AIMIR Research Activities 2019

Giuseppe Amato[1]*, Paolo Bolettieri[1], Fabio Carrara[1], Luca Ciampi[1], Marco Di Benedetto[1], Franca Debole[1], Fabrizio Falchi[1], Claudio Gennaro[1], Gabriele Lagani[1], Fabio Valerio Massoli[1], Nicola Messina[1], Fausto Rabitti[1], Pasquale Savino[1], Lucia Vadicamo[1], Claudio Vairo[1]

Abstract
The Artificial Intelligence for Multimedia Information Retrieval (AIMIR) research group is part of the NeMIS laboratory of the Information Science and Technologies Institute "A. Faedo" (ISTI) of the Italian National Research Council (CNR). The AIMIR group has a long experience in topics related to: Artificial Intelligence, Multimedia Information Retrieval, Computer Vision and Similarity search on a large scale. We aim at investigating the use of Artificial Intelligence and Deep Learning, for Multimedia Information Retrieval, addressing both effectiveness and efficiency. Multimedia information retrieval techniques should be able to provide users with pertinent results, fast, on huge amount of multimedia data. Application areas of our research results range from cultural heritage to smart tourism, from security to smart cities, from mobile visual search to augmented reality. This report summarize the 2019 activities of the research group.

Keywords
Multimedia Information Retrieval – Artificial Intelligence — Computer Vision — Similarity Search

[1] *AIMIR Group, NeMIS Lab, ISTI-CNR, via G. Moruzzi, 1 - 56127 Pisa PI, Italy, EU*
***Corresponding author**: giuseppe.amato@isti.cnr.it

Contents

AIMIR
Artificial Intelligence for
Multimedia Information Retrieval

Introduction

The Artificial Intelligence for Multimedia Information Retrieval (AIMIR) research group of the NeMIS-CNR lab has a long experience in topics related to Artificial Intelligence, Multimedia Information Retrieval, Computer Vision and Similarity search on a large scale. We investigate the use of Artificial Intelligence and Deep Learning for Multimedia Information Retrieval, addressing both issues of effectiveness and efficiency. Multimedia information retrieval techniques should be able to provide users with relevant results, fast, on a huge amount of multimedia data. Application areas of our research results range from cultural heritage to smart tourism, from security to smart cities, from mobile visual search to augmented reality.

In this paper, we report the activities of the AIMIR research group on 2019. The rest of the report is organized as follows. In Section 1, we summarizes the research conducted on our main research fields. In Section 2, we describes the projects in which we were involved during the year. We report the complete list of papers we published in 2019, together with their abstract, in Section 3. The list of Thesis on which we were involved can be found in Section 4. In Section 6, we highlights the datasets we created and made publicly available during 2019. The awards we received along the year are listed in Section 7.

1. Research Topics

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. In the paper reported in Section 4.2 and 3.2.5, we employed a different learning approach, belonging to the family of *Hebbian* learning algorithms, which are biologically plausible.

The algorithm was applied to train deep convolutional neural networks on image the classification task. We also investigated the possibility of training networks in a hybrid fashion, i.e. training some layers using backpropagation and other layers using the Hebbian algorithm. We found that the Hebbian algorithm is useful to train the lower layers (responsible for low level representation learning) and the higher classification layers, although performance losses occur when intermediate layers are converted from backprop to Hebbian-based. On the other hand, Hebbian training can be completed in fewer epochs (e.g. 1-2 vs 10-20 epochs required by gradient descent), which makes the algorithm useful for applications in the context of transfer learning. For example, given a network pre-trained on some task, it is possible to fine-tune the higher layers on a new task with low computational cost.

In future works, we aim at exploring further alternatives of biologically plausible learning algorithms, in order to be able to train all the network layers in a biologically plausible manner.

1.1.2 Abstract Reasoning on Neural Networks

Despite their success, there are still many open problems with current deep architectures: it is known that they cannot generalize well to unseen objects and they lack human-like reasoning capabilities. Humans, as well as animals, can abstract visual perception to recognize some shape patterns never seen before. Unlike humans, convolutional networks are able to learn very precise representations, usually very difficult to generalize to abstract concepts.

We concentrated on a very specific task designed to test abstract reasoning capabilities of neural networks. This task is called *same-different* and consists in predicting if two shapes inside the same image are the same or not. It is a challenging task for convolutional architectures since it is required not to learn specific shape patterns to solve the problem.

A clear understanding of the *same-different* concept, and the ability to actuate complex relational reasoning are considered key features in many scenarios. For example, these concepts may have a great impact on classification problems, on the research of particular patterns in cultural heritage, and the detection of patterns defining aesthetic beauty in images and even music. Indeed, the world is often perceived by humans as a set of recurrent structures composited together, such as the human eyes in a face or the repeating chorus in a song.

Guided by these motivations, we tested a large number of state-of-the-art architectures for object classification (Section 3.2.2), and we analyzed their capability to deal with this challenging task. We employed a carefully designed dataset (SVRT) containing simple 2D images satisfying the *same-different* requirements, together with negative examples. We discovered that novel deep architectures are able to correctly solve this task and they can generalize quite well, while shallow architectures cannot move from the chance accuracy.

In future works, we manage to isolate the important architectural factors that enable architectures to correctly learn this problem. Furthermore, we are planning to tackle the same kind of problems in real-world pictures, and to work with sequential data such as audio files.

1.1.3 Adversarial Machine Learning

Adversarial machine learning is about attempting to fool models through malicious input. The topic has become very popular with the recent advances on Deep Learning. We started working on this topic with a focus on detection of adversarial examples and images in particular in 2017 [1]. In 2018, we started studying the deep features distance space for the same goal of detecting adversarial images and the results were encouraging [2]. We extended these previous works in a journal paper published on MTAP [3]. These activities were also part of the Carrara PhD Thesis (see Section 4.1) defended in 2019. An overview of the same technique was given on ERCIM News [4]. During this year, an extensive analysis of the layer activation in case of adversarial attacks were re-

ported in [5]. We also adapted our techniques to the case of recently proposed ODE-Nets in [6]. Finally, we consider the task of detecting adversarial faces, i.e., malicious faces given to machine learning systems in order to fool the recognition and the verification in particular [7].

1.1.4 Neural Ordinary Differential Equations

Presented in a NeurIPS 2018 best paper [8], Neural Ordinary Differential Equations comprise novel differentiable and learnable models (also referred to as ODE-Nets) whose outputs are defined as the solution of a system of parametric ordinary differential equations. Those models exhibit benefits such as a O(1)-memory consumption and a straight-forward modelling of continuous-time and inhomogeneous data, and when using adaptive ODE solvers, they acquire also other interesting properties, such as input-dependent adaptive computation and the tunability (via a tolerance parameter) of the accuracy-speed trade-off at inference time.

Our research on this topic concerns the evaluation of these novel models when applied to common vision tasks. We analyzed the continuous image representation learned by ODE-Nets and evaluate them in image classification and transfer learning tasks, observing an improved transferability over features extracted by standard residual networks [9]. (More details in Section 3.2.8).

Another important analysis we conducted concerns the vulnerability of ODE-defined networks to adversarial attacks: we conducted experiments on classical (residual), mixed, and proposed ODE-only architectures, and showed that when feature extraction is performed mostly by ODE blocks, the attack success rate of strong adversarial attacks, like Projected Gradient Descent, is lower and can be furthermore reduced by controlling the tolerance of the adaptive ODE solver [6]. (Mode details in Section 3.2.9).

Future work will investigate novel methods to regularize the training of ODE-Nets and explore applications to time-series multimedia data, such as videos or motion capture data.

