEEE Access**
  IEEE, IF 3.476: [43]
- **Journal of the Association for Information Science and Technology**
  Wiley, IF 3.275: [37]
- **International Journal of Multimedia Information Retrieval**
  Springer, IF 3.205: [55]
- **ACM Journal of Computing and Cultural Heritage**
  ACM, IF : 2.047 [38, 80]
- **Heritage**
  MDPI, IF to be assigned: [94]
- **Journal of Imaging**
  MDPI, IF to be assigned: [54]
- **Information**
  MDPI, IF to be assigned: [17]
- **International Journal on Digital Libraries**
  Springer, IF to be assigned: [90, 7]

3.1.1 Behavioral impulsivity is associated with pupillary alterations and hyperactivity in CDKL5 mutant mice

A. Viglione, G. Sagona, F. Carrara, G. Amato, V. Totaro, L. Lupori, E. Putignano, T. Pizzorusso, R. Mazziotti Human Molecular Genetics, Oxford University Press. [93]

Cyclin-dependent kinase-like 5 (Cdkl5) deficiency disorder (CDD) is a severe neurodevelopmental condition caused by mutations in the X-linked Cdkl5 gene. CDD is characterized by early-onset seizures in the first month of life, intellectual disability, motor and social impairment. No effective treatment is currently available and medical management is only symptomatic and supportive. Recently, mouse models of Cdkl5 disorder have demonstrated that mice lacking Cdkl5 exhibit autism-like phenotypes, hyperactivity and dysregulations of the arousal system, suggesting the possibility to use these features as translational biomarkers. In this study, we tested Cdkl5 male and female mutant mice in an appetitive operant conditioning chamber to assess cognitive and motor abilities, and performed pupillometry to assess the integrity of the arousal system. Then, we evaluated the performance of artificial intelligence models to classify the genotype of the animals from the behavioral and physiological phenotype. The behavioral results show that CDD mice display impulsivity, together with low levels of cognitive flexibility and perseverative behaviors. We assessed arousal levels by simultaneously recording pupil size and locomotor activity. Pupillometry reveals in CDD mice a smaller pupil size and an impaired response to unexpected stimuli associated with hyperlocomotion, demonstrating a global defect in arousal modulation. Finally, machine learning reveals that both behavioral and pupillometry parameters can be considered good predictors of CDD. Since early diagnosis is essential to evaluate treatment outcomes and pupillary measures can be performed easily, we proposed the monitoring of pupil size as a promising biomarker for CDD.

3.1.2 Bus violence: an open benchmark for video violence detection on public transport

L. Ciampi, P. Foszner, N. Messina, M. Staniszewski, C. Gennaro, F. Falchi, G. Serao, M. Cogiel, D. Golba, A. Szczesna, G. Amato. Sensors, MDPI. [26]

The automatic detection of violent actions in public places through video analysis is difficult because the employed Artificial Intelligence-based techniques often suffer from generalization problems. Indeed, these algorithms hinge on large quantities of annotated data and usually experience a drastic drop in performance when used in scenarios never seen during the supervised learning phase. In this paper, we introduce and publicly release the Bus Violence *benchmark, the first large-scale collection of video clips for violence detection on public transport, where some actors simulated violent actions inside a moving bus in changing conditions, such as the background or light. Moreover, we conduct a performance analysis of several state-of-the-art video violence detectors pre-trained with general violence detection databases on this newly established use case. The achieved moderate performances reveal the difficulties in generalizing from these popular methods, indicating the need to have this new collection of labeled data, beneficial for specializing them in this new scenario.*

3.1.3 Comparing the performance of Hebbian against backpropagation learning using convolutional neural networks

G. Lagani, F. Falchi, C. Gennaro, G. Amato. Neural Computing and Applications, Springer. [58]

In this paper, we investigate Hebbian learning strategies applied to Convolutional Neural Network (CNN) training. We consider two unsupervised learning approaches, Hebbian Winner-Takes-All (HWTA), and Hebbian Principal Component Analysis (HPCA). The Hebbian learning rules are used to train the layers of a CNN in order to extract features that are then used for classification, without requiring backpropagation (backprop). Experimental comparisons are made with state-of-the-art unsupervised (but backprop-based) Variational Auto-Encoder (VAE) training. For completeness,we consider two supervised Hebbian learning variants (Supervised Hebbian Classifiers—SHC, and Contrastive Hebbian Learning—CHL), for training the final classification layer, which are compared to Stochastic Gradient Descent training. We also investigate hybrid learning methodologies, where some network layers are trained following the Hebbian approach, and others are trained by backprop. We tested our approaches on MNIST, CIFAR10, and CIFAR100 datasets. Our results suggest that Hebbian learning is generally suitable for training early feature extraction layers, or to retrain higher network layers in fewer training epochs than backprop. Moreover, our experiments show that Hebbian learning outperforms VAE training, with HPCA performing generally better than HWTA.

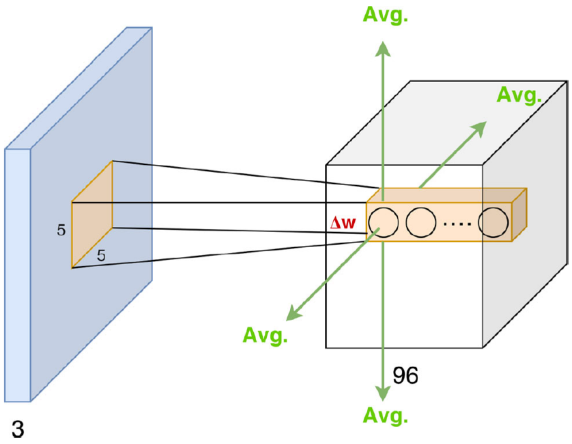

Figure 1. Hebbian Learning on deep CNNs: update averaging over horizontal and vertical dimensions

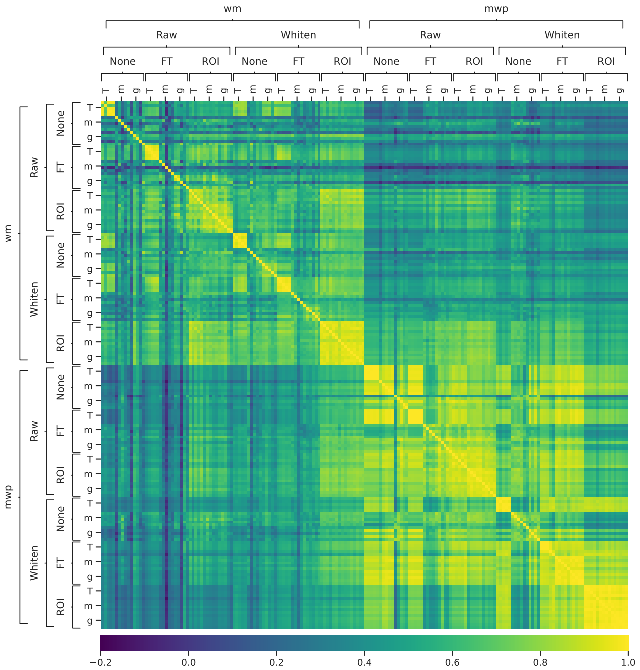

Figure 2. Dementia detection: Pearson Correlation Coefficients among predictions of various transformer-based models. T = ViT, m = MLP-Mixer, g = gMLP. With ROI usage, predictions tend to highly correlate independently from the model or random weight initialization used.

3.1.4 Deep networks for behavioral variant frontotemporal dementia identification from multiple acquisition sources

M. Di Benedetto, F. Carrara, B. Tafuri, S. Nigro, R. De Blasi, F. Falchi, C. Gennaro, G. Gigli, G. Logroscino, G. Amato. Computers in Biology and Medicine, Elsevier. [41]

Behavioral variant frontotemporal dementia (bvFTD) is a neurodegenerative syndrome whose clinical diagnosis remains a challenging task especially in the early stage of the disease. Currently, the presence of frontal and anterior temporal lobe atrophies on magnetic resonance imaging (MRI) is part of the diagnostic criteria for bvFTD. However, MRI data processing is usually dependent on the acquisition device and mostly require human-assisted crafting of feature extraction. Following the impressive improvements of deep architectures, in this study we report on bvFTD identification using various classes of artificial neural networks, and present the results we achieved on classification accuracy and obliviousness on acquisition devices using extensive hyperparameter search. In particular, we will demonstrate the stability and generalization of different deep networks based on the attention mechanism, where data intra-mixing confers models the ability to identify the disorder even on MRI data in inter-device settings, i.e., on data produced by different acquisition devices and without model fine tuning, as shown from the very encouraging performance evaluations that dramatically reach and overcome the 90% value on the AuROC and balanced accuracy metrics.

3.1.5 An embedded toolset for human activity monitoring in critical environments

M. Di Benedetto, F. Carrara, L. Ciampi, F. Falchi, C. Gennaro, G. Amato. Expert Systems with Applications, Elsevier. [40]

In many working and recreational activities, there are scenarios where both individual and collective safety have to be constantly checked and properly signaled, as occurring in dangerous workplaces or during pandemic events like the recent COVID-19 disease. From wearing personal protective equipment to filling physical spaces with an adequate number of people, it is clear that a possibly automatic solution would help to check compliance with the established rules. Based on an off-the-shelf compact and low-cost hardware, we present a deployed real use-case embedded system capable of perceiving people's behavior and aggregations and supervising the appliance of a set of rules relying on a configurable plug-in framework. Working on indoor and outdoor environments, we show that our implementation of counting people aggregations, measuring their reciprocal physical distances, and checking the proper usage of protective equipment is an effective yet open framework for monitoring human activities in critical conditions.

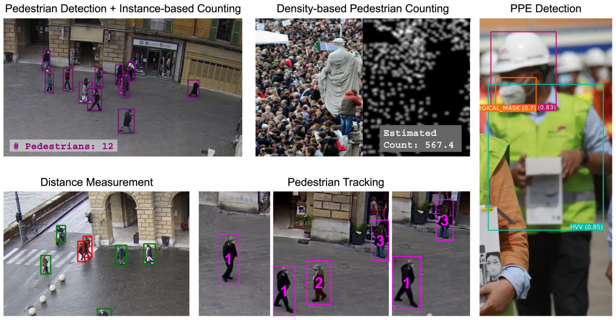

Figure 3. Visualization of output examples of the modules currently available in our embedded toolset for human activity monitoring in critical environments. [40].

3.1.6 An exploratory approach to archaeological knowledge production

C. Thanos, C. Meghini, V. Bartalesi, G. Coro International Journal on Digital Libraries, Springer. [90]

The current scientific context is characterized by intensive digitization of the research outcomes and by the creation of data infrastructures for the systematic publication of datasets and data services. Several relationships can exist among these outcomes. Some of them are explicit, e.g. the relationships of spatial or temporal similarity, whereas others are hidden, e.g. the relationship of causality. By materializing these hidden relationships through a linking mechanism, several patterns can be established. These knowledge patterns may lead to the discovery of information previously unknown. A new approach to knowledge production can emerge by following these patterns. This new approach is exploratory because by following these patterns, a researcher can get new insights into a research problem. In the paper, we report our effort to depict this new exploratory approach using Linked Data and Semantic Web technologies (RDF, OWL). As a use case, we apply our approach to the archaeological domain.

3.1.7 The Face Deepfake Detection Challenge

L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L.M.Q. Bui, M. Fontani, D.A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, N. Messina, G. Amato, G. Perelli, S. Concas, C. Cuccu, G. Orrù, G.L. Marcialis, S. Battiato. Journal of Imaging, MDPI. [54]

Multimedia data manipulation and forgery has never been easier than today, thanks to the power of Artificial Intelligence (AI). AI-generated fake content, commonly called Deepfakes, have been raising new issues and concerns, but also new challenges for the research community. The Deepfake detection task has become widely addressed, but unfortunately, approaches in the literature suffer from generalization issues. In this paper, the Face Deepfake Detection and Reconstruction Challenge is described. Two different tasks were proposed to the participants: (i) creating a Deepfake detector capable of working in an "in the wild" scenario; (ii) creating a method capable of reconstructing original images from Deepfakes. Real images from CelebA and FFHQ and Deepfake images created by StarGAN, StarGAN-v2, StyleGAN, StyleGAN2, AttGAN and GDWCT were collected for the competition. The winning teams were chosen with respect to the highest classification accuracy value (Task I) and "minimum average distance to Manhattan" (Task II). Deep Learning algorithms, particularly those based on the EfficientNet architecture, achieved the best results in Task I. No winners were proclaimed for Task II. A detailed discussion of teams proposed methods with corresponding ranking is presented in this paper.

3.1.8 A Leap among Quantum Computing and Quantum Neural Networks: A Survey

F.V. Massoli, L. Vadicamo, G. Amato, F. Falchi. ACM Computing Surveys [66]

In recent years, Quantum Computing witnessed massive improvements in terms of available resources and algorithms development. The ability to harness quantum phenomena to solve computational problems is a long-standing dream that has drawn the scientific community's interest since the late '80s. In such a context, we propose our contribution. First, we introduce basic concepts related to quantum computations, and then we explain the core functionalities of technologies that implement the Gate Model and Adiabatic Quantum Computing paradigms. Finally, we gather, compare, and analyze the current state-of-the-art concerning Quantum Perceptrons and Quantum Neural Networks implementations.

3.1.9 Learning to count biological structures with raters' uncertainty

L. Ciampi, F. Carrara, V. Totaro, R. Mazziotti, L. Lupori, C. Santiago, G. Amato, T. Pizzorusso, C. Gennaro. Medical Image Analysis, Elsevier [25]

Exploiting well-labeled training sets has led deep learning models to astonishing results for counting biological structures in microscopy images. However, dealing with weak multi-rater annota-

tions, i.e., when multiple human raters disagree due to non-trivial patterns, remains a relatively unexplored problem. More reliable labels can be obtained by aggregating and averaging the decisions given by several raters to the same data. Still, the scale of the counting task and the limited budget for labeling prohibit this. As a result, making the most with small quantities of multi-rater data is crucial. To this end, we propose a two-stage counting strategy in a weakly labeled data scenario. First, we detect and count the biological structures; then, in the second step, we refine the predictions, increasing the correlation between the scores assigned to the samples and the raters' agreement on the annotations. We assess our methodology on a novel dataset comprising fluorescence microscopy images of mice brains containing extracellular matrix aggregates named perineuronal nets. We demonstrate that we significantly enhance counting performance, improving confidence calibration by taking advantage of the redundant information characterizing the small sets of available multi-rater data.

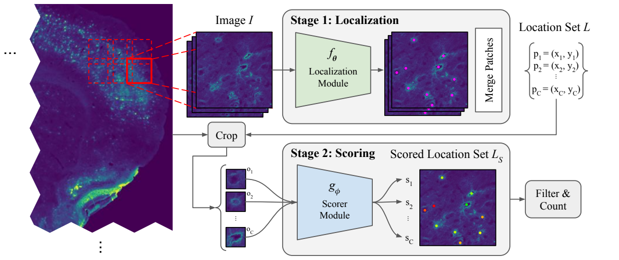

Figure 4. The proposed two-stage counting technique. [25].

3.1.10
Linking different scientific digital libraries in Digital Humanities: the IMAGO case study

V. Bartalesi, N. Pratelli, E. Lenzi International Journal on Digital Libraries, Springer. [7]

In the last years, several scientific digital libraries (DLs) in digital humanities (DH) field have been developed following the Open Science principles. These DLs aim at sharing the research outcomes, in several cases as FAIR data, and at creating linked information spaces. In several cases, to reach these aims the Semantic Web technologies and Linked Data have been used. This paper presents how the current scientific DLs in the DH field can provide the creation of linked information spaces and navigational services that allow users to navigate them, using Semantic Web technologies to formally represent, search and browsing knowledge. To support the argument, we present our experience in developing a scientific DL supporting scholars in creating, evolving and consulting a knowledge base related to Medieval and Renaissance geographical works within the three years (2020–2023) Italian National research project IMAGO—Index Medii Aevi Geographiae Operum. In the presented case study, a linked information space was created to allow users to discover and navigate knowledge across multiple repositories, thanks to the extensive use of ontologies. In particular, the linked information spaces created within the IMAGO project make use of five different datasets, i.e. Wikidata, the MIRABILE digital archive, the Nuovo Soggettario thesaurus, Mapping Manuscript Migration knowledge base and the Pleiades gazetteer. The linking among different datasets allows to considerably enrich the knowledge collected in the IMAGO KB.

3.1.11
Multi-camera vehicle counting using edge-AI

L. Ciampi, C. Gennaro, F. Carrara, F. Falchi, C. Vairo, G. Amato. Expert Systems with Applications, Elsevier. [27]

This paper presents a novel solution to automatically count vehicles in a parking lot using images captured by smart cameras. Unlike most of the literature on this task, which focuses on the analysis of single images, this paper proposes the use of multiple visual sources to monitor a wider parking area from different perspectives. The proposed multi-camera system is capable of automatically estimating the number of cars present in the entire parking lot directly on board the edge devices. It comprises an on-device deep learning-based detector that locates and counts the vehicles from the captured images and a decentralized geometric-based approach that can analyze the inter-camera shared areas and merge the data acquired by all the devices. We conducted the experimental evaluation on an extended version of the CNRPark-EXT dataset, a collection of images taken from the parking lot on the campus of the National Research Council (CNR) in Pisa, Italy. We show that our system is robust and takes advantage of the redundant information deriving from the different cameras, improving the overall performance without requiring any extra geometrical information of the monitored scene.