1.1.5 Intuition

We investigated the relation between intuition, recognition and deep learning. Starting from the famous quote "*Intuition is nothing more and nothing less than recognition*", by Herbert Simon, we considered the ongoing research in deep learning, especially for computer vision, with respect to the research conducted by Daniel Kahneman and Gary Klein. Within the discipline of psychology and decision making, expert intuition has been discussed a lot in recent years, dividing researchers into believers and skeptics. However, after six years of discussion, a believer, Gary Klein, and a skeptic, Daniel Kahneman, wrote an important paper in 2009 whose subtitle was "A failure to disagree". Trying to answer the question When can we trust intuition? they agreed on a set of conditions for trustable intuitive expertise. Among these conditions, the most important ones are: an environment that is sufficiently regular to be predictable; an opportunity to learn these regularities through prolonged practice. In 3.3.1, we discussed the similarities between intuition and deep learning.

1.2 Computer Vision

1.2.1 Learning in Virtual Worlds

In the new spring of artificial intelligence, and in particular 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. For several applications, in the era of big data, the availability of real input examples, to train machine learning algorithms, is not considered an issue. However, for several other applications, there is not such an abundance of training data. 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. In fact, although a series of annotated datasets are available and successfully used to produce important academic results and commercially fruitful products, there is still a huge amount of scenarios where laborious human intervention is needed to produce high quality training sets. For example, such cases include, but are not limited to, safety equipment detection, weapon-wielding detection, and autonomous driven cars.

To overcome these limitations and to provide useful examples in a variety of scenarios, the research community has recently started to leverage on the use of programmable virtual scenarios to generate visual datasets and the needed associated annotations. For example, in an image-based machine learning technique, using a modern rendering engine (i.e., capable of producing photo-realistic imagery) has been proven a valid companion to automatically generate adequate datasets.

We successfully applied the *Virtual World approach* using the *Grand Tefth Auto V* engine for detecftion of personal protection equipment (CBMI Paper, Section 3.2.1), and M.D. Thesis, Section 4.2), and pedestrian (ICIAP Paper, Section 3.2.4).

1.2.2 Visual Counting

The counting problem is the estimation of the number of objects instances in still images or video frames. This task has recently become a hot research topic due to its interdisciplinary and widespread applicability and to its paramount importance for many real-world applications, for instance, counting bacterial cells from microscopic images, estimate the number of people present at an event, counting animals in ecological surveys with the intention of monitoring the population of a certain region, counting the number of trees in an aerial image of a forest, evaluate the number of vehicles in a highway or in a car park, monitoring crowds in surveillance systems, and others.

In humans, studies have demonstrated that, as a consequence of the subitizing ability [10], the brain switches between two techniques in order to count objects [11]. When the observed objects are less than five, the fast and accurate Parallel Individuation System(PIS) is employed, otherwise, the inaccurate and error-prone Approximate Number System(ANS)

is used. Thus, at least for crowded scenes, Computer Vision approaches offer a fast and useful alternative for counting objects.

In principle, the key idea behind objects counting using Computer Vision-based techniques is very simple: density times area. However, objects are not regular across the scene.They cluster in certain regions and are spread out in others. Another factor of complexity is represented by perspective distortions created by different camera viewpoints in various scenes, resulting in large variability of scales of objects. Others challenges points to be considered are inter-object and intra-object occlusions, high similarity of appearance between objects and background elements, different illuminations, and low image quality. In order to overcome these challenges, several machine learning-based approaches (especially supervised) have been suggested. However, these methods are limited to the scenarios where it is possible to collect and to manually annotate a representative set of training images. Systematic differences in the visual appearance or in the geometrical patterns between the training set and new images get in trouble these techniques and may result in systematic gross under- or over-counting.

We proposed some solutions able to count vehicles located in parking lots. In particular, we introduced a detection-based approach able to localize and count vehicles from images taken by a smart camera [12], [13], and another one from images captured by a drone [14].

1.2.3 Face Recognition

Face recognition is a key task in many application fields, such as security and surveillance. Several approaches have been proposed in the last few years to implement the face recognition task. Some approaches are based on local features of the images, such as Local Binary Pattern (LBP) [15] which combines local descriptors of the face in order to build a global representation that can be used to measure the distance with other LBP features. Some other approaches are based on detecting the facial landmarks from the detected face and on measuring the distance between some of these landmarks. Recently, Deep Learning approach and Convolutional Neural Networks (CNNs) have been proposed to address the face verification problem with very good results, such as [16].

We implemented several solutions based on the aforementioned techniques to address the face recognition problem in different application scenarios. For example, we studied the problem of intrusion detection in a monitored environment, and we designed a system to automatically detect unauthorized accesses to a restricted environment by exploiting face recognition on the images acquired by a Raspberry Pi camera placed in front of the entrance of the monitored room [17, 18, 19].

We also have compared the recognition accuracy in performing the face recognition task by using the distance of facial landmarks and some CNN-based approaches. Facial landmarks are very important in forensics because they can be used as objective proof in trials, however, the recognition accuracy of these approaches is much lower than the ones based on Deep Learning [20].

Recently, with the MoDro project (see Section 2.6), we started to investigate the possibility to move the face recognition computation inside embedded devices such as a Raspberry Pi so that they autonomously can perform this task without the need to transmit the video stream to a remote server for the analysis.

Finally, we consider the task of detecting adversarial faces, i.e., malicious faces given to machine learning systems in order to fool the recognition and the verification in particular [7].

1.2.4 Smart Parking

The advancements of Computer Vision and image understanding enables cheap vision-based solution for various application. In the context of smart city applications, we continued exploring systems for distributed vision-based parking lot occupancy detection and car counting.

Thanks to smart cameras and the miniaturization of complex vision models (i.e. CNNs), the detection of the occupancy status can be performed directly on-board of the camera, saving network bandwidth and implementing privacy by design. Moreover, with respect to traditional ground sensors, a single camera can monitor multiple spots and can be re-purposed for additional tasks (e.g. video-surveillance). Seminal work on parking lot detection with miniaturized CNNs in distributed environment has been published (see also Sec. 4.1).

In order to have a more flexible solution, we also proposed a deep learning-based approach that is able to count the cars in captured images without any extra information of the scenes, like the regions of interest (i.e. the position of the parking lots) or the perspective map. This is a key feature since in this way our solution is directly applicable in unconstrained contexts.

Some of the results on these topics have been presented in [21, 22, 13, 12].

Figure 1. Smart Camera parking monitoring

1.2.5 Image Quality Assessment

In the development of novel compression algorithms for images and videos, an essential step concerns the assessment of the quality of compressed images. This is a tedious task, as it usually requires expensive human questionnaires or the usage of computationally expensive algorithms for judging the quality and detect perceivable artifacts in the compressed images. We investigated the use of Deep Learning to predict a quality score for distorted imaged. We showed that deep models are able to approximate currently available algorithms (e.g. HDRVP, DRIM) at a lower computational cost [23]. More details are available in Section 3.1.5.

1.2.6 Understanding Motion Capture Data

Motion capture data represents movements of parts of objects and beings in the 3D space and in time. [Carrara, TODO]

1.3 Multimedia Information Retrieval

1.3.1 Large Scale

Large scale indexing and searching has always been a main focus of the research activities of the AIMIR group [24, 25, 26, 27, 28]. During 2019 we investigated some approaches to transform neural network features into text forms suitable for being indexed by a standard full-text retrieval engine such as Elasticsearch (see Section 3.1.1). This activity started last year with a paper published at SIGIR [29]. The basic idea of our approaches relies on a transformation of neural network features with the twofold aim of promoting the sparsity without the need of unsupervised pre-training. We validate our approach on a recent convolutional neural network feature, namely Regional Maximum Activations of Convolutions (R-MAC), which is a state-of-art descriptor for image retrieval. Its effectiveness has been proved through several instance-level retrieval benchmarks.

1.3.2 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 [30, 31], to compete at the Video Browser Showdown (VBS) [32], an international video search competition that evaluates the performance of interactive video retrievals systems. 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. The system supports four types of queries: query by keywords, query by object location, query by colors, and query by visual similarity. The user interface, shown in Figure 2, provides a text box to specify the keywords, and a canvas for sketching objects and colors to be found in the target video. The system leverages the image tagging system proposed in [33], which is able to label images with about

Figure 2. VISIONE User Interface

15K concepts. Furthermore, we use object detectors from YOLOv3 [34], and YOLO9000 [35] with about 9500 object tags, to detect objects in a video for the object location search. While the R-MAC [36] descriptors are adopted as image global descriptors for the similarity search functionality. In the VBS competition we faced three different content search tasks: *visual Known-Item Search (visual KIS)*, *textual Known-Item Search (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. From the experience at the competition, we ascertained a high efficiency regarding the indexing structure, made to support large scale multimedia access but a lack of effectiveness on keywords search. As a result of the system assessment made after the competition, we decide to invest more effort on the keywords-based search, trying to ameliorate the image annotation part. As forthcoming adjustment of the system, we plan to integrate some largely used datasets (i.e. Places365-CNN [37], SIN concepts [38]) to enrich the text retrieval part, and some automatic tools for scene understanding. Furthermore, we will improve the user interface to make it more usable and collaborative.

1.3.3 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. Methods for *exact similarity search* guarantee to find the true result set, but they scale poorly with the dimensionality of the data (*curse of dimensionality*) and the size of the data set. To overcome these issues, the research community has developed

a wide spectrum of techniques for *approximate similarity 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). One way to speed-up the metric searching is to transform the original space into a more tractable space where the transformed data objects can be efficiently indexed and searched.

In the past, we developed and proposed various techniques to support approximate similarity research in metric spaces, including approaches that rely on transforming the data objects into permutations (*Permutation-based indexing*) [39, 40, 41, 42], low-dimensional Euclidean vectors (*Supermetric search*) [43, 44], or compact binary codes (*Sketching technique*) [45]. Moreover, for a class of metric space that satisfy the so called "4-point property "[46] we derived a new pruning rule named *Hilbert Exclusion* [47], which can be used with any indexing mechanism based on hyperplane partitioning in order to determine subset of data that do not need to be exhaustively inspected.

During 2019, we further investigated the use of some geometrical properties (namely, the 4-point property and the n-point property [46]) to support metric search. In particular, we extensively explored the use of the 4-point property within a variety of different hyperplane partition indexing structures, and we proposed a search structure whose partition and exclusion conditions are tailored, at each node, to suit the individual reference points and data set [48] (see Section 3.1.3). Moreover, we further employed the the 4-point property to define a mechanism of local pivoting that associates each metric object with two reference objects that are best suited to filter that particular object if it is not relevant to a query, maximising the probability of excluding it from a search [49] (see Section 3.2.14). In previous work, we have developed a technique for mapping metric spaces into Euclidean vector spaces, called *nSimplex projection* [44], which can be applied on spaces meeting the the n-point property. Recently, by exploiting this technique we defined a novel approach to generate permutations associated with metric data objects [50]. Our *SPLX-Perms* (i.e., the permutations obtained using nSimplex projection) resulted, in most of the tested cases, more effective than traditional permutations, which makes them particularly suitable for permutation-based indexing techniques (see Section 3.2.13). Finally, we further exploited the nSimplex projection to embed string spaces into Euclidean vector spaces [51] (Section 3.2.16), and to transform metric objects into binary strings [52] (Section 3.2.15).

1.3.4 Relational Content-Based Image 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 tackled the problem introducing a novel task, called R-CBIR (Relational Content-Based Image Retrieval). Given a query image, the objective of is catching images that are similar to the input query not only in terms of detected entities but also with respect to the relationships (spatial and non-spatial) between them. We experimented with different variations of the recently introduced Relation Network architecture to extract relationship-aware visual features. In particular, we approached the problem transferring knowledge from the Relation Network module trained on the R-VQA task using the CLEVR dataset. Under this setup, in [53, 54] we introduced the Two-Stage Relation Network (2S-RN) and the Aggregated Visual Features Relation Network (AVF-RN) modules (Section 3.1.2). The first introduces late-fusion of question features into the visual pipeline in order to produce visual features not conditioned on the particular question. In the latter, we solved the problem of producing a compact and representative visual relationship-aware feature by aggregating all the possible couples of objects directly inside the network for training it end to end. Preliminary results on the CLEVR dataset show interesting relational abilities of these features on the R-CBIR task. Since this task requires compact features, R-CBIR encourages the development of solutions able to produce efficient yet powerful visual relationships-aware features, capable of efficiently describing even complex images.

2. Projects & Activities

2.1 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.

2.2 ADA

In the era of Big Data, manufacturing companies are overwhelmed by a lot of disorganized information: the large amount of digital content that is increasingly available in the manufacturing process makes the retrieval of accurate information a critical issue. In this context, and thanks also to the Industry 4.0 campaign, the Italian manufacturing industries have made a lot of effort to ameliorate their knowledge management system using the most recent technologies, like big data analysis and machine learning methods. In this context, therefore, the main target of the ADA project is to design and develop a platform based on big data analytics systems that allows for the acquisition, organization, and automatic retrieval of information from technical texts and images in the different phases of acquisition, design & development, testing, installation and maintenance of products. In paper [55], we illustrate the work carried out in the ADA project focusing on the image content retrieval part: the images contained in the corporate documents constitute a relevant source of information that could be relevant in the manufacturing work flow. On this context, we developed specific techniques for the extraction, classification, recognition, and tagging of images within technical documentation.

2.3 VIDEMO

Visual Deep Engines for Monitoring (VIDEMO) is a 2 year project funded by Regione Toscana, Istituto di Scienza e Tecnoologie dell'Informazione "A.Faedo" (ISTI) del CNR, Visual Engines srl. VIDEMO is about automatic analysis of images and video using deep learning methods for secure societies. The activities reported in Section 1.2.3 have been mainly conducted in the context of this project by Fabio Valerio Massoli. Fabrizio Falchi is the scientific coordinator of the project.

2.4 VISECH

Visual Engines for Cultural Heritage (VISECH) is a 2 year project funded by Regione Toscana, Istituto di Scienza e Tecnoologie dell'Informazione "A.Faedo" (ISTI) del CNR, Visual Engines srl. VISECH is about automatic analysis of cultural heritage multimedia. Giuseppe Amato is the scientific coordinator of the project.

2.5 MultiForesee

The main objective of this Action, entitled MULTI-modal Imaging of FOREnsic SciEnce Evidence (MULTI-FORESEE)-tools for Forensic Science[1], 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.

Some of the results obtained by this Action have been presented in [20].

2.6 MoDro

The aim of the MoDro project was to start experimentation of the 5G network in the territory of the city of Matera. The CNR, with the Institutes ISTI, IFC, IEIIT, ISASI, participated among the TIM's experimenters, with demos on jammer detection, electronic devices able to disturb frequencies in transmissions, video streaming footage in immersive reality of the city using the drone provided by CNR-ISASI of Lecce, and the face and objects recognition. The latter activity (developed by the AIMIR research group), in particular, is a scientific novelty, since the face and objects recognition is usually done with fixed cameras on the ground, and can prove to be a powerful means in the hands of the police for the safety of citizens (see Figure 3).

The drone flew at a height of 40m, never on people and in weather conditions with strong winds. The results obtained on facial recognition are therefore exceptional, given the conditions in which the data were found.

2.7 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 Nantiontal Virtual Lab on AI aims at proposing a strategic vision and big and long-term projects.