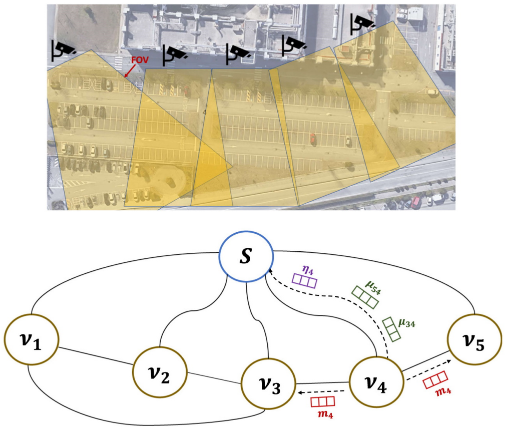

Figure 5. An example of our multi-camera counting system [27].

3.1.12
ICS: Total Freedom in Manual Text Classification Supported by Unobtrusive Machine Learning

A. Esuli. IEEE Access. [43]

We present the Interactive Classification System (ICS), a web-based application that supports the activity of manual text classification. The application uses machine learning to continuously fit automatic classification models that are in turn used to actively support its users with classification suggestions. The key requirement we have established for the development of ICS is to give its users total freedom of action: they can at any time modify any classification schema and any label assignment, possibly reusing any *relevant information from previous activities. We investigate how this requirement challenges the typical scenarios faced in machine learning research, which instead give no active role to humans or place them into very constrained roles, e.g., on-demand labeling in active learning processes, and always assume some degree of batch processing of data. We satisfy the "total freedom" requirement by designing an unobtrusive machine learning model, i.e., the machine learning component of ICS acts as an unobtrusive observer of the users, that never interrupts them, continuously adapts and updates its models in response to their actions, and it is always available to perform automatic classifications. Our efficient implementation of the unobtrusive machine learning model combines various machine learning methods and technologies, such as hash-based feature mapping, random indexing, online learning, active learning, and asynchronous processing.*

3.1.13
Improving the Adversarial Robustness of Neural ODE Image Classifiers by Tuning the Tolerance Parameter

F. Carrara, R. Caldelli, F. Falchi, G. Amato. Information, MDPI. [17]

The adoption of deep learning-based solutions practically pervades all the diverse areas of our everyday life, showing improved performances with respect to other classical systems. Since many applications deal with sensible data and procedures, a strong demand to know the actual reliability of such technologies is always present. This work analyzes the robustness characteristics of a specific kind of deep neural network, the neural ordinary differential equations (N-ODE) network. They seem very interesting for their effectiveness and a peculiar property based on a test-time tunable parameter that permits obtaining a trade-off between accuracy and efficiency. In addition, adjusting such a tolerance parameter grants robustness against adversarial attacks. Notably, it is worth highlighting how decoupling the values of such a tolerance between training and test time can strongly reduce the attack success rate. On this basis, we show how such tolerance can be adopted, during the prediction phase, to improve the robustness of N-ODE to adversarial attacks. In particular, we demonstrate how we can exploit this property to construct an effective detection strategy and increase the chances of identifying adversarial examples in a non-zero knowledge attack scenario. Our experimental evaluation involved two standard image classification benchmarks. This showed that the proposed detection technique provides high rejection of adversarial examples while maintaining most of the pristine samples.

3.1.14
Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown

S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. Þ. Jónsson, J. Lokoč, A. Leibetseder, F. Mejzlík, L. Peška, L. Rossetto, K. Schall, K. Schoeffmann, H.Schuldt, F. Spiess, L. Tran, L. Vadicamo, P. Veselý, S. Vrochidis, and J. Wu IJMIR, Springer. [55]

The Video Browser Showdown addresses difficult video search challenges through an annual interactive evaluation campaign attracting research teams focusing on interactive video retrieval. The campaign aims to provide insights into the performance of participating interactive video retrieval systems, tested by selected search tasks on large video collections. For the first time in its ten year history, the Video Browser Showdown 2021 was organized in a fully remote setting and hosted a record number of sixteen scoring systems. In this paper, we describe the competition setting, tasks and results and give an overview of state-of-the-art methods used by the competing systems. By looking at query result logs provided by ten systems, we analyze differences in retrieval model performances and browsing times before a correct submission. Through advances in data gathering methodology and tools, we provide a comprehensive analysis of ad-hoc video search tasks, discuss results, task design and methodological challenges. We highlight that almost all top performing systems utilize some sort of joint embedding for text-image retrieval and enable specification of temporal context in queries for known-item search. Whereas a combination of these techniques drive the currently top performing systems, we identify several future challenges for interactive video search engines and the Video Browser Showdown competition itself.

3.1.15
Representation of socio-historical context to support the authoring and presentation of multimodal narratives: the Mingei online platform

N. Partarakis, P. Doulgeraki, E. Karuzaki, I. Adami, S. Ntoa, D. Metilli, V. Bartalesi, C. Meghini, Y. Marketakis, D. Kaplanidi, M. Theodoridou, X. Zabulis Journal on computing and cultural heritage, ACM. [80]

In this article, the Mingei Online Platform is presented as an authoring platform for the representation of social and historic context encompassing a focal topic of interest. The proposed representation is employed in the contextualised presentation of a given topic, through documented narratives that support its presentation to diverse audiences. Using the obtained representation, the documentation and digital preservation of social and historical dimensions of Cultural Heritage are demonstrated. The implementation follows the Human-Centred Design approach and has been conducted under an iterative design and evaluation approach involving both usability and domain experts.

3.1.16
A representation protocol for traditional crafts

X. Zabulis, N. Partarakis, C. Meghini, A. Dubois, S. Manitsaris, H. Hauser, N. M. Thalmann, C. Ringas, L. Panesse, N. Cadi, E. Baka, C. Beisswenger, D. Makrygiannis, A. Glushkova, B. E. O. Padilla, D. Kaplanidi, E. Tasiopoulou, C. Cuenca, A. L. Carre, V. Nitti, I. Adami, E. Zidianakis, P. Doulgeraki, E. Karouzaki, V. Bartalesi, D. Metilli Heritage, MDPI. [94]

A protocol for the representation of traditional crafts and the tools to implement this are proposed. The proposed protocol is a method for the systematic collection and organization of digital assets and knowledge, their representation into a formal model, and their utilization for research, education, and preservation. A set of digital tools accompanies this protocol that enables the online curation of craft representations. The proposed approach was elaborated *and evaluated with craft practitioners in three case studies. Lessons learned are shared and an outlook for future work is provided.*

3.1.17 Generalized Funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification

A. Moreo, A. Pedrotti, F. Sebastiani. ACM TOIS [75]

Funnelling (Fun) is a recently proposed method for cross-lingual text classification (CLTC) based on a two-tier learning ensemble for heterogeneous transfer learning (HTL). In this ensemble method, 1st-tier classifiers, each working on a different and language-dependent feature space, return a vector of calibrated posterior probabilities (with one dimension for each class) for each document, and the final classification decision is taken by a meta-classifier that uses this vector as its input. The meta-classifier can thus exploit class-class correlations, and this (among other things) gives Fun an edge over CLTC systems in which these correlations cannot be brought to bear. In this article, we describe Generalized Funnelling (gFun), a generalization of Fun consisting of an HTL architecture in which 1st-tier components can be arbitrary view-generating functions, i.e., language-dependent functions that each produce a language-independent representation ("view") of the (monolingual) document. We describe an instance of gFun in which the meta-classifier receives as input a vector of calibrated posterior probabilities (as in Fun) aggregated to other embedded representations that embody other types of correlations, such as word-class correlations (as encoded by Word-Class Embeddings), word-word correlations (as encoded by Multilingual Unsupervised or Supervised Embeddings), and word-context correlations (as encoded by multilingual BERT). We show that this instance of gFun substantially improves over Fun and over state-of-the-art baselines by reporting experimental results obtained on two large, standard datasets for multilingual multilabel text classification. Our code that implements gFun is publicly available.

3.1.18 Tweet sentiment quantification: An experimental re-evaluation

A. Moreo, F. Sebastiani. PLOS ONE [76]

Sentiment quantification is the task of training, by means of supervised learning, estimators of the relative frequency (also called "prevalence") of sentiment-related classes (such as Positive, Neutral, Negative) in a sample of unlabelled texts. This task is especially important when these texts are tweets, since the final goal of most sentiment classification efforts carried out on Twitter data is actually quantification (and not the classification of individual tweets). It is well-known that solving quantification by means of "classify and count" (i.e., by classifying all unlabelled items by means of a standard classifier and counting the items that have been assigned to a given class) is less than optimal in terms of accuracy, and that more accurate quantification methods exist. Gao and Sebastiani 2016 carried out a systematic comparison of quantification methods on the task of tweet sentiment quantification. In hindsight, we observe that the experimentation carried out in that work was weak, and that the reliability of the conclusions that were drawn from the results is thus questionable. We here re-evaluate those quantification methods (plus a few more modern ones) on exactly the same datasets, this time following a now consolidated and robust experimental protocol (which also involves simulating the presence, in the test data, of class prevalence values very different from those of the training set). This experimental protocol (even without counting the newly added methods) involves a number of experiments 5,775 times larger than that of the original study. Due to the above-mentioned presence, in the test data, of samples characterised by class prevalence values very different from those of the training set, the results of our experiments are dramatically different from those obtained by Gao and Sebastiani, and provide a different, much more solid understanding of the relative strengths and weaknesses of different sentiment quantification methods.

3.1.19 MedLatinEpi and MedLatinLit: Two datasets for the computational authorship analysis of medieval Latin texts

S. Corbara, A. Moreo, F. Sebastiani, M. Tavoni. JOCCH [38]

We present and make available MedLatinEpi and MedLatinLit, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatinEpi and MedLatinLit consist of 294 and 30 curated texts, respectively, labelled by author; MedLatinEpi texts are of epistolary nature, while MedLatinLit texts consist of literary comments and treatises about various subjects. As such, these two datasets lend themselves to supporting research in authorship analysis tasks, such as authorship attribution, authorship verification, or same-author verification. Along with the datasets, we provide experimental results, obtained on these datasets, for the authorship verification task, i.e., the task of predicting whether a text of unknown authorship was written by a candidate author. We also make available the source code of the authorship verification system we have used, thus allowing our experiments to be reproduced, and to be used as baselines, by other researchers. We also describe the application of the above authorship verification system, using these datasets as training data, for investigating the authorship of two medieval epistles whose authorship has been disputed by scholars.

3.1.20 Syllabic quantity patterns as rhythmic features for Latin authorship attribution

S. Corbara, A. Moreo, F. Sebastiani. JASIST [37]

Abstract It is well known that, within the Latin production of written text, peculiar metric schemes were followed not only in poetic compositions, but also in many prose works. Such metric patterns were based on so-called syllabic quantity, that is, on the length of the involved syllables, and there is substantial evidence suggesting that certain authors had a preference for certain metric patterns over others. In this research we investigate the possibility to employ syllabic quantity as a base for deriving rhythmic features for the task of computational authorship attribution of Latin prose texts. We test the impact of these features on the authorship attribution task when combined with other topic-agnostic features. Our experiments, carried out on three different datasets using support vector machines (SVMs) show that rhythmic features based on syllabic quantity are beneficial in discriminating among Latin prose authors.

3.2 Proceedings

In this section, we report the paper we published in alphabetic order of the first author. Our works were presented, and published in the proceedings of the following conferences:

- **CBMI** – International Conference on Content-based Multimedia Indexing. [19, 70, 57]
- **DH** – DH2022 - Digital Humanities 2022 responding to Asian diversity [60]
- **MAD** – International Workshop on Multimedia AI against Disinformation. [30]
- **WebSci** – 14th ACM Web Science Conference. [52]
- **IRCDL** – 18th Italian Research Conference on Digital Libraries. [72, 6]
- **ICIAP** – International Conference on Image Analysis and Processing. [14, 68, **?**]
- **ICIP** – IEEE International Conference on Image Processing. [16]
- **ISCC** – IEEE Symposium on Computers and Communications. [3]
- **I-CiTieS** – 8th Italian Conference on ICT for Smart Cities And Communities. [2]
- **Ital-IA** – Ital-IA secondo Convegno Nazionale CINI sull'Intelligenza Artificiale. [79, 47, 56, 64, 13]
- **MMM** – International Conference on Multimedia Modeling. [1, 63]
- **SEBD** – 30th Italian Symposium on Advanced Database Systems. [33]
- **SISAP** – International Conference on Similarity Search and Applications. [20, 91, 59]
- **TPDL** – 26th International Conference on Theory and Practice of Digital Libraries. [32, 83]
- **VISIGRAPP** – International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. [67, 23]
- **VTC2022-Spring** – 95th IEEE Vehicular Technology Conference. [65]
- **CIRCLE** – 2nd Joint Conference of the Information Retrieval Communities in Europe [73]
- **ECML/PKDD** – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases [12]
- **CLEF** – Conference and Labs of the Evaluation Forum, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization [34, 45, 46]
- **ECIR** – 44th European Conference on Information Retrieval [44]
- **NLDB** – Natural Language Processing and Information Systems, 27th International Conference on Applications of Natural Language to Information Systems [35]

3.2.1
Actor-Critic Scheduling for Path-Aware Air-to-Ground Multipath Multimedia Delivery

A. Machumilane, A. Gotta, P. Cassarà, C. Gennaro, G. Amato 95th IEEE Vehicular Technology Conference: VTC2022-Spring 2022 [65]

Reinforcement Learning (RL) has recently found wide applications in network traffic management and control because some of its variants do not require prior knowledge of network models. In this paper, we present a novel scheduler for real-time multimedia delivery in multipath systems based on an Actor-Critic (AC) RL algorithm. We focus on a challenging scenario of real-time video streaming from an Unmanned Aerial Vehicle (UAV) using multiple wireless paths. The scheduler acting as an RL agent learns in real-time the optimal policy for path selection, path rate allocation and redundancy estimation for flow protection. The scheduler, implemented as a module of the GStreamer framework, can be used in real or simulated settings. The simulation results show that our scheduler can target a very low loss rate at the receiver by dynamically adapting in real-time the scheduling policy to the path conditions without performing training or relying on prior knowledge of network channel models.

3.2.2
ALADIN: Distilling Fine-Grained Alignment Scores for Efficient Image-Text Matching and Retrieval

N. Messina, M. Stefanini, M. Cornia, L. Baraldi, F. Falchi, G. Amato, R. Cucchiara International Conference on Content-based Multimedia Indexing (CBMI). [70]

Image-text matching is gaining a leading role among tasks involving the joint understanding of vision and language. In literature, this task is often used as a pre-training objective to forge architectures able to jointly deal with images and texts. Nonetheless, it has a direct downstream application: cross-modal retrieval, which consists in finding images related to a given query text or vice-versa. Solving this task is of critical importance in cross-modal search engines. Many recent methods proposed effective solutions to the image-text matching problem, mostly using recent large vision-language (VL) Transformer networks. However, these models are often computationally expensive, especially at inference time. This prevents their adoption in large-scale cross-modal retrieval scenarios, where results should be provided to the user almost instantaneously. In this paper, we propose to fill in the gap between effectiveness and efficiency by proposing an ALign And DIstill Network (ALADIN). ALADIN first produces high-effective scores by aligning at fine-grained level images and texts. Then, it learns a shared embedding space – where an efficient kNN search can be performed – by distilling the relevance scores obtained from the fine-grained alignments. We obtained remarkable results on MS-COCO, showing that our method can compete with state-of-the-art VL Transformers while being almost 90 times faster. The code for reproducing our results is available at `https://github.com/mesnico/ALADIN`.

3.2.3
Combining efficientnet and vision transformers for video deepfake detection

D.A. Coccomini, N. Messina, C. Gennaro, F. Falchi International Conference on Image Analysis and Processing (ICIAP)

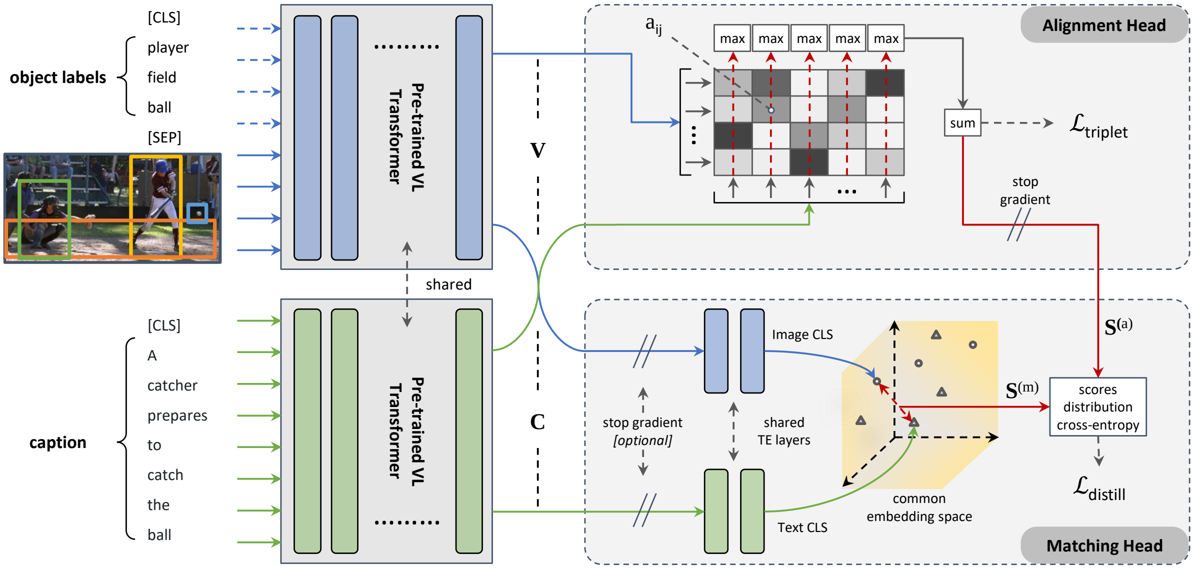

Figure 6. Overview of the ALADIN architecture [57].