[1] https://multiforesee.com/

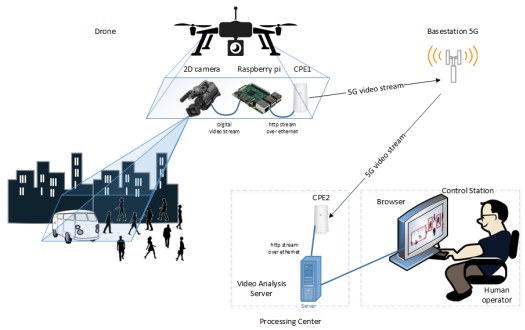


Figure 3. Face and Objects Recognition with 2D camera

3. Papers

3.1 Journals

3.1.1 Large-scale instance-level image retrieval

G. Amato, F. Carrara, F. Falchi, C. Gennaro, L. Vadicamo. In Press on Information Processing & Management special issue on Deep Learning for Information Retrieval.[56]. Abstract:

"*The great success of visual features learned from deep neural networks has led to a significant effort to develop efficient and scalable technologies for image retrieval. Nevertheless, its usage in large-scale Web applications of content-based retrieval is still challenged by their high dimensionality. To overcome this issue, some image retrieval systems employ the product quantization method to learn a large-scale visual dictionary from a training set of global neural network features. These approaches are implemented in main memory, preventing their usage in big-data applications. The contribution of the work is mainly devoted to investigating some approaches to transform neural network features into text forms suitable for being indexed by a standard full-text retrieval engine such as Elasticsearch. The basic idea of our approaches relies on a transformation of neural network features with the twofold aim of promoting the sparsity without the need of unsupervised pre-training. We validate our approach on a recent convolutional neural network feature, namely Regional Maximum Activations of Convolutions (R-MAC), which is a state-of-art descriptor for image retrieval. Its effectiveness has been proved through several instance-level retrieval benchmarks. An extensive experimental evaluation conducted on the standard benchmarks shows the effectiveness and efficiency of the proposed approach and how it compares to state-of-the-art main-memory indexes.*"

An interesting advantage of our approach is the possibility of implementation of search systems based on multimodal descriptions. In fact, the surrogate text approach allows in a straightforward way to deal with multimodal queries in the "text + image" form, such as look for images similar to a given one and containing the word "flower" in the metadata. Since features have also been transformed into text form, it is possible to create a single text query that contains both the metadata and the surrogate text corresponding to the visual query. This approach has been extensively used in [57, 41].

3.1.2 Learning visual features for relational CBIR

N. Messina, G. Amato, F. Carrara, F. Falchi, C. Gennaro In Press on International Journal of Multimedia Information Retrieval special issue on Deep Learning in Image and Video Retrieval[54]. Abstract:

"*Recent works in deep-learning research highlighted remarkable relational reasoning capabilities of some carefully designed architectures. In this work, we employ a relationship-aware deep learning model to extract compact visual features used relational image descriptors. In particular, we are interested in relational content-based image retrieval (R-CBIR), a task consisting in finding images containing similar inter-object relationships. Inspired by the relation networks (RN) employed in relational visual question answering (R-VQA), we present novel architectures to explicitly capture relational information from images in the form of network activations that can be subsequently extracted and used as visual features. We describe a two-stage relation network module (2S-RN), trained on the R-VQA task, able to collect non-aggregated visual features. Then, we propose the aggregated visual features relation network (AVF-RN) module that is able to produce better relationship-aware features by learning the aggregation directly inside the network. We employ an R-CBIR ground-truth built by exploiting scene-graphs similarities available in the CLEVR dataset in order to rank images in a relational fashion. Experiments show that features extracted from our 2S-RN model provide an improved retrieval performance with respect to standard non-relational methods. Moreover, we demonstrate that the features extracted from the novel AVF-RN can further improve the performance measured on the R-CBIR task, reaching the state-of-the-art on the proposed dataset.*"

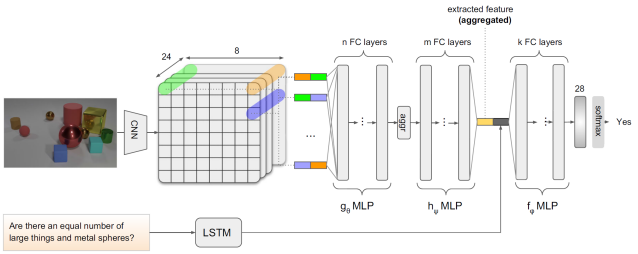


Figure 4. Architecture overview from [54].

3.1.3 Supermetric search

R. Connor, L. Vadicamo, F. A. Cardillo, F. Rabitti [48]. Abstract:

"*Metric search is concerned with the efficient evaluation of queries in metric spaces. In general, a large space of objects is arranged in such a way that, when a further object is presented as a query, those objects most similar to the query can be efficiently found. Most mechanisms rely upon the triangle inequality property of the metric governing the space. The triangle inequality property is equivalent to a finite embedding property, which states that any three points of the space can be isometrically embedded in two-dimensional Euclidean space. In this paper, we examine a class of semimetric space which is finitely four-embeddable in three-dimensional Euclidean space. In mathematics this property has been extensively studied and is generally known as the four-point property. All spaces with the four-point property are metric spaces, but they also have some stronger geometric guarantees. We coin the term supermetri space as, in terms of metric search, they are significantly*

more tractable. Supermetric spaces include all those governed by Euclidean, Cosine, Jensen–Shannon and Triangular distances, and are thus commonly used within many domains. In previous work we have given a generic mathematical basis for the supermetric property and shown how it can improve indexing performance for a given exact search structure. Here we present a full investigation into its use within a variety of different hyperplane partition indexing structures, and go on to show some more of its flexibility by examining a search structure whose partition and exclusion conditions are tailored, at each node, to suit the individual reference points and data set present there. Among the results given, we show a new best performance for exact search using a well-known benchmark."

3.1.4 Distributed video surveillance using smart cameras

H. Kavalionak, C. Gennaro, G. Amato, C. Vairo, C. Perciante, C. Meghini, F. Falchi. Journal of Grid Computing [58]. Abstract:

"*Video surveillance systems have become an indispensable tool for the security and organization of public and private areas. Most of the current commercial video surveillance systems rely on a classical client/server architecture to perform face and object recognition. In order to support the more complex and advanced video surveillance systems proposed in the last years, companies are required to invest resources in order to maintain the servers dedicated to the recognition tasks. In this work, we propose a novel distributed protocol for a face recognition system that exploits the computational capabilities of the surveillance devices (i.e. cameras) to perform the recognition of the person. The cameras fall back to a centralized server if their hardware capabilities are not enough to perform the recognition. In order to evaluate the proposed algorithm we simulate and test the 1NN and weighted kNN classification algorithms via extensive experiments on a freely available dataset. As a prototype of surveillance devices we have considered Raspberry PI entities. By means of simulations, we show that our algorithm is able to reduce up to 50% of the load from the server with no negative impact on the quality of the surveillance service.*" The activity presented in this paper follows up the activities we addressed in [59, 60, 61, 62].