[28]

Deepfakes are the result of digital manipulation to forge realistic yet fake imagery. With the astonishing advances in deep generative models, fake images or videos are nowadays obtained using variational autoencoders (VAEs) or Generative Adversarial Networks (GANs). These technologies are becoming more accessible and accurate, resulting in fake videos that are very difficult to be detected. Traditionally, Convolutional Neural Networks (CNNs) have been used to perform video deepfake detection, with the best results obtained using methods based on EfficientNet B7. In this study, we focus on video deep fake detection on faces, given that most methods are becoming extremely accurate in the generation of realistic human faces. Specifically, we combine various types of Vision Transformers with a convolutional EfficientNet B0 used as a feature extractor, obtaining comparable results with some very recent methods that use Vision Transformers. Differently from the state-of-the-art approaches, we use neither distillation nor ensemble methods. Furthermore, we present a straightforward inference procedure based on a simple voting scheme for handling multiple faces in the same video shot. The best model achieved an AUC of 0.951 and an F1 score of 88.0%, very close to the state-of-the-art on the DeepFake Detection Challenge (DFDC). The code for reproducing our results is publicly available here: `https://tinyurl.com/cnn-vit-dfd`.

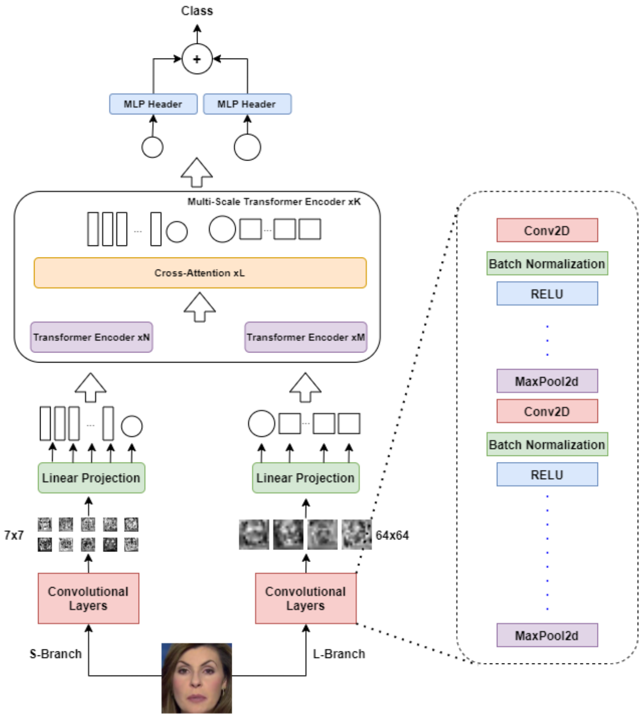

Figure 7. Overview of the Convolutional Cross Vision Transformer architecture [28].

3.2.4
Deep Features for CBIR with Scarce Data using Hebbian Learning

G. Lagani, D. Bacciu, C. Gallicchio, F. Falchi, C. Gennaro, G. Amato International Conference on Content-based Multimedia Indexing (CBMI). [57]

Features extracted from Deep Neural Networks (DNNs) have proven to be very effective in the context of Content Based Image Retrieval (CBIR). Recently, biologically inspired Hebbian learning algorithms have shown promises for DNN training. In this contribution, we study the performance of such algorithms in the development of feature extractors for CBIR tasks. Specifically, we consider a semi-supervised learning strategy in two steps: first, an unsupervised pre-training stage is performed using Hebbian learning on the image dataset; second, the network is fine-tuned using supervised Stochastic Gradient Descent (SGD) training. For the unsupervised pre-training stage, we explore the nonlinear Hebbian Principal Component Analysis (HPCA) learning rule. For the supervised fine-tuning stage, we assume sample efficiency scenarios, in which the amount of labeled samples is just a small fraction of the whole dataset. Our experi-

mental analysis, conducted on the CIFAR10 and CIFAR100 datasets, shows that, when few labeled samples are available, our Hebbian approach provides relevant improvements compared to various alternative methods.

3.2.5 AI and Computer Vision for Smart Cities

G. Amato, F. Carrara, L. Ciampi, M. Di Benedetto, C. Gennaro, F. Falchi, N. Messina, C. Vairo Italian Conference on ICT for Smart Cities And Communities(I-CiTies 2022). [2]

Artificial Intelligence (AI) is increasingly employed to develop public services that make life easier for citizens. In this abstract, we present some research topics and applications carried out by the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR of Pisa about the study and development of AI-based services for Smart Cities dedicated to the interaction with the physical world through the analysis of images gathered from city cameras. Like no other sensing mechanism, networks of city cameras can observe the world and simultaneously provide visual data to AI systems to extract relevant information and make/suggest decisions helping to solve many real-world problems. Specifically, we discuss some solutions in the context of smart mobility, parking monitoring, infrastructure management, and surveillance systems.

3.2.6 AIMH Lab for Trustworthy AI

N. Messina, F. Carrara, D. Coccomini, F. Falchi, C. Gennaro, G. Amato Ital-IA 2020 - Workshop su AI Responsabile ed Affidabile. [47]

In this short paper, we report the activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR related to Trustworthy AI. Artificial Intelligence is becoming more and more pervasive in our society, controlling recommendation systems in social platforms as well as safety-critical systems like autonomous vehicles. In order to be safe and trustworthy, these systems require to be easily interpretable and transparent. On the other hand, it is important to spot fake examples forged by malicious AI generative models to fool humans (through fake news or deep-fakes) or other AI systems (through adversarial examples). This is required to enforce an ethical use of these powerful new technologies. Driven by these concerns, this paper presents three crucial research directions contributing to the study and the development of techniques for reliable, resilient, and explainable deep learning methods. Namely, we report the laboratory activities on the detection of adversarial examples, the use of attentive models as a way towards explainable deep learning, and the detection of deepfakes in social platforms.

3.2.7 AIMH Lab for the Industry

F. Carrara, L. Ciampi, M. Di Benedetto, F. Falchi, C. Gennaro, F.V. Massoli, G. Amato Ital-IA 2022 - Workshop su AI per l'Industria, Online conference [47]

In this short paper, we report the activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR related to Industry. The massive digitalization affecting all the stages of product design, production, and control calls for data-driven algorithms helping in the coordination of humans, machines, and digital resources in Industry 4.0. In this context, we developed AI-based Computer-Vision technologies of general interest in the emergent digital paradigm of the fourth industrial revolution, fo-cusing on anomaly detection and object counting for computer-assisted testing and quality control. Moreover, in the automotive sector, we explore the use of virtual worlds to develop AI systems in otherwise practically unfeasible scenarios, showing an application for accident avoidance in self-driving car AI agents.

3.2.8 AIMH Lab: Smart Cameras for Public Administration

L. Ciampi, D. Cafarelli, F. Carrara, M. Di Benedetto, F. Falchi, C. Gennaro, F.V. Massoli, N. Messina, G. Amato Ital-IA 2022 - Workshop su AI per la Pubblica Amministrazione [56]

In this short paper, we report the activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR related to Public Administration. In particular, we present some AI-based public services serving the citizens that help achieve common goals beneficial to the society, putting humans at the epicenter. Through the automatic analysis of images gathered from city cameras, we provide AI applications ranging from smart parking and smart mobility to human activity monitoring.

3.2.9 AIMH Lab for Healthcare and Wellbeing

M. Di Benedetto, F. Carrara, L. Ciampi, F. Falchi, C. Gennaro, G. Amato Ital-IA 2022 - Workshop AI per la Medicina e la Salute [64]

In this work we report the activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR related to Healthcare and Wellbeing. By exploiting the advances of recent machine learning methods and the compute power of desktop and mobile platforms, we will show how artificial intelligence tools can be used to improve healthcare systems in various parts of disease treatment. In particular we will see how deep neural networks can assist doctors from diagnosis (e.g., cell counting, pupil and brain analysis) to communication to patients with Augmented Reality .

3.2.10 AIMH Lab for Cybersecurity

C. Vairo, D.A. Coccomini, F. Falchi, C. Gennaro, F.V. Massoli, N. Messina, G. Amato Ital-IA 2022 - Workshop su AI per Cybersecurity [13]

In this short paper, we report the activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR related to Cy-bersecurity. We discuss about our active research fields, their applications and challenges. We focus on face recognition and detection of adversarial examples and deep fakes. We also present our activities on the detection of persuasion techniques combining image and text analysis.

3.2.11 Approximate Nearest Neighbor Search on Standard Search Engines

F. Carrara, L. Vadicamo, C. Gennaro, G. Amato 15th International Conference on Similarity Search and Applications (SISAP 2022) [20]

Approximate search for high-dimensional vectors is commonly addressed using dedicated techniques often combined with hardware acceleration provided by GPUs, FPGAs, and other custom in-memory silicon. Despite their effectiveness, harmonizing those optimized solutions with other types of searches often poses technological difficulties. For example, to implement a combined text+image multimodal search, we are forced first to query the index of high-dimensional image descriptors and then filter the results based on the textual query or vice versa. This paper proposes a text surrogate technique to translate real-valued vectors into text and index them with a standard textual search engine such as Elasticsearch or Apache Lucene. This technique allows us to perform approximate kNN searches of high-dimensional vectors alongside classical full-text searches natively on a single textual search engine, enabling multimedia queries without sacrificing scalability. Our proposal exploits a combination of vector quantization and scalar quantization. We compared our approach to the existing literature in this field of research, demonstrating a significant improvement in performance through preliminary experimentation.

3.2.12
Counting or Localizing? Evaluating cell counting and detection in microscopy images.

L. Ciampi, F. Carrara, G. Amato, C. Gennaro. In 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022). [23]

Image-based automatic cell counting is an essential yet challenging task, crucial for the diagnosing of many diseases. Current solutions rely on Convolutional Neural Networks and provide astonishing results. However, their performance is often measured only considering counting errors, which can lead to masked mistaken estimations; a low counting error can be obtained with a high but equal number of false positives and false negatives. Consequently, it is hard to determine which solution truly performs best. In this work, we investigate three general counting approaches that have been successfully adopted in the literature for counting several different categories of objects. Through an experimental evaluation over three public collections of microscopy images containing marked cells, we assess not only their counting performance compared to several state-of-the-art methods but also their ability to correctly localize the counted cells. We show that commonly adopted counting metrics do not always agree with the localization performance of the tested models, and thus we suggest integrating the proposed evaluation protocol when developing novel cell counting solutions.

3.2.13
Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection

D.A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, G. Amato 1st International Workshop on Multimedia AI against Disinformation [30]

Deepfake Generation Techniques are evolving at a rapid pace, making it possible to create realistic manipulated images and videos and endangering the serenity of modern society. The continual emergence of new and varied techniques brings with it a further problem to be faced, namely the ability of deepfake detection models to update themselves promptly in order to be able to identify manipulations carried out using even the most recent methods. This is an extremely complex problem to solve, as training a model requires large amounts of data, which are difficult to obtain if the deepfake generation method is too recent. Moreover, continuously retraining a network would be unfeasible. In this paper, we ask ourselves if, among the various deep learning techniques, there is one that is able to generalise the concept of deepfake to such an extent that it does not remain tied to one or more specific deepfake generation methods used in the training set. We compared a Vision Transformer with an EfficientNetV2 on a cross-forgery context based on the ForgeryNet dataset. From our experiments, It emerges that EfficientNetV2 has a greater tendency to specialize often obtaining better results on training methods while Vision Transformers exhibit a superior generalization ability that makes them more competent even on images generated with new methodologies.

3.2.14
Diary of our initiatory journey on the continent of data citation in SSH

C. Concordia, E. Gray, N. Larrousse DH2022 - Digital Humanities 2022 responding to Asian diversity [60]

If citation is a common practice for publications, it is relatively new for data especially in SSH. This paper will present the work carried out during the SSHOC project about data citation in general and more precisely how to make them actionable. The metaphor of a travel journal of an expedition seemed appropriate to us to present this work carried out during the SSHOC project.

3.2.15
FastHebb: Scaling Hebbian Training of Deep Neural Networks to ImageNet Level

G. Lagani, C. Gennaro, H. Fassold, G. Amato 15th International Conference on Similarity Search and Applications (SISAP 2022) [59]

Learning algorithms for Deep Neural Networks are typically based on supervised end-to-end Stochastic Gradient Descent (SGD) training with error backpropagation (backprop). Backprop algorithms require a large number of labelled training samples to achieve high performance. However, in many realistic applications, even if there is plenty of image samples, very few of them are labelled, and semi-supervised sample-efficient training strategies have to be used. Hebbian learning represents a possible approach towards sample efficient training; however, in current solutions, it does not scale well to large datasets. In this paper, we present FastHebb, an efficient and scalable solution for Hebbian learning which achieves higher efficiency by 1) merging together update computation and aggregation over a batch of inputs, and 2) leveraging efficient matrix multiplication algorithms on GPU. We validate our approach on different computer vision benchmarks, in a semi-supervised learning scenario. FastHebb outperforms previous solutions by up to 50 times in terms of training speed, and notably, for the first time, we are able to bring Hebbian algorithms to ImageNet scale.

3.2.16
A Geographical Extension for NOnt Ontology

N. Pratelli Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries [83]

Digital Libraries (DLs) are full of narratives. Besides its richness, DLs services often retrieve narrative components, but not the narratives as a whole. To formally represent this knowledge and to create and visualize narratives the Digital Humanities group of ISTI-CNR has developed the Narrative Ontology (NOnt) and the Narrative Building and Visualising Tool (NBVT). In this context, my research aims to investigate the possibility to introduce the geospatial dimension of narratives in NOnt. Moreover, my research aims to extend the functionalities of NBVT to enrich the narratives with geospatial information. As a case study, I have chosen to create narratives about mountain ecosystems and economic value chains produced within the Mountain Valorization through Interconnectedness and Green Growth (MOVING) European project (2020-2023). Currently, my research is still at an early stage and I have started to conduct a state-of-the-art study of the geospatial and spatiotemporal RDF/S representation techniques. Eventually, I will evaluate the extension of NOnt.

3.2.17
Investigating binary partition power in metric query

R. Connor, A. Dearle, L. Vadicamo 30th Italian Symposium on Advanced Database Systems, SEBD [33]

It is generally understood that, as dimensionality increases, the minimum cost of metric query tends from $O(\log n)$ to $O(n)$ in both space and time, where n is the size of the data set. With low dimensionality, the former is easy to achieve; with very high dimensionality, the latter is inevitable. We previously described BitPart as a novel mechanism suitable for performing exact metric search in "high(er)" dimensions. The essential tradeoff of BitPart is that its space cost is linear with respect to the size of the data, but the actual space required for each object may be small as $\log_2 n$ bits, which allows even very large data sets to be queried using only main memory. Potentially the time cost still scales with $O(\log n)$. Together these attributes give exact search which outperforms indexing structures if dimensionality is within a certain range. In this article, we reiterate the design of BitPart in this context. The novel contribution is an in-depth examination of what the notion of "high(er)" means in practical terms. To do this we introduce the notion of exclusion power, *and show its application to some generated data sets across different dimensions.*

3.2.18
Learning to Detect Fallen People in Virtual Worlds

F. Carrara, L. Pasco, C. Gennaro, F. Falchi International Conference on Content-based Multimedia Indexing (CBMI). [19]

Falling is one of the most common causes of injury in all ages, especially in the elderly, where it is more frequent and severe. For this reason, a tool that can detect a fall in real time can be helpful in ensuring appropriate intervention and avoiding more serious damage. Some approaches available in the literature use sensors, wearable devices, or cameras with special features such as thermal or depth sensors. In this paper, we propose a Computer Vision deep-learning based approach for human fall detection based on largely available standard RGB cameras. A typical limitation of this kind of approaches is the lack of generalization to unseen environments. This is due to the error generated during human detection and, more generally, due to the unavailability of large-scale datasets that specialize in fall detection problems with different environments and fall types. In this work, we mitigate these limitations with a general-purpose object detector trained using a virtual world dataset in addition to real-world images. Through extensive experimental evaluation, we verified that by training our models on synthetic images as well, we were able to improve their ability to generalize. Code to reproduce results is available at https://github.com/lorepas/fallen-people-detection.