3.1.5 Efficient Evaluation of Image Quality via Deep-Learning Approximation of Perceptual Metrics

A. Artusi, F. Banterle, F. Carrara. A. Moreo. IEEE Transaction on Image Processing [23]. Abstract:

"*Image metrics based on Human Visual System (HVS) play a remarkable role in the evaluation of complex image processing algorithms. However, mimicking the HVS is known to be complex and computationally expensive (both in terms of time and memory), and its usage is thus limited to a few applications and to small input data. All of this makes such metrics not fully attractive in real-world scenarios. To address these issues, we propose Deep Image Quality Metric (DIQM), a deep-learning approach to learn the global image quality feature (mean-opinion-score). DIQM can emulate existing visual metrics efficiently, reducing the computational costs by more than an order of magnitude with respect to existing implementations.*"

3.1.6 LSTM-based real-time action detection and prediction in human motion streams

F. Carrara. P. Elias, J. Sedmidubsky, P. Zezula. Multimedia Tools and Applications [63]. Abstract:

"*Motion capture data digitally represent human movements by sequences of 3D skeleton configurations. Such spatio-temporal data, often recorded in the stream-based nature, need to be efficiently processed to detect high-interest actions, for example, in human-computer interaction to understand hand gestures in real time. Alternatively, automatically annotated parts of a continuous stream can be persistently stored to become searchable, and thus reusable for future retrieval or pattern mining. In this paper, we focus on multi-label detection of user-specified actions in unsegmented sequences as well as continuous streams. In particular, we utilize the current advances in recurrent neural networks and adopt a unidirectional LSTM model to effectively encode the skeleton frames within the hidden network states. The model learns what subsequences of encoded frames belong to the specified action classes within the training phase. The learned representations of classes are then employed within the annotation phase to infer the probability that an incoming skeleton frame belongs to a given action class. The computed probabilities are finally compared against a learned threshold to automatically determine the beginnings and endings of actions. To further enhance the annotation accuracy, we utilize a bidirectional LSTM model to estimate class probabilities by considering not only the past frames but also the future ones. We extensively evaluate both the models on the three use cases of real-time stream annotation, offline annotation of long sequences, and early action detection and prediction. The experiments demonstrate that our models outperform the state of the art in effectiveness and are at least one order of magnitude more efficient, being able to annotate 10 k frames per second.*"

3.2 Proceedings

3.2.1 Learning Safety Equipment Detection using Virtual Worlds

M. Di Benedetto, E. Meloni, G. Amato, F. Falchi, C. Gennaro. In 2019 International Conference on Content-Based Multimedia Indexing (CBMI) [64]. Abstract:

" *Nowadays, the possibilities offered by state-of-the-art deep neural networks allow the creation of systems capable of recognizing and indexing visual content with very high accuracy. Performance of these systems relies on the availability of high quality training sets, containing a large number of examples (e.g. million), in addition to the the machine learning tools themselves. For several applications, very good training sets can be obtained, for example, crawling (noisily) annotated images from the internet, or by analyzing user interaction (e.g.: on social networks). However, there are several applications for which high quality training sets are not easy to be obtained/created. Consider, as an example, a security scenario where one wants to automatically detect rarely occurring threatening events. In this respect, recently, researchers investigated the possibility of using a visual virtual environment, capable of artificially generating controllable and photo-realistic contents, to create training sets for applications with little available training images. We explored this idea to generate synthetic photo-realistic training sets to train classifiers to recognize the proper use of individual safety equipment*

(e.g.: worker protection helmets, high-visibility vests, ear protection devices) during risky human activities. Then, we performed domain adaptation to real images by using a very small image data set of real-world photographs. We show that training with the generated synthetic training set and using the domain adaptation step is an effective solution to address applications for which no training sets exist. "

3.2.2 Testing Deep Neural Networks on the Same-Different Task

N. Messina, G. Amato, F. Carrara, F. Falchi, C. Gennaro. In 2019 International Conference on Content-Based Multimedia Indexing (CBMI) [65]. Abstract:

" *Developing abstract reasoning abilities in neural networks is an important goal towards the achievement of human-like performances on many tasks. As of now, some works have tackled this problem, developing ad-hoc architectures and reaching overall good generalization performances. In this work we try to understand to what extent state-of-the-art convolutional neural networks for image classification are able to deal with a challenging abstract problem, the so-called same-different task. This problem consists in understanding if two random shapes inside the same image are the same or not. A recent work demonstrated that simple convolutional neural networks are almost unable to solve this problem. We extend their work, showing that ResNet-inspired architectures are able to learn, while VGG cannot converge. In light of this, we suppose that residual connections have some important role in the learning process, while the depth of the network seems not so relevant. In addition, we carry out some targeted tests on the converged architectures to figure out to what extent they are able to generalize to never seen patterns. However, further investigation is needed in order to understand what are the architectural peculiarities and limits as far as abstract reasoning is concerned.* "

3.2.3 Improving Multi-Scale Face Recognition using VGGFace2

F.V. Massoli, F. Falchi, G. Amato, C. Gennaro, C. Vairo. In New Trends in Image Analysis and Processing - ICIAP 2019, BioFor International Workshop[2] [66]. Abstract:

"*Convolutional neural networks have reached extremely high performances on the Face Recognition task. These models are commonly trained by using high-resolution images and for this reason, their discrimination ability is usually degraded when they are tested against low-resolution images. Thus, Low-Resolution Face Recognition remains an open challenge for deep learning models. Such a scenario is of particular interest for surveillance systems in which it usually happens that a low-resolution probe has to be matched with higher resolution galleries. This task can be especially hard to accomplish since the probe can have resolutions as low as 8, 16 and 24 pixels per side while the typical input of state-of-the-art neural network is 224. In this paper, we described the training campaign we used to fine-tune a ResNet-50 architecture, with Squeeze-and-Excitation blocks, on the tasks of very low and mixed resolutions face recognition. For the training process we used the VGGFace2 dataset and then we tested the performance of the final model on the IJB-B dataset; in particular, we tested the neural network on the 1:1 verification task. In our experiments we considered two different scenarios: (1) probe and gallery with same resolution; (2) probe and gallery with mixed resolutions. Experimental results show that with our approach it is possible to improve upon state-of-the-art models performance on the low and mixed resolution face recognition tasks with a negligible loss at very high resolutions.*"

[2] http://www.grip.unina.it/biofor2019/

3.2.4 Learning Pedestrian Detection from VirtualWorlds

G. Amato, L. Ciampi, F. Falchi, C. Gennaro, N. Messina. In Image Analysis and Processing - ICIAP 2019, 20th International Conference, Trento, Italy, September 9–13, 2019 [67]. Abstract:

"*In this paper, we present a real-time pedestrian detection system that has been trained using a virtual environment. This is a very popular topic of research having endless practical applications and recently, there was an increasing interest in deep learning architectures for performing such a task. However, the availability of large labeled datasets is a key point for an effective train of such algorithms. For this reason, in this work, we introduced ViPeD, a new synthetically generated set of images extracted from a realistic 3D video game where the labels can be automatically generated exploiting 2D pedestrian positions extracted from the graphics engine. We exploited this new synthetic dataset fine-tuning a state-of-the-art computationally efficient Convolutional Neural Network (CNN). A preliminary experimental evaluation, compared to the performance of other existing approaches trained on real-world images, shows encouraging results.*"

3.2.5 Hebbian Learning Meets Deep Learning

G. Amato, F. Carrara, F. Falchi, C. Gennaro, G. Lagani. In Image Analysis and Processing - ICIAP 2019, 20th International Conference, Trento, Italy, September 9–13, 2019 [68]. Abstract:

"*Neural networks are said to be biologically inspired since they mimic the behavior of real neurons. However, several processes in state-of-the-art neural networks, including Deep Convolutional Neural Networks (DCNN), are far from the ones found in animal brains. One relevant difference is the training process. In state-of-the-art artificial neural networks, the training process is based on back-propagation and Stochastic Gradient Descent (SGD) optimization. However, studies in neuroscience strongly suggest that this kind of processes does not occur in the biological brain. Rather, learning methods based on Spike-Timing-Dependent Plasticity (STDP) or the Hebbian learning rule seem to be more plausible, according to neuroscientists. In this paper, we investigate the use of the Hebbian learning rule when training Deep Neural Networks for image classification by proposing a novel weight update rule for shared kernels in DCNNs. We perform experiments using the CIFAR-10 dataset in which we employ Hebbian learning, along with SGD, to train parts of the model or whole networks for the task of image classification, and we discuss their performance thoroughly considering both effectiveness and efficiency aspects.*"

3.2.6 Counting Vehicles with Deep Learning in Onboard UAV Imagery

G. Amato, L. Ciampi, F. Falchi, C. Gennaro. In International Symposium on Computers and Communications - ISCC 2019, Barcelona, Spain, June 30 - July 3, 2019 [14]. Abstract:

"*The integration of mobile and ubiquitous computing with deep learning methods is a promising emerging trend that aims at moving the processing task closer to the data source rather than bringing the data to a central node. The advantages of this approach range from bandwidth reduction, high scalability, to high reliability, just to name a few. In this paper, we pro-pose a real-time deep learning approach to automatically detect and count vehicles in videos taken from a UAV (Unmanned Aerial Vehicle). Our solution relies on a convolutional neural network-based model fine-tuned to the specific domain of applications that is able to precisely localize instances of the vehicles using a regression approach, straight from image pixels to bounding box coordinates, reasoning globally about the image when making predictions and implicitly encoding contextual information. A comprehensive experimental evaluation on real-world datasets shows that our approach results in state-of-the-art performances.Furthermore, our solution achieves real-time performances by running at a speed of 4 Frames Per Second on an NVIDIA Jetson TX2 board, showing the potentiality of this approach for real-time processing in UAVs.*"

3.2.7 A wireless smart camera network for parking monitoring

G. Amato, P. Bolettieri, D. Moroni, F.Carrara, L. Ciampi, G. Pieri, C. Gennaro, G.R. Leone, C. Vairo. In 3rd IEEE International Workshop on Cooperative Sensing for Smart MObility (COSSMO), Abu Dhabi, UAE, December 09, 2018 [13]. Abstract:

"*In this paper we present a Wireless Sensor Network(WSN), which is intended to provide a scalable solution for active cooperative monitoring of wide geographical areas.The system is designed to use different smart-camera prototypes: where the connection to the power grid is available a powerful embedded hardware implements a Deep Neural Network, otherwise a fully autonomous energy-harvesting node based on a low-energy custom board employs lightweight image analysis algorithms. Parking lots occupancy monitoring in the historical city of Lucca (Italy) is the application where the implemented smart cameras have been deployed. Traffic monitoring and surveillance are possible new scenarios for the system.*"

3.2.8 Evaluation of continuous image features learned by ODE Nets

F. Carrara, G. Amato, F. Falchi, C. Gennaro. In Image Analysis and Processing - ICIAP 2019, 20th International Conference, Trento, Italy, September 9–13, 2019 [9]. Abstract:

"*Deep-learning approaches in data-driven modeling relies on learning a finite number of transformations (and representations) of the data that are structured in a hierarchy and are often instantiated as deep neural networks (and their internal activations). State-of-the-art models for visual data usually implement deep residual learning: the network learns to predict a finite number of discrete updates that are applied to the internal network state to enrich it. Pushing the residual learning idea to the limit, ODE Net—a novel network formulation involving continuously evolving internal representations that gained the best paper award at NeurIPS 2018—has been recently proposed. Differently from traditional neural networks, in this model the dynamics of the internal states are defined by an ordinary differential equation with learnable parameters that defines a continuous transformation of the input representation. These representations can be computed using standard ODE solvers, and their dynamics can be steered to learn the input-output mapping by adjusting the ODE parameters via standard gradient-based optimization. In this work, we investigate the image representation learned in the continuous hidden states of ODE Nets. In particular, we train image classifiers including ODE-defined continuous layers and perform preliminary experiments to assess the quality, in terms of transferability and generality, of the learned image representations and compare them to standard representation extracted from residual networks. Experiments on CIFAR-10 and Tiny-ImageNet-200 datasets show that representations extracted from ODE Nets are more transferable and suggest an improved robustness to overfit.*"

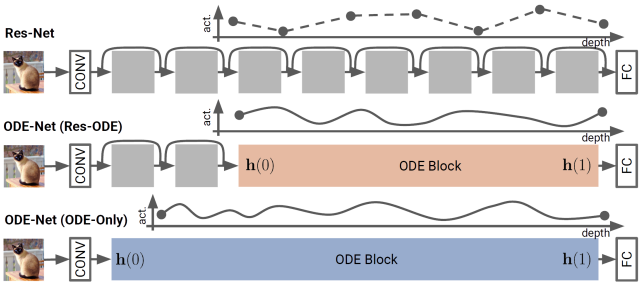


Figure 5. Architecture overview from [9].

3.2.9 On the Robustness to Adversarial Examples of Neural ODE Image Classifiers

F. Carrara, R.Caldelli, F. Falchi, G. Amato. In IEEE International Workshop on Information Forensics and Security - WIFS 2019, Delft, Netherlands, December 9–12, 2019 [6]. Abstract:

"*The vulnerability of deep neural networks to adversarial attacks currently represents one of the most challenging open problems in the deep learning field. The NeurIPS 2018 work that obtained the best paper award proposed a new paradigm for defining deep neural networks with continuous internal activations. In this kind of networks, dubbed Neural ODE Networks, a continuous hidden state can be defined via parametric ordinary differential equations, and its dynamics can be adjusted to build representations for a given task, such as image classification. In this paper, we analyze the robustness of image classifiers implemented as ODE Nets to adversarial attacks and compare it to standard deep models. We show that Neural ODE are natively more robust to adversarial attacks with respect to state-of-the-art residual networks, and some of their intrinsic properties, such as adaptive computation cost, open new directions to further increase the robustness of deep-learned models. Moreover, thanks to the continuity of the hidden state, we are able to follow the perturbation injected by manipulated inputs and pinpoint*

the part of the internal dynamics that is most responsible for the misclassification."

3.2.10 Exploiting CNN Layer Activations to Improve Adversarial Image Classification

R. Caldelli, R. Becarelli, F. Carrara, F. Falchi, G. Amato. In 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, September 22–25, 2019 [5]. Abstract:

"*Neural networks are now used in many sectors of our daily life thanks to efficient solutions such instruments provide for diverse tasks. Leaving to artificial intelligence the chance to make choices on behalf of humans inevitably exposes these tools to be fraudulently attacked. In fact, adversarial examples, intentionally crafted to fool a neural network, can dangerously induce a misclassification though appearing innocuous for a human observer. On such a basis, this paper focuses on the problem of image classification and proposes an analysis to better insight what happens inside a convolutional neural network (CNN) when it evaluates an adversarial example. In particular, the activations of the internal network layers have been analyzed and exploited to design possible countermeasures to reduce CNN vulnerability. Experimental results confirm that layer activations can be adopted to detect adversarial inputs.*"

3.2.11 Face Verification and Recognition for Digital Forensics and Information Security

G. Amato, F. Falchi, C. Gennaro, F.V. Massoli, N. Passalis, A. Tefas, A. Trivilini, C. Vairo. In 7th International Symposium on Digital Forensics and Security (ISDFS), Barcelos, Portugal, 10-12 June 2019 [69]. Abstract:

"*In this paper, we present an extensive evaluation of face recognition and verification approaches performed by the European COST Action MULTI-modal Imaging of FOREnsic SciEnce Evidence (MULTI-FORESEE). The aim of the study is to evaluate various face recognition and verification methods, ranging from methods based on facial landmarks to state-of-the-art off-the-shelf pre-trained Convolutional Neural Networks (CNN), as well as CNN models directly trained for the task at hand. To fulfill this objective, we carefully designed and implemented a realistic data acquisition process, that corresponds to a typical face verification setup, and collected a challenging dataset to evaluate the real world performance of the aforementioned methods. Apart from verifying the effectiveness of deep learning approaches in a specific scenario, several important limitations are identified and discussed through the paper, providing valuable insight for future research directions in the field.*".