3.2.19
Linking different scientific digital libraries in Digital Humanities: the IMAGO case study

V. Bartalesi, N. Pratelli, E. Lenzi IRCDL 2022 - 18th Italian Research Conference on Digital Libraries, Padua, Italy, 24-25/02/2022 [6]

The geography of the world created during the Middle Ages and Renaissance (VI-XV centuries) was crucial to the development of Western thought in the European history. Until now, to the best of our knowledge, Medieval and Renaissance geographic Latin literature has not been studied using digital methods. The Italian National research project IMAGO - Index Medii Aevi Geographiae Operum - (2020-2023) aims at providing a systematic overview of this literature using Semantic Web technologies and the Linked Open Data paradigm. As the first step to develop tools to support scholars in creating, evolving and consulting a knowledge base (KB) of the geographic works, we created an OWL 2 DL ontology. To maximize its interoperability, we developed the ontology as an extension of two reference vocabularies: the CIDOC CRM and FRBRoo (including its in-progress reformulation LRMoo). In this paper, we briefly present the project, the ontology, and the automatic and semi-automatic tools we developed to populate it. The final aim of the project is the creation of a Web application allowing scholars to freely access and visualise the data collected in the IMAGO knowledge base.

3.2.20
MOBDrone: A Drone Video Dataset for Man OverBoard Rescue

D. Cafarelli, L. Ciampi, L. Vadicamo, C. Gennaro, A. Berton, M. Paterni, C. Benvenuti, M. Passera, F. Falchi International Conference on Image Analysis and Processing (ICIAP) [14]

Modern Unmanned Aerial Vehicles (UAV) equipped with cameras can play an essential role in speeding up the identification and rescue of people who have fallen overboard, i.e., man overboard (MOB). To this end, Artificial Intelligence techniques can be leveraged for the automatic understanding of visual data acquired from drones. However, detecting people at sea in aerial imagery is challenging primarily due to the lack of specialized annotated datasets for training and testing detectors for this task. To fill this gap, we introduce and publicly release the MOBDrone benchmark, a collection of more than 125K drone-view images in a marine environment *under several conditions, such as different altitudes, camera shooting angles, and illumination. We manually annotated more than 180K objects, of which about 113K man overboard, precisely localizing them with bounding boxes. Moreover, we conduct a thorough performance analysis of several state-of-the-art object detectors on the MOBDrone data, serving as baselines for further research.*

3.2.21 On pushing DeepFake Tweet Detection capabilities to the limits

M. Gambini, T. Fagni, F. Falchi, M. Tesconi 14th ACM Web Science Conference 2022 [52]

The recent advances in natural language generation provide an additional tool to manipulate public opinion on social media. Even though there has not been any report of malicious exploit of the newest generative techniques so far, disturbing human-like scholarly examples of GPT-2 and GPT-3 can be found on social media. Therefore, our paper investigates how the state-of-the-art deepfake social media text detectors perform at recognizing GPT-2 tweets as machine-written, also trying to improve the state-of-the-art by hyper-parameter tuning and ensembling the most promising detectors; finally, our work concentrates on studying the detectors' capabilities to generalize over tweets generated by the more sophisticated and complex evolution of GPT-2, that is GPT-3. Results demonstrate that hyper-parameter optimization and ensembling advance the state-of-the-art, especially on the detection of GPT-2 tweets. However, all tested detectors dramatically decreased their accuracy on GPT-3 tweets. Despite this, we found out that even though GPT-3 tweets are much closer to human-written tweets than the ones produced by GPT-2, they still have latent features in common share with other generative techniques like GPT-2, RNN and other older methods. All things considered, the research community should quickly devise methods to detect GPT-3 social media texts, as well as older generative methods.

3.2.22 On the Expected Exclusion Power of Binary Partitions for Metric Search

L. Vadicamo, A. Alan, R. Connor 15th International Conference on Similarity Search and Applications (SISAP 2022) [91]

The entire history and, we dare say, future of similarity search is governed by the underlying notion of partition. A partition is an equivalence relation defined over the space, therefore each element of the space is contained within precisely one of the equivalence classes of the partition. All attempts to search a finite space efficiently, whether exactly or approximately, rely on some set of principles which imply that if the query is within one equivalence class, then one or more other classes either cannot, or probably do not, contain any of its solutions. In most early research, partitions relied only on the metric postulates, and logarithmic search time could be obtained on low dimensional spaces. In these cases, it was straightforward to identify multiple partitions, each of which gave a relatively high probability of identifying subsets of the space which could not contain solutions. Over time the datasets being searched have become more complex, leading to higher dimensional spaces. It is now understood that even an approximate search in a very high-dimensional space is destined to require $\mathcal{O}(n)$ time and space. Almost entirely missing from the research literature however is any analysis of exactly when this effect takes over. In this paper, we make a start on tackling this important issue. Using a quantitative approach, we aim to shed some light on the notion of the exclusion power of partitions, in an attempt to better understand their nature with respect to increasing dimensionality.

3.2.23 Recurrent Vision Transformer for Solving Visual Reasoning Problems

N. Messina, G. Amato, F. Carrara, C. Gennaro, F. Falchi International Conference on Image Analysis and Processing (ICIAP) [68]

Although convolutional neural networks (CNNs) showed remarkable results in many vision tasks, they are still strained by simple yet challenging visual reasoning problems. Inspired by the recent success of the Transformer network in computer vision, in this paper, we introduce the Recurrent Vision Transformer (RViT) model. Thanks to the impact of recurrent connections and spatial attention in reasoning tasks, this network achieves competitive results on the same-different visual reasoning problems from the SVRT dataset. The weight-sharing both in spatial and depth dimensions regularizes the model, allowing it to learn using far fewer free parameters, using only 28k training samples. A comprehensive ablation study confirms the importance of a hybrid CNN + Transformer architecture and the role of the feedback connections, which iteratively refine the internal representation until a stable prediction is obtained. In the end, this study can lay the basis for a deeper understanding of the role of attention and recurrent connections for solving visual abstract reasoning tasks. The code for reproducing our results is publicly available here: `https://tinyurl.com/recvit`.

3.2.24 Reinforced Damage Minimization in Critical Events for Self-driving Vehicles

F. Merola, F. Falchi, C. Gennaro, M. Di Benedetto 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP-VISIGRAPP) [67]

Self-driving systems have recently received massive attention in both academic and industrial contexts, leading to major improvements in standard navigation scenarios typically identified as well-maintained urban routes. Critical events like road accidents or unexpected obstacles, however, require the execution of specific emergency actions that deviate from the ordinary driving behavior and are therefore harder to incorporate in the system. In this context, we propose a system that is specifically built to take control of the vehicle and perform an emergency maneuver in case of a dangerous scenario. The presented architecture is based on a deep reinforcement learning algorithm, trained in a simulated environment and using raw sensory data as input. We evaluate the system's performance on several typical pre-accident scenario and show promising results, with the vehicle being able to consistently perform an avoidance maneuver to nullify or minimize the incoming damage.

3.2.25
A Spatio-Temporal Attentive Network for Video-Based Crowd Counting

M. Avvenuti, M. Bongiovanni, L. Ciampi, F. Falchi, C. Gennaro, N. Messina IEEE Symposium on Computers and Communications (ISCC) [3]

Automatic people counting from images has recently drawn attention for urban monitoring in modern Smart Cities due to the ubiquity of surveillance camera networks. Current computer vision techniques rely on deep learning-based algorithms that estimate pedestrian densities in still, individual images. Only a bunch of works take advantage of temporal consistency in video sequences. In this work, we propose a spatio-temporal attentive neural network to estimate the number of pedestrians from surveillance videos. By taking advantage of the temporal correlation between consecutive frames, we lowered state-of-the-art count error by 5% and localization error by 7.5% on the widely-used FDST benchmark.

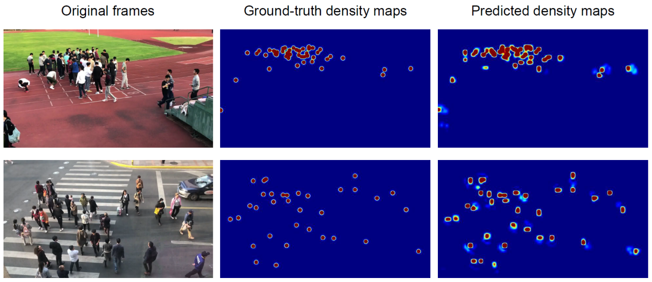

Figure 8. Some qualitative examples showing ground-truth and predicted density maps [3].

3.2.26
The SSH Data Citation Service, a tool to explore and collect citation metadata

C. Concordia, E. Gray, N. Larrousse 26th International Conference on Theory and Practice of Digital Libraries, TPDL [32]

This paper presents the SSH Data Citation Service (DCS), a software tool that provides functionalities to find, collect and analyse metadata related to digital objects, in particular datasets, referred to in citation strings. Starting from the citation string of a dataset, the DCS aggregates metadata related to the data from different sources: the repository hosting the dataset, PID Registration Agencies and Knowledge Graphs and gives a unified view of information about datasets coming from these sources.The DCS has been designed and developed in the SSHOC project as a tool to help investigate approaches adopted for data citation by Social Sciences and Humanities organisations managing data repositories, and as an utility to help data managers to create citation metadata.

3.2.27
A task category space for user-centric comparative multimedia search evaluations

J. Lokoč, W. Bailer, K. Barthel, C. Gurrin, S. Heller, B. Jónsson, L. Peška, L. Rossetto, K. Schoeffmann, L. Vadicamo, S. Vrochidis, J. Wu International Conference on Multimedia Modeling (MMM). [63]

In the last decade, user-centric video search competitions have facilitated the evolution of interactive video search systems. So far, these competitions focused on a small number of search task categories, with few attempts to change task category configurations. Based on our extensive experience with interactive video search contests, we have analyzed the spectrum of possible task categories and propose a list of individual axes that define a large space of possible task categories. Using this concept of category space, new user-centric video search competitions can be designed to benchmark video search systems from different perspectives. We further analyse the three task categories considered so far at the Video Browser Showdown and discuss possible (but sometimes challenging) shifts within the task category space.

3.2.28
Towards Unsupervised Machine Learning Approaches for Knowledge Graphs

F. Minutella, F. Falchi, P. Manghi, M. De Bonis, N. Messina. 18th Italian Research Conference on Digital Libraries (IRCDL). [72]

Nowadays, a lot of data is in the form of Knowledge Graphs aiming at representing information as a set of nodes and relationships between them. This paper proposes an efficient framework to create informative embeddings for node classification on large knowledge graphs. Such embeddings capture how a particular node of the graph interacts with his neighborhood and indicate if it is either isolated or part of a bigger clique. Since a homogeneous graph is necessary to perform this kind of analysis, the framework exploits the metapath approach to split the heterogeneous graph into multiple homogeneous graphs. The proposed pipeline includes an unsupervised attentive neural network to merge different metapaths and produce node embeddings suitable for classification. Preliminary experiments on the IMDb dataset demonstrate the validity of the proposed approach, which can defeat current state-of-the-art unsupervised methods.

3.2.29
Tuning Neural ODE Networks to Increase Adversarial Robustness in Image Forensics

R. Caldelli, F. Carrara, F. Falchi IEEE International Conference on Image Processing (ICIP). [16]

Although deep-learning-based solutions are pervading different application sectors, many doubts have arisen about their reliability and, above all, their security against threats that can mislead their decision mechanisms. In this work, we considered a particular kind of deep neural network, the Neural Ordinary Differential Equations (N-ODE) networks, which have shown intrinsic robustness against adversarial samples by properly tuning their tolerance parameter at test time. Their behaviour has never been investigated in image forensics tasks such as distinguishing between an original and an altered image. Following this direction, we demonstrate how tuning the tolerance parameter during the prediction phase can control and increase N-ODE's robustness versus adversarial attacks. We performed experiments on basic image transformations used to generate tampered data, providing encouraging results in terms of adversarial rejection and preservation of the correct classification of pristine *images.*

3.2.30
VISIONE at video browser showdown 2022

G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, C. Vairo International Conference on Multimedia Modeling (MMM). [1]

VISIONE is a content-based retrieval system that supports various search functionalities (text search, object/color-based search, semantic and visual similarity search, temporal search). It uses a full-text search engine as a search backend. In the latest version of our system, we modified the user interface, and we made some changes to the techniques used to analyze and search for videos.

3.2.31
Active Learning and the Saerens-Latinne-Decaestecker Algorithm: An Evaluation

A. Molinari, A. Esuli, F. Sebastiani 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE). [73]

The Saerens-Latinne-Decaestecker (SLD) algorithm is a method whose goal is improving the quality of the posterior probabilities (or simply "posteriors") returned by a probabilistic classifier in scenarios characterized by prior probability shift (PPS) between the training set and the unlabelled ("test") set. This is an important task, (a) because posteriors are of the utmost importance in downstream tasks such as, e.g., multiclass classification and cost-sensitive classification, and (b) because PPS is ubiquitous in many applications. In this paper we explore whether using SLD can indeed improve the quality of posteriors returned by a classifier trained via active learning (AL), a class of machine learning (ML) techniques that indeed tend to generate substantial PPS. Specifically, we target AL via relevance sampling (ALvRS) and AL via uncertainty sampling (ALvUS), two AL techniques that are very well-known especially because, due to their low computational cost, are suitable to being applied in scenarios characterized by large datasets. We present experimental results obtained on the RCV1-v2 dataset, showing that SLD fails to deliver better-quality posteriors with both ALvRS and ALvUS, thus contradicting previous findings in the literature, and that this is due not to the amount of PPS that these techniques generate, but to how the examples they prioritize for annotation are distributed.

3.2.32
Investigating Topic-Agnostic Features for Authorship Tasks in Spanish Political Speeches

S. Corbara, B. Chulvi Ferriols, P. Rosso, A. Moreo. NLDB [35]

Authorship Identification is the branch of authorship analysis concerned with uncovering the author of a written document. Methods devised for Authorship Identification typically employ stylometry (the analysis of unconscious traits that authors exhibit while writing), and are expected not to make inferences grounded on the topics the authors usually write about (as reflected in their past production). In this paper, we present a series of experiments evaluating the use of feature sets based on rhythmic and psycholinguistic patterns for Authorship Verification and Attribution in Spanish political language, via different approaches of text distortion used to actively mask the underlying topic. We feed these feature sets to a SVM learner, and show that they lead to results that are comparable to those obtained by the BETO transformer when the latter is trained on the original text, i.e., when potentially learning from topical information.

3.2.33
Rhythmic and Psycholinguistic Features for Authorship Tasks in the Spanish Parliament: Evaluation and Analysis

S. Corbara, B. Chulvi Ferriols, P. Rosso, A. Moreo. CLEF [34]

Among the many tasks of the authorship field, Authorship Identification aims at uncovering the author of a document, while Author Profiling focuses on the analysis of personal characteristics of the author(s), such as gender, age, etc. Methods devised for such tasks typically focus on the style of the writing, and are expected not to make inferences grounded on the topics that certain authors tend to write about. In this paper, we present a series of experiments evaluating the use of topic-agnostic feature sets for Authorship Identification and Author Profiling tasks in Spanish political language. In particular, we propose to employ features based on rhythmic and psycholinguistic patterns, obtained via different approaches of text masking that we use to actively mask the underlying topic. We feed these feature sets to a SVM learner, and show that they lead to results that are comparable to those obtained by a BETO transformer, when the latter is trained on the original text, i.e., potentially learning from topical information. Moreover, we further investigate the results for the different authors, showing that variations in performance are partially explainable in terms of the authors' political affiliation and communication style.

3.2.34
Ordinal quantification through regularization

M. Bunse, A. Moreo, F. Sebastiani, M. Senz. ECML/PKDD [12]

Quantification, i.e., the task of training predictors of the class prevalence values in sets of unlabelled data items, has received increased attention in recent years. However, most quantification research has concentrated on developing algorithms for binary and multiclass problems in which the classes are not ordered. We here study the ordinal case, i.e., the case in which a total order is defined on the set of $n > 2$ classes. We give three main contributions to this field. First, we create and make available two datasets for ordinal quantification (OQ) research that overcome the inadequacies of the previously available ones. Second, we experimentally compare the most important OQ algorithms proposed in the literature so far. To this end, we bring together algorithms that are proposed by authors from very different research fields, who were unaware of each other's developments. Third, we propose three OQ algorithms, based on the idea of preventing ordinally implausible estimates through regularization. Our experiments show that these algorithms outperform the existing ones if the ordinal plausibility assumption holds.

3.3 Magazines

In this section, we report the papers we have published in magazines.

3.3.1
Syllabic quantity: Rhythmic clues for Latin authorship

S. Corbara, A. Moreo, F. Sebastiani.
Information Matters. [36].

Techniques for uncovering who's behind anonymous identities, and unmasking forged ones, are known as "Authorship Identification"(AID, for short), which are part of the field of "Authorship Analysis", whose goal is to infer characteristics (such as the gender, the age, or the native language) of the authors of written documents.

3.4 Editorials

In this section, we report proceedings and books for which we have acted as editors.

3.4.1
CLEF 2022

A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings. Lecture Notes in Computer Science 13390, Springer 2022. [5].

3.4.2
SISAP 2022

D. Hiemstra, M.F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (eds.), Similarity Search and Applications - 15th International Conference, SISAP 2022, Bologna, Italy, October 5-7, 2022, Proceedings, Lecture Notes in Computer Science 13590, Springer, 2022. [88].

3.4.3
LQ2022 @ ECML/PKDD 2022

J.J. del Coz, P. González, A. Moreo, F. Sebastiani (eds.), Proceedings of the 2nd International Workshop on Learning to Quantify (LQ 2022), Co-located with ECML/PKDD 2022, Grenoble, France, September 23, 2022. [39].

3.5 Preprints

In this section, we report the papers published in preprint form on publicly accessible archives, in alphabetic order by first author.

3.5.1
Transformer-Based Multi-modal Proposal and Re-Rank for Wikipedia Image-Caption Matching.

N. Messina, D.A. Coccomini, A. Esuli, F. Falchi arXiv:2206.10436. [69]

With the increased accessibility of web and online encyclopedias, the amount of data to manage is constantly increasing. In Wikipedia, for example, there are millions of pages written in multiple languages. These pages contain images that often lack the textual context, remaining conceptually floating and therefore harder to find and manage. In this work, we present the system we designed for participating in the Wikipedia Image-Caption Matching challenge on Kaggle, whose objective is to use data associated with images (URLs and visual data) to find the correct caption among a large pool of available ones. A system able to perform this task would improve the accessibility and completeness of multimedia content on large online encyclopedias. Specifically, we propose a cascade of two models, both powered by the recent Transformer model, able to efficiently and effectively infer a relevance score between the query image data and the captions. We verify through extensive experimentation that the proposed two-model approach is an effective way to handle a large pool of images and captions while maintaining bounded the overall computational complexity at inference time. Our approach achieves remarkable results, obtaining a normalized Discounted Cumulative Gain (nDCG) value of 0.53 on the private leaderboard of the Kaggle challenge.

3.5.2
Deep learning for structural health monitoring: An application to heritage structures

F. Carrara, F. Falchi, M. Girardi, N. Messina, C. Padovani, D. Pellegrini arXiv:2211.10351. [18]

Thanks to recent advancements in numerical methods, computer power, and monitoring technology, seismic ambient noise provides precious information about the structural behavior of old buildings. The measurement of the vibrations produced by anthropic and environmental sources and their use for dynamic identification and structural health monitoring of buildings initiated an emerging, cross-disciplinary field engaging seismologists, engineers, mathematicians, and computer scientists. In this work, we employ recent deep learning techniques for time-series forecasting to inspect and detect anomalies in the large dataset recorded during a long-term monitoring campaign conducted on the San Frediano bell tower in Lucca. We frame the problem as an unsupervised anomaly detection task and train a Temporal Fusion Transformer to learn the normal dynamics of the structure. We then detect the anomalies by looking at the differences between the predicted and observed frequencies.

3.5.3
MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection

D.A. Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, Claudio Gennaro arXiv:arxiv.2211.10996. [31]

In this paper, we introduce MINTIME, a video deepfake detection approach that captures spatial and temporal anomalies and handles instances of multiple people in the same video and variations in face sizes. Previous approaches disregard such information either by using simple a-posteriori aggregation schemes, i.e., average or max operation, or using only one identity for the inference, i.e., the largest one. On the contrary, the proposed approach builds on a Spatio-Temporal TimeSformer combined with a Convolutional Neural Network backbone to capture spatio-temporal anomalies from the face sequences of multiple identities depicted in a video. This is achieved through an Identity-aware Attention mechanism that attends to each face sequence independently based on a masking operation and facilitates video-level aggregation. In addition, two novel embeddings are employed: (i) the Temporal Coherent Positional Embedding that encodes each face sequence's temporal information *and (ii) the Size Embedding that encodes the size of the faces as a ratio to the video frame size. These extensions allow our system to adapt particularly well in the wild by learning how to aggregate information of multiple identities, which is usually disregarded by other methods in the literature. It achieves state-of-the-art results on the ForgeryNet dataset with an improvement of up to 14% AUC in videos containing multiple people and demonstrates ample generalization capabilities in cross-forgery and cross-dataset settings.*

3.5.4
Unravelling Interlanguage Facts via Explainable Machine Learning

B. Berti, A. Esuli, F. Sebastiani arXiv:2208.01468. [9]

Native language identification (NLI) is the task of training (via supervised machine learning) a classifier that guesses the native language of the author of a text. This task has been extensively researched in the last decade, and the performance of NLI systems has steadily improved over the years. We focus on a different facet of the NLI task, i.e., that of analysing the internals of an NLI classifier trained by an explainable *machine learning algorithm, in order to obtain explanations of its classification decisions, with the ultimate goal of gaining insight into which linguistic phenomena "give a speaker's native language away". We use this perspective in order to tackle both NLI and a (much less researched) companion task, i.e., guessing whether a text has been written by a native or a non-native speaker. Using three datasets of different provenance (two datasets of English learners' essays and a dataset of social media posts), we investigate which kind of linguistic traits (lexical, morphological, syntactic, and statistical) are most effective for solving our two tasks, namely, are most indicative of a speaker's L1. We also present two case studies, one on Spanish and one on Italian learners of English, in which we analyse individual linguistic traits that the classifiers have singled out as most important for spotting these L1s. Overall, our study shows that the use of explainable machine learning can be a valuable tool for the scholar who investigates interlanguage facts and language transfer.*

3.5.5
Multi-Label Quantification

A. Moreo, M. Francisco, F. Sebastiani arXiv:2211.08063. [74]

Quantification, variously called "supervised prevalence estimation" or "learning to quantify", is the supervised learning task of generating predictors of the relative frequencies (a.k.a. "prevalence values") of the classes of interest in unlabelled data samples. While many quantification methods have been proposed in the past for binary problems and, to a lesser extent, single-label multiclass problems, the multi-label setting (i.e., the scenario in which the classes of interest are not mutually exclusive) remains by and large unexplored. A straightforward solution to the multi-label quantification problem could simply consist of recasting the problem as a set of independent binary quantification problems. Such a solution is simple but naive, since the independence assumption upon which it rests is, in most cases, not satisfied. In these cases, knowing the relative frequency of one class could be of help in determining the prevalence of other related classes. We propose the first truly multi-label quantification methods, i.e., methods for inferring estimators of class prevalence values that strive to leverage the stochastic dependencies among the classes of interest in order to predict their relative frequencies more accurately. We show empirical evidence that natively multi-label solutions outperform the naive approaches by a large margin. The code to reproduce all our experiments is available online.

4. Dissertations

4.1 PhD Thesis

4.1.1
Deep Learning Techniques for Visual Counting

Ciampi Luca, PhD Information Science, University of Pisa, 2022 [21].

In this thesis, I investigated and enhanced Deep Learning (DL)-based techniques for the visual counting task, which automatically estimates the number of objects, such as people or vehicles, present in images and videos. Specifically, I tackled the problem related to the lack of data needed for training current DL-based solutions by exploiting synthetic data gathered from video games, employing Domain Adaptation strategies between different data distributions, and taking advantage of the redundant information characterizing datasets labeled by multiple annotators. Furthermore, I addressed the engineering challenges coming out of the adoption of DL-based techniques in environments with limited power resources, mainly due to the high computational budget the AI-based algorithms require.

4.1.2
Relational Learning in Computer Vision

Messina Nicola, PhD Information Science, University of Pisa, 2022 [22].

This thesis tackles important shortcomings of recent Deep Learning technologies in processing image and text data in a relational and joint manner. In fact, although deep neural networks obtained impressive results on many tasks, they cannot perform non-local processing by explicitly relating potentially interconnected visual or textual entities. In this work, we introduce a challenging variant of the Content-Based Image Retrieval (CBIR) task, called Relational CBIR. In R-CBIR, we aim to retrieve images also having similar relationships among the multiple objects present in the images. Then, we move a step further, considering real-world images and focusing on cross-modal visual-textual retrieval. We use the Transformer Encoder, a recently introduced module that relies on the power of self-attention, to relate different sentence words and image regions, with large-scale retrieval as the main goal. We show that the obtained features contain very high-level semantics and defeat current image descriptors on the challenging Semantic CBIR task. We then propose some solutions for scaling the search to possibly millions of images or texts. In the end, we deploy the developed networks in a large-scale interactive video retrieval software, called VISIONE, developed in our laboratory. Sticking to the multi-modal Transformer framework, we tackle another critical task in the modern Internet: detecting persuasion techniques in memes spread on social networks during disinformation campaigns. Finally, we probe current state-of-the-art CNNs on challenging visual reasoning benchmarks requiring non-local spatial comparisons. After understanding the drawbacks

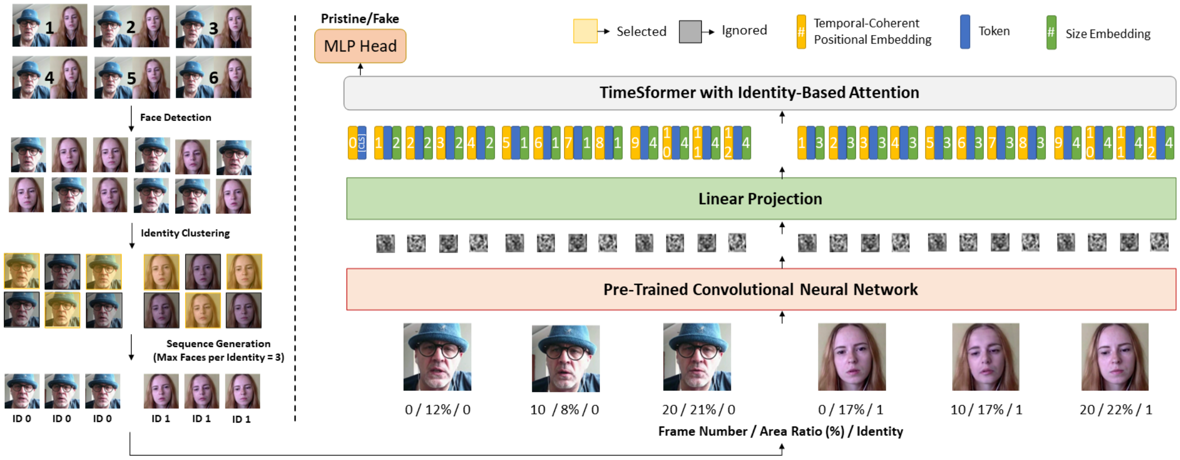

Figure 9. Overview of the MINTIME architecture [31].

of CNNs on these tasks, we propose a hybrid CNN-Transformer architecture, constraining the model complexity and reaching higher data efficiency.

4.2 Master of Science Dissertations

4.2.1
AI in the Loop in Animal Robot Interaction

Edoardo Fazzari, Master of Science in Artificial Intelligence and Data Engineering, 2022. [48]. Advisors: F. Falchi, C. Stefanini, D. Romano, F. Carrara

Animal-robot interaction is an interesting topic that have emerged very recently and it is still under development, thus offering many areas of research that has not yet been investigated. The main contribution of this thesis is to introduce artificial intelligence methods in the loop in animal-robot interaction. Specifically, this thesis deals with the classification of the interaction that occurred between an insect (Orthoptera: Ensifera) and two chemical agents (sugar and ammonia) through the readings of the animal's antennae. This was carried out in two steps: first, a deep neural network was developed and trained, using state-of-the art computer vision techniques, to locate the position of specific parts of the cricket in order to generate reliable numeric sequences; subsequently, these obtained sequences were used to identify the type of interaction that happened, using one-dimensional convolutional deep learning techniques.

4.2.2
Cross-modal learning for sentiment analysis of social media images

Alessio Serra, Master of Science in Artificial Intelligence and Data Engineering, 2022. [87]. Advisors: F. Falchi, M. Tesconi, M. Avvenuti, F. Carrara, M. Gambini

In this thesis is presented a broad overview of the Sentiment Analysis problem in the machine learning field, both for textual and visual media, with a focus on all techniques and methods used for the experimental part. In the latter, a big Twitter Visual Sentiment Analysis dataset was built crawling about 3.5 images from the social media for three months. This was achieved without the need for a human annotator, thus minimizing the effort required and allowing for the creation of a huge data set. The cross-modal learning approach used confirmed that, even if the textual information associated to images is often noisy and ambiguous, it can still be useful to build a reliable dataset, whose size is limited only by the number of images available. This large dataset can help the future research to train robust visual models and its size would be particularly advantageous, as the number of parameters of current SOTA models is exponentially growing along with their need of data to avoid overfitting problems. The effectiveness of the T4SA 2.0 was tested fine-tuning the Vision-Transformer model, which achieved incredible results on other manual annotated visual datasets, even beating the current State Of The Art.

4.2.3
Action Detection in Rugby Video Sequences in both Professional and Amateur Settings

Federica Baldi, Master of Science in Artificial Intelligence and Data Engineering, 2022. [4]. Advisors: F. Falchi, C. Gennaro, L. Marchesotti

Among the several problems that are part of Computer Vision and, more specifically, Video Understanding, we find action detection, i.e., the task of classifying and temporally locating actions that appear in a video. This is a widespread field of research at the moment, probably also because of the countless real-world applications that could benefit from it. In recent years and owing mainly to the advent of new Deep Learning technologies, exceptional progress has been made. However, as it is currently almost utopian to be able to devise a generic architecture that works well regardless of the context, there is still a lot of research to be done in the different application areas. One of the most popular yet challenging one is the analysis of sports videos. Within this context, the objective of this thesis is to address the problem of automatic action detection in rugby video sequences, filmed in both professional and amateur settings. So far, and to the best of our knowledge, research in the specific use case of rugby is lagging far behind that of other sports, and there is neither a Machine Learning method nor datasets published for this purpose. Therefore, first, a dataset was created by collecting and annotating rugby matches at various levels of professionalism. The intent is to make the dataset publicly available so as to jumpstart research in this area as well. Then, and after careful inspection of *the state of the art, experiments on the new dataset were performed by adapting one of the existing approaches in the literature. The approach, originally designed for soccer, allowed us to create a baseline on the new dataset and evaluate the differences between the performance obtained on the professional segment and that obtained on the amateur one. Finally, after analyzing the results, possible future lines of research were proposed.*

4.2.4
AI-Assisted 3D Human Head Keypoint Matching for Image Pair Alignment

Lorenzo Biondi. Master of Science in Artificial Intelligence and Data Engineering, 2022. [10]. Advisors: F. Fabrizio, M. Di Benedetto, N. Tonellotto, F. Carrara

This thesis work is about identifying, through a Computer Vision approach, specific keypoints within MR ("Magnetic Resonance") volumes of human heads. Such keypoints can be exploited to align the volumes with other targets, such as the Augmented Reality representation of the users' heads or other scanned images. The first phase of the study focused on the identification of training and test datasets, the selection of four non-coplanar points of interest for the alignment, and the identification of the most appropriate neural network architecture to individuate such points. In particular, different 3D CNN ("Convolutional Neural Network") architectures were tested, with different sets of hyperparameters and exploiting multiple volume sampling frequencies. Moreover, we designed a customized data augmentation procedure in order to enlarge the available training samples and better the network generalization capabilities. Once obtained a model with sufficient performances for the alignment, the following steps were the implementation of the prediction phase in JavaScript and, finally, the image-to-image alignment. The dataset collection and keypoints annotation, together with the final model, are an evident contribution of this thesis work to the state of the art.

4.2.5
Design, implementation, and test of tools for multi-camera vehicle tracking based on artificial intelligence

Gaetano Emanuele Valenti. Master of Science in Artificial Intelligence and Data Engineering, 2022. [92]. Advisors: G. Claudio, F. Falchi, N. Messina, L. Ciampi

With the continuous expansion of the city scale, the management of city has become more and more challenging. Thanks to the development of computer vision technology and surveillance network throughout the city, there are many new options for city management, especially in traffic management. Among them, multi-camera vehicle tracking is one of the important tasks. It aims to track the vehicles over large areas in multiple surveillance camera networks. The latter is a truly complex task given the high variability of possible scenarios in cities, both urban and suburban. The difficulty is increased considering the various possible weather conditions, vehicles that are very similar to each other, observed from different perspectives and with frequent lighting variations. Such problems are currently difficult to solve given the unavailability of enough data. Aiming to fill this gap and thus enable the handling of increasingly challenging scenarios, this thesis work aims to design, implement and validate a tool for data extraction by exploiting video game that provides a very realistic simulation of reality and allows totally arbitrary data extraction that can be used in multiple use cases. Such a tool is subsequently used for the extraction of a synthetic dataset with the objective of fine tuning and improving the performance of a well-known challenge presented by NVIDIA AI City Challenge. The latter is focused entirely in the application of AI to improve the efficiency of operations in city environments. The previously extracted synthetic dataset has been used to fine-tune a model to perform feature extraction for one of the most important step of multi-camera tracking, the "re-identification" module.It has the objective to distinguish each individual entity, in this case vehicles, uniquely and consequently extract highly discriminative features. This fine-tuning process has contributed to an improvement in the MOT's main classification metric, the IDF1.This result opens up a wide range of opportunities in the use of the tool made with the possibility of using such data independently or in support of real data.