3.2.12 An Image Retrieval System for Video

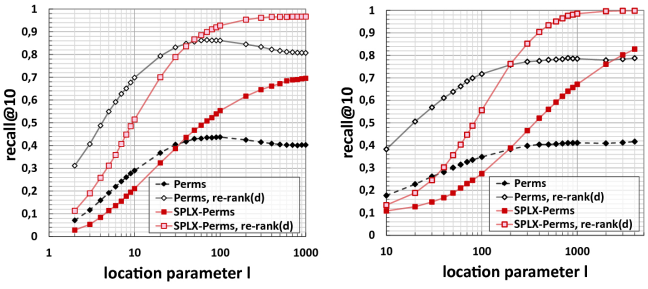
P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, C. Vairo. In Similarity Search and Applications 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11807. [31]. Abstract:

"*Since the 1970's the Content-Based Image Indexing and Retrieval (CBIR) has been an active area. Nowadays, the rapid increase of video data has paved the way to the advancement of the technologies in many different communities for the creation of Content-Based Video Indexing and Retrieval (CBVIR). However, greater attention needs to be devoted to the development of effective tools for video search and browse. In this paper, we present Visione, a system for large-scale video retrieval. The system integrates several content-based analysis and retrieval modules, including a keywords search, a spatial object-based search, and a visual similarity search. From the tests carried out by users when they needed to find as many correct examples as possible, the similarity search proved to be the most promising option. Our implementation is based on state-of-the-art deep learning approaches for content analysis and leverages highly efficient indexing techniques to ensure scalability. Specifically, we encode all the visual and textual descriptors extracted from the videos into (surrogate) textual representations that are then efficiently indexed and searched using an off-the-shelf text search engine using similarity functions.*"

3.2.13 SPLX-Perm: A Novel Permutation-Based Representation for Approximate Metric Search

L. Vadicamo, R. Connor, F. Falchi, C. Gennaro, F. Rabitti. In Similarity Search and Applications 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11807 [50]. Abstract:

"*Many approaches for approximate metric search rely on a permutation-based representation of the original data objects. The main advantage of transforming metric objects into permutations is that the latter can be efficiently indexed and searched using data structures such as inverted-files and prefix trees. Typically, the permutation is obtained by ordering the identifiers of a set of pivots according to their distances to the object to be represented. In this paper, we present a novel approach to transform metric objects into permutations. It uses the object-pivot distances in combination with a metric transformation, called n-Simplex projection. The resulting permutation-based representation, named SPLX-Perm, is suitable only for the large class of metric space satisfying the n-point property. We tested the proposed approach on two benchmarks for similarity search. Our preliminary results are encouraging and open new perspectives for further investigations on the use of the n-Simplex projection for supporting permutation-based indexing.*"



(a) SISAP colors, Jensen-Shannon dist., $n = 1,000$ (b) YFCC100M(1M), Euclidean dist., $n = 4,000$

Figure 6. *Recall*@10 varying the permutation prefix l [50].

3.2.14 Query Filtering with Low-Dimensional Local Embeddings

E. Chávez, R. Connor, L. Vadicamo. In Similarity Search and Applications 12th International Conference, SISAP 2019,

Newark, NJ, USA, October 2–4, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11807 [49]. Abstract:

"*The concept of local pivoting is to partition a metric space so that each element in the space is associated with precisely one of a fixed set of reference objects or pivots. The idea is that each object of the data set is associated with the reference object that is best suited to filter that particular object if it is not relevant to a query, maximising the probability of excluding it from a search. The notion does not in itself lead to a scalable search mechanism, but instead gives a good chance of exclusion based on a tiny memory footprint and a fast calculation. It is therefore most useful in contexts where main memory is at a premium, or in conjunction with another, scalable, mechanism.*

In this paper we apply similar reasoning to metric spaces which possess the four-point property, which notably include Euclidean, Cosine, Triangular, Jensen-Shannon, and Quadratic Form. In this case, each element of the space can be associated with two reference objects, and a four-point lower-bound property is used instead of the simple triangle inequality. The probability of exclusion is strictly greater than with simple local pivoting; the space required per object and the calculation are again tiny in relative terms.

We show that the resulting mechanism can be very effective. A consequence of using the four-point property is that, for m reference points, there are $\binom{m}{2}$ *pivot pairs to choose from, giving a very good chance of a good selection being available from a small number of distance calculations. Finding the best pair has a quadratic cost with the number of references; however, we provide experimental evidence that good heuristics exist. Finally, we show how the resulting mechanism can be integrated with a more scalable technique to provide a very significant performance improvement, for a very small overhead in build-time and memory cost.* "

3.2.15 Metric Embedding into the Hamming Space with the n-Simplex Projection

L. Vadicamo, V. Mic, P. Zezula [52]. In Similarity Search and Applications 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11807 Abstract:

"*Transformations of data objects into the Hamming space are often exploited to speed-up the similarity search in metric spaces. Techniques applicable in generic metric spaces require expensive learning, e.g., selection of pivoting objects. However, when searching in common Euclidean space, the best performance is usually achieved by transformations specifically designed for this space. We propose a novel transformation technique that provides a good trade-off between the applicability and the quality of the space approximation. It uses the n-Simplex projection to transform metric objects into a low-dimensional Euclidean space, and then transform this space to the Hamming space. We compare our approach theoretically and experimentally with several techniques of the metric embedding into the Hamming space. We focus on the applicability, learning cost, and the quality of search space approximation.*"

3.2.16 Modelling string structure in vector spaces

C. Connor, A. Dearle, L. Vadicamo. In 27th Italian Symposium on Advanced Database Systems (SEBD), Castiglione della Pescaia (GR), Italy, 16-19 June 2019. [51]. Abstract:

"*Searching for similar strings is an important and frequent database task both in terms of human interactions and in absolute world-wide CPU utilisation. A wealth of metric functions for string comparison exist. However, with respect to the wide range of classification and other techniques known within vector spaces, such metrics allow only a very restricted range of techniques. To counter this restriction, various strategies have been used for mapping string spaces into vector spaces, approximating the string distances within the mapped space and therefore allowing vector space techniques to be used. In previous work we have developed a novel technique for mapping metric spaces into vector spaces, which can therefore be applied for this purpose. In this paper we evaluate this technique in the context of string spaces, and compare it to other published techniques for mapping strings to vectors. We use a publicly available English lexicon as our experimental data set, and test two different string metrics over it for each vector mapping. We find that our novel technique considerably outperforms previously used technique in preserving the actual distance.*"

3.2.17 CNN-based system for low resolution face recognition

F. V. Massoli, G. Amato, F. Falchi, C. Gennaro, and C. Vairo. In 27th Italian Symposium on Advanced Database Systems (SEBD), Castiglione della Pescaia (GR), Italy, 16-19 June 2019. [70]. Abstract:

"*Since the publication of the AlexNet in 2012, Deep Convolutional Neural Network models became the most promising and powerful technique for image representation. Specifically, the ability of their inner layers to extract high level abstractions of the input images, called deep features vectors, has been employed. Such vectors live in a high dimensional space in which an inner product and thus a metric is defined. The latter allows to carry out similarity measurements among them. This property is particularly useful in order to accomplish tasks such as Face Recognition. Indeed, in order to identify a person it is possible to compare deep features, used as face descriptors, from different identities by means of their similarities. Surveillance systems, among others, utilize this technique. To be precise, deep features extracted from probe images are matched against a database of descriptors from known identities. A critical point is that the database typically contains features extracted from high resolution images while the probes, taken by surveillance cameras, can be at a very low resolution. Therefore, it is mandatory to have a neural network which is able to extract deep features that are robust with respect to resolution variations. In this paper we discuss a CNN-based pipeline that we built for the task of Face Recognition among images with different resolution. The entire system relies on the ability of a CNN to extract deep features that can be used to perform a similarity search in order to fulfill the face recognition task.* "

3.2.18 VISIONE at VBS2019

G. Amato, P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, C. Vairo. In 25th International Conference on MultiMedia Modeling (MMM), Thessaloniki, Greece, 8-11 January 2019. Lecture Notes in Computer Science, vol.