4.2.6

Development of deep learning models for fight detection in videos

Gianluca Serao. Master of Science in Artificial Intelligence and Data Engineering, 2022. [86]. Advisors: G. Claudio, G. Amato, L. Ciampi, N. Messina

Video action recognition has gained much attention in recent years by the research community for its importance in many everyday applications such as human-machine interactions and surveillance. Video action recognition aims to classify human actions using a sequence of still images. This task is not easy: different actions may share similar patterns, making them difficult to distinguish. Moreover, videos pose several issues also from the appearance point of view, like camera motion and illuminance changes. Therefore, it is crucial to jointly model motion and appearance information in this context. Early methods were mainly based on handcrafted features and motion estimation techniques. Afterwards, with the rise of deep learning, most methods were based on deep networks, such as Convolutional Neural Networks (CNN) and Residual Networks (ResNet). Initially, the most successful methods used simultaneously multiple networks based on 2D convolution to model the temporal dimension. Lately, networks based on 3D convolution have gained popularity as they can model the temporal and appearance information at once. In this thesis work, we develop a tool that can detect violent actions, mainly composed of fights, within videos. To this aim, we use different state-of-the-art ResNets based on 3D convolution or its approximation. Along with these networks, we use other computer vision techniques such as optical flow and ad-hoc video sampling strategies. We train these architectures on several datasets, using different kinds of input: plain and optical flow videos. Finally, we test the models on datasets never seen before to evaluate the generalisation capabilities and compare them. Tests show that our methods reach state-of-the-art performances with all the datasets.

4.2.7
Design and implementation of a system for automatic pedestrian counting in video streams based on artificial intelligence techniques

Marco Bongiovanni. Master of Science in Artificial Intelligence and Data Engineering, 2022. [11]. Advisors: G. Claudio, F. Falchi, N. Messina, L. Ciampi

Crowd counting aims to estimate the total number of people in images or videos. In recent years, it has become a hot topic in computer vision, thanks to its several real-world applications, such as in public safety and disaster management. However, the crowd counting task involves some difficult challenges: people density not being uniform in the scene, inter-object occlusion, changes in light conditions or perspective between different scenes. Most of the existing approaches use Convolutional Neural Networks (CNN) to estimate people density maps from static input images. Even though crowded video sequences are usually available, only very few methods proposed in the literature take advantage of the temporal correlation between neighbor frames in the same video. In this thesis, in order to exploit temporal consistency, we regress people flows between consecutive frames instead of considering frames as independent images and estimating people density within them. Moreover, we exploit some state-of-the-art self-attention mechanisms, such as TimeSformer and Self-Attention convolution, to improve the counting capabilities. The proposed architecture performs better than the benchmark architecture in the video-based crowd counting literature. Furthermore, by testing both networks on datasets they had never seen before, our network also proves to have better generalization capability.

4.2.8
Design and Implementation of Facial Expression Recognition System

Stefano Poleggi. Master of Science in Artificial Intelligence and Data Engineering, 2022. [82]. Advisors: F. Falchi, D. Cafarelli, C. Gennaro, G. Amato

Development of a facial expression recognition system capable of predicting emotions based on a dimensional approach (Valence-Arousal). For the development of the neural network on which the system is based, a training technique was employed that is capable of training robust networks as the resolution of the input images decreases.

4.2.9
Design and implementation of a deep learning system for knowledge graph analysis

Filippo Minutella, Master of Science in Artificial Intelligence and Data Engineering, 2022. [71]. Advisors: F. Falchi, P. Manghi, M. De Bonis, N. Messina

Nowadays a lot of data is in the form of Knowledge Graphs, i.e. a set of nodes and relationships between them. Many companies exclude relationships or don't use them to their full potential in order to convert naturally graph-like data into tabular data so that it can be organized in the usual databases and analyzed using simple, familiar processes. This conversion process has the advantage of simplification but brings with it a loss of information that cannot always be ignored. After a review of techniques aimed at performing different tasks on graph data types, some of these were used in the analysis of the data provided by OpenAIRE. OpenAIRE is a platform to support Open Science in Europe and it provides a Research Graph, which is a graph composed of scientific resources linked to their authors, where they have been published, and the keywords in them. For the analysis of the Research Graph, it has been used a metapath approach in order to allow the analysis of a heterogeneous graph by transforming it into a series of homogeneous graphs. Such graphs are simpler to be analyzed and they allow to focus the analysis on a single type of element of the graph. A framework was developed to analyze the Research Graph and to highlight the anomalies in the dataset. The framework integrates the metapath approach and a neural network to perform Node Classification and Node Embedding, and the results were compared with the methods of Graph Neural Networks in the literature. The result of our work is a method that can leverage the node attributes and graph metapaths to perform Node Classification or Node Embedding by identifying the most significant information. The result of the work presented in this thesis is a framework that is scalable, easy to understand and fast. Moreover, it performs better than other unsupervised methods available in the literature.

5. Resources

In this section, we report contributions of AIMH having to do with the creation of datasets (Section 5.1), the publication of code (Section 5.2), and the design of shared tasks (Section 5.3)

5.1 Datasets

5.1.1
Bus Violence: a large-scale benchmark for video violence detection in public transport

P. Foszner, M. Staniszewski, A. Szczesna, M. Cogiel, D. Golba, L. Ciampi, N. Messina, C. Gennaro, F. Falchi, G. Amato, G. Serao [49]

The Bus Violence dataset is a large-scale collection of videos depicting violent and non-violent situations in public transport environments. This benchmark was gathered from multiple cameras located inside a moving bus where several people simulated violent actions, such as stealing an object from another person, fighting between passengers, etc. It contains 1,400 video clips manually annotated as having or not violent scenes, making it one of the biggest benchmarks for video violence detection in the literature. Specifically, videos are recorded from three cameras at 25 Frames Per Second (FPS) - two cameras located in the corners of the bus (with resolution 960x540 px) and one fisheye in the middle (1280x960 px). The clips have a minimum length of 16 frames and a maximum of 48 frames, capturing a very precise action (either violence or non-violence). The dataset is perfectly balanced, containing 700 videos of violence and 700 videos of non-violence. The dataset is freely available at `https://zenodo.org/record/7044203`.

5.1.2
Crowd simulation (CrowdSim2) for tracking and object detection

P. Foszner, A. Szczesna, A. Cygan, B. Bizon, M. Cogiel, D. Golba, L. Ciampi, N. Messina E. Macioszek, M. Staniszewski

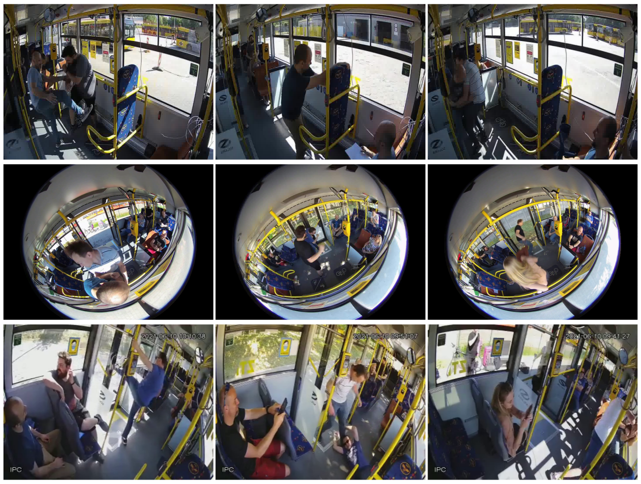

Figure 10. Samples from the Bus Vioelnce dataset [49].

[89]

CrowdSim2 is a crowd simulation tool designed in Unity for the purpose of generating massive synthetic data. Such generated data from crowd simulation enables the validation of various methods in terms of tracking multiple people and detecting objects (specifically, pedestrians and cars). The dataset is freely available at `https://zenodo.org/record/7262220`.

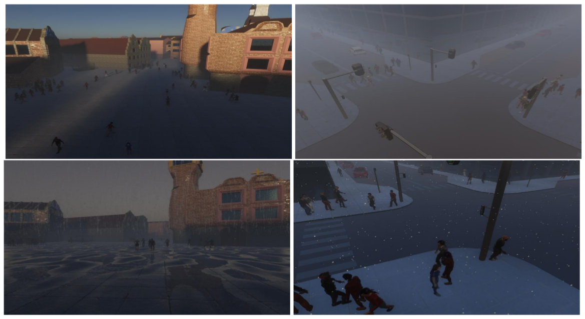

Figure 11. Samples from the CrowdSim2 dataset [89].

5.1.3
MOBDrone: a large-scale drone-view dataset for man overboard detection

D. Cafarelli, L. Ciampi, L. Vadicamo, C. Gennaro, A. Berton, M. Paterni, C. Benvenuti, M. Passera, F. Falchi [15].

The Man OverBoard Drone (MOBDrone) dataset is a large-scale collection of aerial footage images. It contains 126,170 frames extracted from 66 video clips gathered from one UAV flying at an altitude of 10 to 60 meters above the mean sea level. Images are manually annotated with more than 180K bounding boxes localizing objects belonging to 5 categories — person, boat, lifebuoy, surfboard, wood. More than 113K of these bounding boxes belong to the person category and localize people in the water simulating the need to be rescued. The dataset is freely available at `https://zenodo.org/record/5996890`.

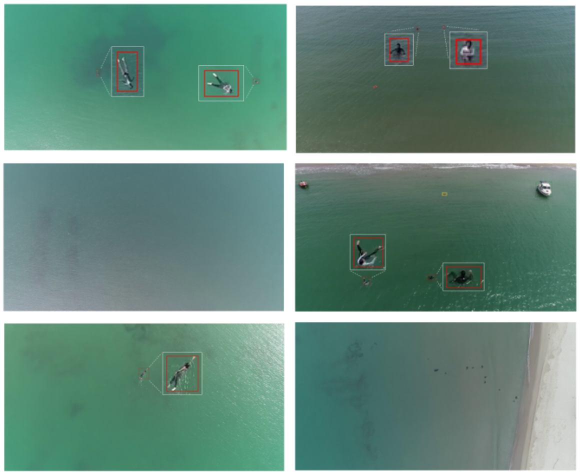

Figure 12. MOBDrone dataset examples [15].

5.1.4
PNN - A Multi-Rater Benchmark for Perineuronal Nets Detection and Counting in Fluorescence Microscopy Images

L. Ciampi, F. Carrara, V. Totaro, R. Mazziotti, L. Lupori, C. Santiago, G. Amato, T. Pizzorusso, C. Gennaro [24].

PNN is a dataset of fluorescence microscopy images of mice brain slices stained against perineuronal nets (PNNs). PNNs are dot-annotated by experts for evaluating cell detection and counting. The dataset is composed of two subsets: a large single-rater subset (PNN-SR) and a smaller multi-rater subset (PNN-MR). The annotation procedure in PNN-MR has been performed by seven different raters, and all the annotations of each rater are reported. This enables modeling the confidence of a given detection using the agreement between raters on that sample as training signal. The dataset is freely available at `https://doi.org/10.5281/zenodo.5567032`.

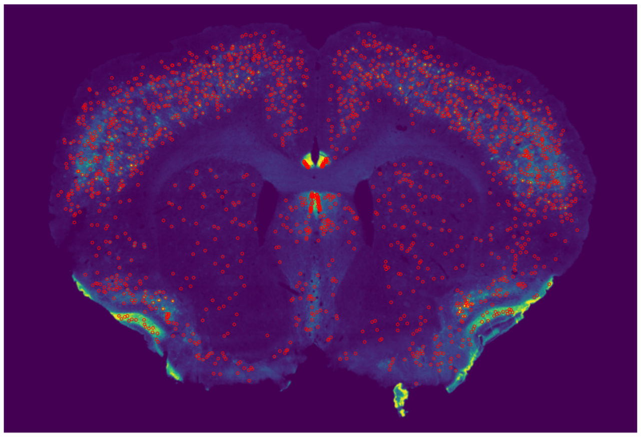

Figure 13. A sample from the single-rater subset (PNN-SR) with dot annotations in red. [24].

5.1.5
VWFP: Virtual World Fallen People Dataset for Visual Fallen People Detection

F. Carrara, L. Pasco, C. Gennaro, F. Falchi [81].

A synthetic dataset for visual fallen people detection comprising images extracted from the highly photo-realistic video game Grand Theft Auto V developed by Rockstar North. Each image is labeled by the game engine providing bounding boxes and statuses (fallen or non-fallen) of people present in the scene. The dataset comprises 6,071 synthetic images depicting 7,456 fallen and 26,125 non-fallen pedestrian instances in various looks, camera positions, background scenes, lightning, and occlusion conditions. The dataset is freely available at `https://zenodo.org/record/6394684`.

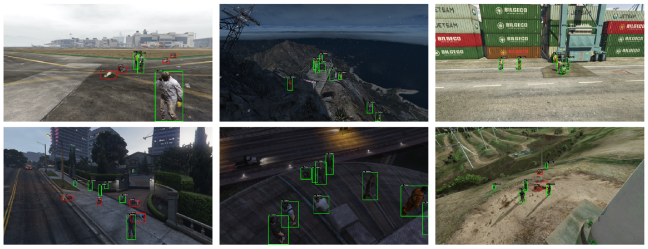

Figure 14. Samples from the Virtual World Fallen People (VWFP) Dataset. Green and red bounding boxes represent non-fallen and fallen people, respectively [81].

5.1.6
Product Reviews for Ordinal Quantification

M. Bunse, A. Moreo, F. Sebastiani, M. Senz. [12].

This data set comprises a labeled training set, validation samples, and testing samples for ordinal quantification. It appears in our research paper "Ordinal Quantification Through Regularization", which we have published at ECML-PKDD 2022. The data is extracted from the McAuley data set of product reviews in Amazon, where the goal is to predict the 5-star rating of each textual review. We have sampled this data according to two protocols that are suited for quantification research. The goal of quantification is not to predict the star rating of each individual instance, but the distribution of ratings in sets of textual reviews. More generally speaking, quantification aims at estimating the distribution of labels in unlabeled samples of data. The first protocol is the artificial prevalence protocol (APP), where all possible distributions of labels are drawn with an equal probability. The second protocol, APP-OQ, is a variant thereof, where only the smoothest 20% of all APP samples are considered. This variant is targeted at ordinal quantification, where classes are ordered and a similarity of neighboring classes can be assumed. 5-star ratings of product reviews lie on an ordinal scale and, hence, pose such an ordinal quantification task. This data set comprises two representations of the McAuley data. The first representation consists of TF-IDF features. The second representation is a RoBERTa embedding. This second representation is dense, while the first is sparse. In our experience, logistic regression classifiers work well with both representations. RoBERTa embeddings yield more accurate predictors than the TF-IDF features.
The dataset is freely available at `https://zenodo.org/record/7081208#.ZAhcf9LMKiA`.

5.1.7
Cherenkov Telescope Data for Ordinal Quantification

M. Bunse, A. Moreo, F. Sebastiani, M. Senz. [12].

This labeled data set is targeted at ordinal quantification. It appears in our research paper "Ordinal Quantification Through Regularization", which we have published at ECML-PKDD 2022. The goal of quantification is not to predict the label of each individual instance, but the distribution of labels in unlabeled sets of data. With the scripts provided, you can extract the relevant features and labels from the public data set of the FACT Cherenkov telescope. These features are precisely the ones that domain experts from astro-particle physics employ in their analyses. The labels stem from a binning of a continuous energy label, which is common practice in these analyses. Our first protocol is the artificial prevalence protocol (APP), where all possible distributions of labels are drawn with an equal probability. The second protocol, APP-OQ, is a variant thereof, where only the smoothest 20% of all APP samples are considered. This variant is targeted at ordinal quantification tasks, where classes are ordered and a similarity of neighboring classes can be assumed. The labels of the FACT data lie on an ordinal scale and, hence, pose such an ordinal quantification task.
The dataset is freely available at `https://zenodo.org/record/7090095#.ZAhcgNLMKiA`.

5.1.8
Two Datasets for the Computational Authorship Analysis of Medieval Latin Texts

S. Corbara, A. Moreo, F. Sebastiani, M. Tavoni [38].

We present and make available MedLatinEpi and MedLatinLit, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatinEpi and MedLatinLit consist of 294 and 30 curated texts, respectively, labelled by author; MedLatinEpi texts are of epistolary nature, while MedLatinLit texts consist of literary comments and treatises about various subjects. As such, these two datasets lend themselves to supporting research in authorship analysis tasks, such as authorship attribution, authorship verification, or same-author verification. In a related paper, along with the datasets we provide experimental results, obtained on these datasets, for the authorship verification task, i.e., the task of predicting whether a text of unknown authorship was written by a candidate author or not.
The dataset is freely available at `https://zenodo.org/record/4298503#.ZAhde9LMKiA`.

5.2 Code

5.2.1
Bus violence: an open benchmark for video violence detection on public transport

L. Ciampi, P. Foszner, N. Messina, M. Staniszewski, C. Gennaro, F. Falchi, G. Serao, M. Cogiel, D. Golba, A. Szczesna, G. Amato [26].

Code for replicating the experiments in [26]. The provided code tests some state-of-the-art video violence detectors over the new Bus Violence benchmark we introduced in the paper. `https://github.com/ciampluca/bus-violence-benchmark-eval`

5.2.2
Counting or Localizing? Evaluating cell counting and detection in microscopy images

L. Ciampi, F. Carrara, G. Amato, C. Gennaro [23]

Code for replicating the experiments in [23]. The provided code trains three commonly adopted cell counting approaches based on detection, density estimation, and segmentation. Through an experimental evaluation over some public collection of microscopy images containing marked cells we show that these counting strategies do not always agree with counting and localization performance. `https://github.com/ciampluca/counting_perineuronal_nets/tree/visapp-counting-cells`

5.2.3
HDN Annotation Tool

V. Bartalesi, N. Pratelli, D. Metilli, C. Meghini [8]

To facilitate the process of populating the ontology developed within the Hypermedia Dante Network (HDN) project (PRIN 2020-2023), we implemented a semi-automatic tool called HDN Annotation Tool. The tool supports scholars to build a knowledge base of the primary sources of Dante Alighieri's Divine Comedy. The tool was developed using a Python backend with the Django framework, and a frontend built with HTML5, JavaScript, and the Bootstrap library. It takes as input the JSON file, where the knowledge automatically extracted from the corpus of the Dartmouth Dante Project (DDP) is stored and shows the relevant information in the corresponding fields of the tool interface. After analyzing the commentaries of the DDP, scholars use the interface of the tool to insert knowledge about primary sources. The tool is accessible through the HDN-Lab, which is the Virtual Research Environment (VRE) of the project, hosted on the D4Science infrastructure. `https://dante.d4science.org/`

5.2.4
IMAGO Annotation Tool

N. Pratelli, V. Bartalesi, E. Lenzi [84]

To populate the ontology developed within the Index Medii Aevi Geographiae Operum (IMAGO) – Italian National Research Project (2020-23), we developed a semi-automatic Web tool, called IMAGO Annotation Tool, to allow scholars to insert knowledge about Medieval and Reinassance works through a user-friendly interface. The tool was created to reduce the time to insert knowledge and to avoid the insertion of mistakes thanks to the use of predefined lists of works, authors, libraries, places, geographic coordinates, and literary genres. Each field of the interface maps a class of the IMAGO ontology. The frontend interface is built using HTML5, CSS3, JavaScript, and the Bootstrap library, using a Python backend, e.i., a Django framework and a PostgreSQL DB. Once the data about a work is inserted through the tool interface, this is encoded as an OWL knowledge base and stored in a triple store. The data is first exported to a JSON object. Indeed our software uses a JSON schema to represent the data, structured according to the IMAGO ontology classes. The JSON object is processed by Java software, which transforms it into an OWL graph encoded in RDF/XML and Turtle formats. This software carries out its task by relying on the Apache Jena library. The graph is finally stored in a Fuseki triple store, and it can be queried through a SPARQL endpoint. `https://imagoarchive.it/tool/annotation/`

5.2.5
Interactive Classification System (ICS)

A. Esuli [43]

The Interactive Classification System (ICS), is a web-based application that supports the activity of manual text classification, i.e., labeling documents according to their content. The system is designed to give total freedom of action to its users: they can at any time modify any classification schema and any label assignment, possibly reusing any relevant information from previous activities. The application uses machine learning to actively support its users with classification suggestions The machine learning component of the system is an unobtrusive observer of the users' activities, never interrupting them, constantly adapting and updating its models in response to their actions, and always available to perform automatic classifications. `https://github.com/aesuli/ics`

5.2.6
Learning to count biological structures with raters' uncertainty

L. Ciampi, F. Carrara, V. Totaro, R. Mazziotti, L. Lupori, C. Santiago, G. Amato, T. Pizzorusso, C. Gennaro. [25]

Code for replicating the experiments in [25]. The provided code trains and tests the proposed two-stage counting pipeline that takes advantage from multi-rating data. `https://github.com/ciampluca/counting_perineuronal_nets`

5.2.7
MOBDrone: a Drone Video Dataset for Man OverBoard Rescue

D. Cafarelli, L. Ciampi, L. Vadicamo, C. Gennaro, A. Berton, M. Paterni, C. Benvenuti, M. Passera, F. Falchi. [14]

Code for replicating the experiments in [14]. The provided code tests some state-of-the-art object detectors over the new MOBDrone benchmark we introduced in the paper. `https://github.com/ciampluca/MOBDrone_eval`

5.2.8
Combining efficientnet and vision transformers for video deepfake detection

D.A. Coccomini, N. Messina, C. Gennaro, F. Falchi International Conference on Image Analysis and Processing (ICIAP) [28]

Code for replicating the experiments in [28] is available at: `https://github.com/davide-coccomini/Combining-EfficientNet-and-Vision-Transformers-for-Video-Deepfake-Detection`

5.2.9
Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection

D.A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, G. Amato 1st International Workshop on Multimedia AI against Disinformation [30]

Code for replicating the experiments in [30] is available at: `https://github.com/davide-coccomini/Cross-Forgery-Video-Deepfake-Detection`

5.2.10
MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection

D.A. Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, Claudio Gennaro [31]

Code for replicating the experiments in [31] is available at: `https://github.com/davide-coccomini/MINTIME-Multi-Identity-size-iNvariant-TIMEsformer-for-Video-Deepfake-Detection`

5.2.11
A Spatio-Temporal Attentive Network for Video-Based Crowd Counting

M. Avvenuti, M. Bongiovanni, L. Ciampi, F. Falchi, C. Gennaro, N. Messina. [3]

Code for replicating the experiments in [3]. The provided code trains and tests the proposed spatiotemporal attentive neural network to estimate the number of pedestrians from surveillance videos. `https://github.com/mesnico/attentive_video_crowd_counting`

5.2.12
The SSH Data Citation Service, a tool to explore and collect citation metadata

C. Concordia [32]

The source code of the SSH Data Citation Service `https://gitea-s2i2s.isti.cnr.it/concordia/sshoc-citationservice`

5.2.13
The SSH Open Marketplace Data Library

C. Concordia [61]

A Python library to download and process the SSH Open Marketplace dataset, and a set of notebooks providing examples and use cases to use this library. The library has been designed to be used by the SSH Open Marketplace Editorial Team and provides a set of functions that can be used in Python Notebooks or programs. `https://github.com/SSHOC/marketplace-curation`

5.2.14
Story Map Building and Visualising Tool (SMBVT)

E. Lenzi, V. Bartalesi, N. Pratelli, G. Coro, P. Pagano [62]

In the context of the MOVING (MOuntain Valorisation through INterconnectedness and Green growth) project, we released an open-source software - the MOVING Story Map Building and Visualization Tool (SMBVT) - that allows users to create and visualise story maps within a collaborative environment and using a user-friendly Web interface. The tool uses Semantic Web technologies and the Narrative Ontology to represent the stories of the MOVING mountain Value Chains. The MOVING community access SMBVT through The MOVING story map Virtual Research Environment and creates the events of the story. For each event, the user can add: a title, a textual description, start and end dates, the geographic coordinates, a media object (i.e. a video or image), notes, and digital objects. The tool takes Wikidata as reference KB and assigns Wikidata Internationalized Resource Identifiers (IRIs) to the story components (i.e. the entities that take part in an event). All the knowledge collected by SMBVT is stored in a JSON Postgres DB. When a story is completed, the tool automatically creates the corresponding visualisation using StoryMapJS library and makes available a corresponding URL that can be freely shared. Finally, SMBVT saves the collected knowledge as a Web Ontology Language (OWL) graph and publishes it as a Linked Open Data. `https://github.com/EmanueleLenzi92/SMBVT`

5.2.15
MedieValla: an authorship verification tool written in Python for medieval Latin

S. Corbara, A. Moreo, F. Sebastiani, M. Tavoni [38].

Code to reproduce the experiments we have conducted on two datasets for the computational authorship analysis of medieval Latin texts. See `https://doi.org/10.5281/zenodo.3903295` *for further details on the dataset.*
The code is freely available at `https://zenodo.org/record/4783678#.ZAheBtLMKiA`*.*

5.3 Shared Tasks

5.3.1
A Concise Overview of LeQua@ CLEF 2022: Learning to Quantify.
A Detailed Overview of LeQua@ CLEF 2022: Learning to Quantify.

A. Esuli, A. Moreo, F. Sebastiani, G. Sperduti. CLEF 2022, Working Notes of the 13th Conference and Labs of the Evaluation Forum. "Concise" [45], "Detailed" [46].

LeQua 2022 is a new lab for the evaluation of methods for "learning to quantify" in textual datasets, ie, for training predictors of the relative frequencies of the classes of interest $\mathscr{Y} = \{y_1, \ldots, y_n\}$ *in sets of unlabelled textual documents. While these predictions could be easily achieved by first classifying all documents via a text classifier and then counting the numbers of documents assigned to the classes, a growing body of literature has shown this approach to be suboptimal, and has proposed better methods. The goal of this lab is to provide a setting for the comparative evaluation of methods for learning to quantify, both in the binary setting and in the single-label multiclass setting; this is the first time that an evaluation exercise solely dedicated to quantification is organized. For both the binary setting and the single-label multiclass setting, data were provided to participants both in ready-made vector form and in raw document form. In this overview article we describe the structure of the lab, we report the results obtained by the participants in the four proposed tasks and subtasks, and we comment on the lessons that can be learned from these results.*

5.3.2
LeQua@CLEF2022: Learning to Quantify

A. Esuli, A. Moreo, F. Sebastiani. ECIR 2022. [44]

LeQua 2022 is a new lab for the evaluation of methods for "learning to quantify" in textual datasets, i.e., for training predictors *of the relative frequencies of the classes of interest in sets of unlabelled textual documents. While these predictions could be easily achieved by first classifying all documents via a text classifier and then counting the numbers of documents assigned to the classes, a growing body of literature has shown this approach to be suboptimal, and has proposed better methods. The goal of this lab is to provide a setting for the comparative evaluation of methods for learning to quantify, both in the binary setting and in the single-label multiclass setting. For each such setting we provide data either in ready-made vector form or in raw document form.*

6. Services

6.1 Services in conferences

In this section, we report the conference in which we were involved in the organization.

6.1.1 SISAP 2022

15th International Conference on Similarity Search and Applications, SISAP 2022, Bologna, Italy, October 5-7, 2022.

- Fabrizio Falchi, Program Committee Co-Chair

6.1.2 SEBD 2022

30th Symposium on Advanced Database System - Tirrenia (Pisa), Italy - 19-22 June 2022.

- Giuseppe Amato, General Co-Chair
- Clauydio Gennaro, Program Committee Co-Chair
- Fabrizio Falchi, Local Chair
- Lucia Vadicamo, Doctoral Consortium Co-Chair
- Fabio Carrara and Andrea Esuli, Publicity Chairs
- Valentina Bartalesi, Publication chair
- Catherine Bosio, Logistics Chair
- Alessandro Nardi, Treasury Chair
- Paolo Bolettieri, Web Chair

6.1.3 CLEF 2022

13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022

- Fabrizio Sebastiani, Program Committee Co-Chair

7. Awards

7.1 ERCIM Cor Baayen Young Researcher Award

The ERCIM Evaluation Committee for the Cor Baayen Young Researcher Award unanimously selected Fabio Carrara from ISTI-CNR as the winner for 2022.

Motivations: *Fabio's research activity stands out for its quality and interdisciplinarity. Worthy of note is his work on adversarial attack detection, proposing solutions based on the analysis of the features extracted by the various layers of deep neural networks. He also researched the application of appropriately simplified deep neural networks on resource-constrained devices, such as smart cameras. His research results are not only theoretical but also have significant application and technology transfer implications, as for example, the miniaturised models for parking occupancy detection (*`http://cnrpark.it/`*).*

Figure 15. ERCIM Cor Baayen Youn Researcher Award.

7.2 Video Browser Showdown

The VISIONE content-based video retrieval system, ranked first in the KIS Visual Category of Video Browser Showdown, The Video Retrieval Competition, with the approach described in [1].

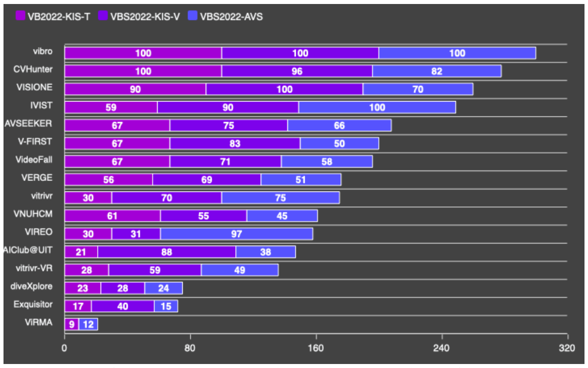

Figure 16. Video Browser Showdown, The Video Retrieval Competition, results .

7.3 ISTI Young Research Awards

The esearchers of the AIMH Lab that won the ISTI Young Research Awards "Matteo Delle Piane" in 2022 are:

- Lucia Vadicamo, in the Advanced category
- Nicola Messina, in the Beginner category

The esearchers of the AIMH Lab that won an ISTI Grant for Young Mobility in 2022 are:

- Luca Ciampi, first call
- Nicola Messina, second call

References

[1] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. Visione at video browser showdown 2022. In *MultiMedia Modeling*, pages 543–548, Cham, 2022. Springer International Publishing.

[2] Giuseppe Amato, Fabio Carrara, Luca Ciampi, Marco Di Benedetto, Claudio Gennaro, Fabrizio Falchi, Nicola Messina, and Claudio Vairo. Ai and computer vision for smart cities. In *Proceedings of the 8th Italian Conference on ICT for Smart Cities And Communities*, 2022.

[3] Marco Avvenuti, Marco Bongiovanni, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Nicola Messina. A spatio- temporal attentive network for video-based crowd counting. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6, 2022.

[4] Federica Baldi. Action detection in rugby video sequences in both professional and amateur settings. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[5] Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*. Springer, 2022.

[6] Valentina Bartalesi, Nicolò Pratelli, and Emanuele Lenzi. A knowledge base of medieval and renaissance geographic latin works. In *IRCDL 2022 - 18th Italian Research Conference on Digital Libraries, Padua, Italy, 24-25/02/2022*. CEUR-WS.org, Aachen, DEU, 2022.

[7] Valentina Bartalesi, Nicolò Pratelli, and Emanuele Lenzi. Linking different scientific digital libraries in digital humanities: the imago case study. *International Journal on Digital Libraries*, 23(4):303–317, 2022.

[8] Valentina Bartalesi, Nicolò Pratelli, Daniele Metilli, and Carlo Meghini. HDN annotation tool, 2022.

[9] Barbara Berti, Andrea Esuli, and Fabrizio Sebastiani. Unravelling interlanguage facts via explainable machine learning. *CoRR*, abs/2208.01468, 2022.

[10] Lorenzo Biondi. Ai-assisted 3d human head keypoint matching for image pair alignment. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[11] Marco Bongiovanni. Design and implementation of a system for automatic pedestrian counting in video streams based on artificial intelligence techniques. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[12] Mirko Bunse, Alejandro Moreo, Fabrizio Sebastiani, and Martin Senz. Ordinal quantification through regularization. In *Proceedings of the 33rd European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML / PKDD 2022)*, Grenoble, FR, 2022. Forthcoming.

[13] Vairo C., Coccomini D. A., Falchi F., Gennaro C., Massoli F. V., Messina N., and Amato G. Aimh lab for cybersecurity. In *Ital-IA 2022 - Workshop su AI per Cybersecurity, 10/02/2022*, 2022.

[14] Donato Cafarelli, Luca Ciampi, Lucia Vadicamo, Claudio Gennaro, Andrea Berton, Marco Paterni, Chiara Benvenuti, Mirko Passera, and Fabrizio Falchi. Mobdrone: A drone video dataset for man overboard rescue. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 633–644, Cham, 2022. Springer International Publishing.

[15] Donato Cafarelli, Luca Ciampi, Lucia Vadicamo, Claudio Gennaro, Andrea Berton, Marco Paterni, Chiara Benvenuti, Mirko Passera, and Fabrizio Falchi. MOBDrone: a large-scale drone-view dataset for man overboard detection, February 2022.

[16] Roberto Caldelli, Fabio Carrara, and Fabrizio Falchi. Tuning neural ode networks to increase adversarial robustness in image forensics. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1496–1500, 2022.

[17] Fabio Carrara, Roberto Caldelli, Fabrizio Falchi, and Giuseppe Amato. Improving the adversarial robustness of neural ode image classifiers by tuning the tolerance parameter. *Information*, 13(12), 2022.

[18] Fabio Carrara, Fabrizio Falchi, Maria Girardi, Nicola Messina, Cristina Padovani, and Daniele Pellegrini. Deep learning for structural health monitoring: An application to heritage structures, 2022.

[19] Fabio Carrara, Lorenzo Pasco, Claudio Gennaro, and Fabrizio Falchi. Learning to detect fallen people in virtual worlds. In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, CBMI '22, page 126–130, New York, NY, USA, 2022. Association for Computing Machinery.

[20] Fabio Carrara, Lucia Vadicamo, Claudio Gennaro, and Giuseppe Amato. Approximate nearest neighbor search on standard search engines. In *International Conference on Similarity Search and Applications*, pages 214–221. Springer, 2022.

[21] Luca Ciampi. *Deep Learning Techniques for Visual Counting*. PhD thesis, Dottorato in Ingegneria dell'informazione, University of Pisa, Italy, 2022.

[22] Luca Ciampi. *Relational Learning in Computer Vision*. PhD thesis, Dottorato in Ingegneria dell'informazione, University of Pisa, Italy, 2022.

[23] Luca Ciampi, Fabio Carrara, Giuseppe Amato, and Claudio Gennaro. Counting or localizing? evaluating cell counting and detection in microscopy images. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2022.

[24] Luca Ciampi, Fabio Carrara, Valentino Totaro, Raffaele Mazziotti, Leonardo Lupori, Carlos Santiago, Giuseppe Amato, Tommaso Pizzorusso, and Claudio Gennaro. A Multi-Rater Benchmark for Perineuronal Nets Detection and Counting in Fluorescence Microscopy Images, October 2021.

[25] Luca Ciampi, Fabio Carrara, Valentino Totaro, Raffaele Mazziotti, Leonardo Lupori, Carlos Santiago, Giuseppe Amato, Tommaso Pizzorusso, and Claudio Gennaro. Learning to count biological structures with raters' uncertainty. *Medical Image Analysis*, 80:102500, aug 2022.

[26] Luca Ciampi, Paweł Foszner, Nicola Messina, Michał Staniszewski, Claudio Gennaro, Fabrizio Falchi, Gianluca Serao, Michał Cogiel, Dominik Golba, Agnieszka Szczesna, and Giuseppe Amato. Bus violence: An open benchmark for video violence detection on public transport. *Sensors*, 22(21), 2022.

[27] Luca Ciampi, Claudio Gennaro, Fabio Carrara, Fabrizio Falchi, Claudio Vairo, and Giuseppe Amato. Multi-camera vehicle counting using edge-ai. *Expert Systems with Applications*, 207:117929, 2022.

[28] Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection, 2022.

[29] Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Cross-forgery analysis of vision transformers and cnns for deepfake image detection. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, MAD '22, page 52–58, New York, NY, USA, 2022. Association for Computing Machinery.

[30] Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Cross-forgery analysis of vision transformers and cnns for deepfake image detection. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, MAD '22, page 52–58, New York, NY, USA, 2022. Association for Computing Machinery.

[31] Davide Alessandro Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. Mintime: Multi-identity size-invariant video deepfake detection, 2022.

[32] Cesare Concordia, Edward J. Gray, and Nicolas Larrousse. The ssh data citation service, a tool to explore and collect citation metadata. In *TPDL 2022 - 26th International Conference on Theory and Practice of Digital Libraries*, pages 351–356. Springer, 2022.

[33] Richard Connor, Alan Dearle, and Lucia Vadicamo. Investigating binary partition power in metric query. In *Proceedings of the 30th Italian Symposium on Advanced Database Systems, SEBD*, pages 415–426, 2022.

[34] Silvia Corbara, Berta Chulvi, Paolo Rosso, and Alejandro Moreo. Rhythmic and psycholinguistic features for authorship tasks in the spanish parliament: Evaluation and analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 79–92. Springer, 2022.

[35] Silvia Corbara, Berta Chulvi Ferriols, Paolo Rosso, and Alejandro Moreo. Investigating topic-agnostic features for authorship tasks in spanish political speeches. In *Natural Language Processing and Information Systems - 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15-17, 2022, Proceedings*, volume 13286 of *Lecture Notes in Computer Science*, pages 394–402. Springer, 2022.

[36] Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. Syllabic quantity: Rhythmic clues for latin authorship. *Information Matters*, 2(6), 2022. https://bit.ly/3b0fB0y.

[37] Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. Syllabic quantity patterns as rhythmic features for Latin authorship attribution. *Journal of the Association for Information Science and Technology*, 74(1):128–141, 2023.

[38] Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. MedLatinEpi and MedLatinLit: Two datasets for the computational authorship analysis of medieval Latin texts. *ACM Journal of Computing and Cultural Heritage*, 15(3):57:1–57:15, 2022.

[39] Juan José del Coz, Pablo González, Alejandro Moreo, and Fabrizio Sebastiani, editors. *Proceedings of the 2nd International Workshop on Learning to Quantify (LQ 2021)*. Grenoble, FR, 2022.

[40] Marco Di Benedetto, Fabio Carrara, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. An embedded toolset for human activity monitoring in critical environments. *Expert Systems with Applications*, 199:117125, 2022.

[41] Marco Di Benedetto, Fabio Carrara, Benedetta Tafuri, Salvatore Nigro, Roberto De Blasi, Fabrizio Falchi, Claudio Gennaro, Giuseppe Gigli, Giancarlo Logroscino, and Giuseppe Amato. Deep networks for behavioral variant

frontotemporal dementia identification from multiple acquisition sources. *Computers in Biology and Medicine*, 148:105937, 2022.

[42] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[43] Andrea Esuli. ICS: Total freedom in manual text classification supported by unobtrusive machine learning. *IEEE Access*, 10:64741–64760, 2022.

[44] Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. Lequa@clef2022: Learning to quantify. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 374–381. Springer, 2022.

[45] Andrea Esuli, Alejandro Moreo, Fabrizio Sebastiani, and Gianluca Sperduti. A concise overview of LeQua 2022: Learning to quantify. In *Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022)*, pages 362–381, Bologna, IT, 2022.

[46] Andrea Esuli, Alejandro Moreo, Fabrizio Sebastiani, and Gianluca Sperduti. A detailed overview of LeQua 2022: Learning to quantify. In *Working Notes of the 13th Conference and Labs of the Evaluation Forum (CLEF 2022)*, Bologna, IT, 2022.

[47] Carrara F., Ciampi L., Di Benedetto M., Falchi F., Gennaro C., Massoli F. V., and Amato G. Aimh lab for the industry. In *Ital-IA 2022 - Workshop su AI per l'Industria, Online conference, 10/02/2022*, 2022.

[48] Edoardo Fazzari. Ai in the loop in animal robot interaction. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[49] Paweł Foszner, Michał Staniszewski, Agnieszka Szczęsna, Michał Cogiel, Dominik Golba, Luca Ciampi, Nicola Messina, Claudio Gennaro, Fabrizio Falchi, Giuseppe Amato, and Gianluca Serao. Bus Violence: a large-scale benchmark for video violence detection in public transport, September 2022.

[50] Paweł Foszner, Agnieszka Szczęsna, Luca Ciampi, Nicola Messina, Adam Cygan, Bartosz Bizoń, Michał Cogiel, Dominik Golba, Elżbieta Macioszek, and Michał Staniszewski. Crowdsim2: an open synthetic benchmark for object detectors. In *18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023)*, 2023. accepted.

[51] Paweł Foszner, Agnieszka Szczęsna, Luca Ciampi, Nicola Messina, Adam Cygan, Bartosz Bizoń, Michał Cogiel, Dominik Golba, Elżbieta Macioszek, and Michał Staniszewski. Development of a realistic crowd simulation environment for fine-grained validation of people tracking methods. In *18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023)*, 2023. accepted.

[52] Margherita Gambini, Tiziano Fagni, Fabrizio Falchi, and Maurizio Tesconi. On pushing deepfake tweet detection capabilities to the limits. In *14th ACM Web Science Conference 2022*, WebSci '22, page 154–163, New York, NY, USA, 2022. Association for Computing Machinery.

[53] Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh M. Q. Bui, Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Giuseppe Amato, Gianpaolo Perelli, Sara Concas, Carlo Cuccu, Giulia Orrù, Gian Luca Marcialis, and Sebastiano Battiato. The face deepfake detection challenge. *Journal of Imaging*, 8(10), 2022.

[54] Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh M. Q. Bui, Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Giuseppe Amato, Gianpaolo Perelli, Sara Concas, Carlo Cuccu, Giulia Orrù, Gian Luca Marcialis, and Sebastiano Battiato. The face deepfake detection challenge. *Journal of Imaging*, 8(10), 2022.

[55] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, et al. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown. *International Journal of Multimedia Information Retrieval*, 11(1):1–18, 2022.

[56] Ciampi L., Cafarelli D., Carrara F., Di Benedetto M., Falchi F., Gennaro C., Massoli F. V., Messina N., and Amato G. Aimh lab: Smart cameras for public administration. In *Ital-IA 2022 - Workshop su AI per la Pubblica Amministrazione, Online conference, 10/02/2022*, 2022.

[57] Gabriele Lagani, Davide Bacciu, Claudio Gallicchio, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Deep features for cbir with scarce data using hebbian learning. In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, CBMI '22, page 136–141, New York, NY, USA, 2022. Association for Computing Machinery.

[58] Gabriele Lagani, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Comparing the performance of hebbian against backpropagation learning using convolutional

neural networks. *Neural Comput. Appl.*, 34(8):6503–6519, 2022.

[59] Gabriele Lagani, Claudio Gennaro, Hannes Fassold, and Giuseppe Amato. Fasthebb: Scaling hebbian training of deep neural networks to imagenet level. In Tomáš Skopal, Fabrizio Falchi, Jakub Lokoč, Maria Luisa Sapino, Ilaria Bartolini, and Marco Patella, editors, *Similarity Search and Applications*, pages 251–264, Cham, 2022. Springer International Publishing.

[60] Nicolas Larrousse and Cesare Gray, Edward J.and Concordia. Diary of our initiatory journey on the continent of data citation in ssh. In *Digital Humanities 2022 responding to Asian diversity, DH22, online conference*, 2022.

[61] Barbot Laure, Gray Edward, Fischer Frank, Ďurčo Matej, König Alexander, Puren Marie, Buddenbohm Stefan, Concordia Cesare, and Illmayer Klaus. The ssh open marketplace: a multi-voiced story. In *DARIAH Annual Event 2022*, 2022.

[62] Emanuele Lenzi, Valentina Bartalesi, Nicolò Pratelli, Gianpaolo Coro, and Pasquale Pagano. Story map building and visualising tool (SMBVT), 2022.

[63] Jakub Lokoč, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Ladislav Peška, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, Stefanos Vrochidis, et al. A task category space for user-centric comparative multimedia search evaluations. In *International conference on multimedia modeling*, pages 193–204. Springer, 2022.

[64] Di Benedetto M., Carrara F., Ciampi L., Falchi F., Gennaro C., and Amato G. Aimh lab for healthcare and wellbeing. In *Ital-IA 2022 - Workshop AI per la Medicina e la Salute, Online conference, 10/02/2022*, 2022.

[65] Achilles Machumilane, Alberto Gott, Pietro Cassará, Claudio Gennaro, and Giuseppe Amato. Actor-critic scheduling for path-aware air-to-ground multipath multimedia delivery. In *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pages 1–5. IEEE, 2022.

[66] Fabio Valerio Massoli, Lucia Vadicamo, Giuseppe Amato, and Fabrizio Falchi. A leap among quantum computing and quantum neural networks: A survey. *ACM Comput. Surv.*, 55(5), dec 2022.

[67] Francesco Merola., Fabrizio Falchi., Claudio Gennaro., and Marco Di Benedetto. Reinforced damage minimization in critical events for self-driving vehicles. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 258–266. INSTICC, SciTePress, 2022.

[68] Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Recurrent vision transformer for solving visual reasoning problems. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 50–61, Cham, 2022. Springer International Publishing.

[69] Nicola Messina, Davide Alessandro Coccomini, Andrea Esuli, and Fabrizio Falchi. Transformer-based multimodal proposal and re-rank for wikipedia image-caption matching, 2022.

[70] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. Aladin: Distilling fine-grained alignment scores for efficient image-text matching and retrieval. In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, CBMI '22, page 64–70, New York, NY, USA, 2022. Association for Computing Machinery.

[71] Filippo Minutella. Design and implementation of a deep learning system for knowledge graph analysis. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[72] Filippo Minutella, Fabrizio Falchi, Paolo Manghi, Michele De Bonis, and Nicola Messina. Towards unsupervised machine learning approaches for knowledge graphs. In Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, and Gianmaria Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[73] Alessio Molinari, Andrea Esuli, and Fabrizio Sebastiani. Active learning and the saerens-latinne-decaestecker algorithm: An evaluation. In Lynda Tamine, Enrique Amigó, and Josiane Mothe, editors, *Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022), Samatan, Gers, France, July 4-7, 2022*, volume 3178 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[74] Alejandro Moreo, Manuel Francisco, and Fabrizio Sebastiani. Multi-label quantification. *CoRR*, abs/2211.08063, 2022.

[75] Alejandro Moreo, Andrea Pedrotti, and Fabrizio Sebastiani. Generalized Funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification. *ACM Transactions on Information Systems*, 41(2):1–37, 2023 (online 2022).

[76] Alejandro Moreo and Fabrizio Sebastiani. Tweet sentiment quantification: An experimental re-evaluation. *PLOS ONE*, 17(9):1–23, September 2022.

[77] Aloia N., Amato G., Bartalesi V., Benedetti F., Bolettieri P., Cafarelli D., Carrara F., Casarosa V., Coccomini D., Ciampi L., Concordia C., Corbara S., Di Benedetto M., Esuli A., Falchi F., Gennaro C., Lagani G., Massoli F. V., Meghini C., Messina N., Metilli D., Molinari

A., Moreo A., Nardi A., Pedrotti A., Pratelli N., Rabitti F., Savino P., Sebastiani F., Sperduti G., Thanos C., Trupiano L., Vadicamo L., and Vairo C. Aimh research activities 2021. Technical report, ISTI Annual Report, ISTI-2021-AR/003, pp.1–34, 2021, 2021.

[78] Aloia N., Amato G., Bartalesi V., Benedetti F., Bolettieri P., Carrara F., Casarosa V., Ciampi L., Concordia C., Corbara S., Esuli A., Falchi F., Gennaro C., Lagani G., Massoli F. V., Meghini C., Messina N., Metilli D., Molinari A., Moreo A., Nardi A., Pedrotti A., Pratelli N., Rabitti F., Savino P., Sebastiani F., Thanos C., Trupiano L., Vadicamo L., and Vairo C. Aimh research activities 2020. Technical report, ISTI Annual Report, ISTI-2020-AR/001, 2020, 2020.

[79] Messina N., Carrara F., Coccomini D., Falchi F., Gennaro C., and Amato G. Aimh lab for trustworthy ai. In *Ital-IA 2020 - Workshop su AI Responsabile ed Affidabile, Online conference, 10/02/2022*, 2022.

[80] Nikolaos Partarakis, Paraskevi Doulgeraki, Effie Karuzaki, Ilia Adami, Stavroula Ntoa, Daniele Metilli, Valentina Bartalesi, Carlo Meghini, Yannis Marketakis, Danai Kaplanidi, Maria Theodoridou, and Xenophon Zabulis. Representation of socio-historical context to support the authoring and presentation of multimodal narratives: The mingei online platform. *J. Comput. Cult. Herit.*, 15(1), 2022.

[81] Lorenzo Pasco. Design and development of artificial intelligence-based techniques for automatic detection of fallen people. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[82] Stefano Poleggi. Design and implementation of facial expression recognition system. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[83] Nicolò Pratelli. A geographical extension for nont ontology. In *Proceedings of the Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries (TPDL-WS-DC 2022), Padua, Italy, September 20, 2022*. CEUR-WS.org, Aachen, DEU, 2022.

[84] Nicolò Pratelli, Valentina Bartalesi, and Emanuele Lenzi. IMAGO annotation tool, 2022.

[85] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[86] Gianluca Serao. Development of deep learning models for fight detection in videos. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[87] Alessio Serra. Cross-modal learning for sentiment analysis of social media images. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[88] Tomás Skopal, Fabrizio Falchi, Jakub Lokoc, Maria Luisa Sapino, Ilaria Bartolini, and Marco Patella, editors. *Similarity Search and Applications - 15th International Conference, SISAP 2022, Bologna, Italy, October 5-7, 2022, Proceedings*, volume 13590 of *Lecture Notes in Computer Science*. Springer, 2022.

[89] Agnieszka Szczęsna, Paweł Foszner, Adam Cygan, Bartosz Bizoń, Michał Cogiel, Dominik Golba, Luca Ciampi, Nicola Messina, Elżbieta Macioszek, and Michał Staniszewski. Crowd simulation (CrowdSim2) for tracking and object detection, February 2023.

[90] Costantino Thanos, Carlo Meghini, Valentina Bartalesi, and Gianpaolo Coro. An exploratory approach to archaeological knowledge production. *International Journal on Digital Libraries*, 23(3):231–239, 2022.

[91] Lucia Vadicamo, Alan Dearle, and Richard Connor. On the expected exclusion power of binary partitions for metric search. In *International Conference on Similarity Search and Applications*, pages 104–117. Springer, 2022.

[92] Gaetano Emanuele Valenti. Design, implementation, and test of tools for multi-camera vehicle tracking based on artificial intelligence. Master's thesis, M.Sc. in Artificial Intelligence and Data Engineering, University of Pisa, 2022.

[93] Aurelia Viglione, Giulia Sagona, Fabio Carrara, Giuseppe Amato, Valentino Totaro, Leonardo Lupori, Elena Putignano, Tommaso Pizzorusso, and Raffaele Mazziotti. Behavioral impulsivity is associated with pupillary alterations and hyperactivity in cdkl5 mutant mice. *Human Molecular Genetics*, 31(23):4107–4120, 2022.

[94] Xenophon Zabulis, Nikolaos Partarakis, Carlo Meghini, Arnaud Dubois, Sotiris Manitsaris, Hansgeorg Hauser, Nadia Magnenat Thalmann, Chris Ringas, Lucia Panesse, Nedjma Cadi, Evangelia Baka, Cynthia Beisswenger, Dimitrios Makrygiannis, Alina Glushkova, Brenda Elizabeth Olivas Padilla, Danae Kaplanidi, Eleana Tasiopoulou, Catherine Cuenca, Anne-Laure Carre, Vito Nitti, Ilia Adami, Emmanouil Zidianakis, Paraskevi Doulgeraki, Effie Karouzaki, Valentina Bartalesi, and Daniele Metilli. A representation protocol for traditional crafts. *Heritage*, 5(2):716–741, 2022.