11296 [30]. Abstract:

"*This paper presents VISIONE, a tool for large–scale video search. The tool can be used for both known-item and ad-hoc video search tasks since it integrates several content-based analysis and retrieval modules, including a keyword search, a spatial object-based search, and a visual similarity search. Our implementation is based on state-of-the-art deep learning approaches for the content analysis and leverages highly efficient indexing techniques to ensure scalability. Specifically, we encode all the visual and textual descriptors extracted from the videos into (surrogate) textual representations that are then efficiently indexed and searched using an off-the-shelf text search engine.*"

3.2.19 Ital-IA

We participated at Ital-IA 2019, first CINI Conference on Artificial Intelligence, March, 18-19, 2019, Rome, Italy, with three short papers:

- *Intelligenza Artificiale per Ricerca in Big Multimedia Data.* F. Carrara, G. Amato, F. Debole, M. Di Benedetto, F. Falchi, C. Gennaro, N. Messina [71]
- *Intelligenza Artificiale e Analisi Visuale per la Cyber Security.* C. Vairo, G. Amato, L. Ciampi, F. Falchi, C. Gennaro, F.V. Massoli [72]
- *Intelligenza Artificiale, Retrieval e Beni Culturali.* L. Vadicamo, G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, F. Rabitti [73]

3.3 Magazines

3.3.1 About Deep Learning, Intuition and Thinking

F. Falchi. In ERCIM News 116 - Special theme: Transparency in Algorithmic Decision Making[74]. Abstract:

"*Intuition is nothing more and nothing less than recognition", is a famous quote by Herbert Simon, who received the Turing Award in 1975 and the Nobel Prize in 1978. As explained by Daniel Kahneman, another Nobel Prize winner, in his book Thinking, Fast and Slow, and during his talk at Google in 2011: "There is really no difference between the physician recognising a particular disease from a facial expression and a little child learning, pointing to something and saying doggie. The little child has no idea what the clues are but he just said, he just knows this is a dog without knowing why he knows". These milestones should be used as a guideline to help understanding decision making in recent AI algorithms and thus their transparency. [continue]*"

3.3.2 Detecting Adversarial Inputs by Looking in the black box

F. Carrara, F. Falchi, G. Amato, R. Becarelli, R. Caldelli. In ERCIM News 116 - Special theme: Transparency in Algorithmic Decision Making[4]. Abstract:

"*The astonishing and cryptic effectiveness of Deep Neural Networks comes with the critical vulnerability to adversarial inputs — samples maliciously crafted to confuse and hinder machine learning models. Insights into the internal representations learned by deep models can help to explain their decisions and estimate their confidence, which can enable us to trace, characterise, and filter out adversarial attacks. [continue]*"

3.4 Misc

3.4.1 AI in the media and creative industries

G. Amato, M. Behrmann, F. Bimbot, B. Caramiaux, F. Falchi, A. Garcia, J. Geurts, J. Gibert, G. Gravier, H. Holken, H. Koenitz, S. Lefebvre, A. Liutkus, F. Lotte, A. Perkis, R. Redondo, E. Turrin, T. Vieville, E. Vincent. Position paper [75] of the New European Media (NEM) [3] initiative. Abstract:

" *Thanks to the Big Data revolution and increasing computing capacities, Artificial Intelligence (AI) has made an impressive revival over the past few years and is now omnipresent in both research and industry. The creative sectors have always been early adopters of AI technologies and this continues to be the case. As a matter of fact, recent technological developments keep pushing the boundaries of intelligent systems in creative applications: the critically acclaimed movie" Sunspring", released in 2016, was entirely written by AI technology, and the first-ever Music Album, called" Hello World", produced using AI has been released this year. Simultaneously, the exploratory nature of the creative process is raising important technical challenges for AI such as the ability for AI-powered techniques to be accurate under limited data resources, as opposed to the conventional" Big Data" approach, or the ability to process, analyse and match data from multiple modalities (text, sound, images, etc.) at the same time. The purpose of this white paper is to understand future technological advances in AI and their growing impact on creative industries. This paper addresses the following questions: Where does AI operate in creative Industries? What is its operative role? How will AI transform creative industries in the next ten years? This white paper aims to provide a realistic perspective of the scope of AI actions in creative industries, proposes a vision of how this technology could contribute to research and development works in such context, and identifies research and development challenges.*"

3.5 Preprints

3.5.1 Detection of Face Recognition Adversarial Attacks

F.V. Massoli, F. Carrara, G. Amato, F. Falchi. Preprint uploaded on ArXiv.[76]

" *Deep Learning methods have become state-of-the-art for solving tasks such as Face Recognition (FR). Unfortunately, despite their success, it has been pointed out that these learning models are exposed to adversarial inputs - images to which an imperceptible amount of noise for humans is added to maliciously fool a neural network - thus limiting their adoption in real-world applications. While it is true that an enormous effort has been spent in order to train robust models against this type of threat, adversarial detection techniques have recently started to draw attention within the scientific community. A detection approach has the advantage that it does not require to re-train any model, thus it can be added on top of any system. In this context, we present our work on adversarial samples detection in forensics mainly focused on detecting attacks against FR systems in which the learning model is typically used*

[3] https://nem-initiative.org/

only as a features extractor. Thus, in these cases, train a more robust classifier might not be enough to defence a FR system. In this frame, the contribution of our work is four-fold: i) we tested our recently proposed adversarial detection approach against classifier attacks, i.e. adversarial samples crafted to fool a FR neural network acting as a classifier; ii) using a k-Nearest Neighbor (kNN) algorithm as a guidance, we generated deep features attacks against a FR system based on a DL model acting as features extractor, followed by a kNN which gives back the query identity based on features similarity; iii) we used the deep features attacks to fool a FR system on the 1:1 Face Verification task and we showed their superior effectiveness with respect to classifier attacks in fooling such type of system; iv) we used the detectors trained on classifier attacks to detect deep features attacks, thus showing that such approach is generalizable to different types of offensives."

3.5.2 Detection of Face Recognition Adversarial Attacks

F.V. Massoli, G. Amato, F. Falchi. Preprint uploaded on ArXiv.[7]

" *Convolutional Neural Networks have reached extremely high performances on the Face Recognition task. Largely used datasets, such as VGGFace2, focus on gender, pose and age variations trying to balance them to achieve better results. However, the fact that images have different resolutions is not usually discussed and resize to 256 pixels before cropping is used. While specific datasets for very low resolution faces have been proposed, less attention has been payed on the task of cross-resolution matching. Such scenarios are of particular interest for forensic and surveillance systems in which it usually happens that a low-resolution probe has to be matched with higher-resolution galleries. While it is always possible to either increase the resolution of the probe image or to reduce the size of the gallery images, to the best of our knowledge an extensive experimentation of cross-resolution matching was missing in the recent deep learning based literature. In the context of low- and cross-resolution Face Recognition, the contributions of our work are: i) we proposed a training method to fine-tune a state-of-the-art model in order to make it able to extract resolution-robust deep features; ii) we tested our models on the benchmark datasets IJB-B/C considering images at both full and low resolutions in order to show the effectiveness of the proposed training algorithm. To the best of our knowledge, this is the first work testing extensively the performance of a FR model in a cross-resolution scenario; iii) we tested our models on the low resolution and low quality datasets QMUL-SurvFace and TinyFace and showed their superior performances, even though we did not train our model on low-resolution faces only and our main focus was cross-resolution; iv) we showed that our approach can be more effective with respect to preprocessing faces with super resolution techniques.*"

4. Thesis

4.1 PhD Thesis

Deep Learning for Image Classification and Retrieval: Analysis and Solutions to Current Limitations

Carrara Fabio, PhD Information Science, University of Pisa, 2019 [77]. Abstract:

" *The large diffusion of cheap cameras and smartphones led to an exponential daily production of digital visual data, such as images and videos. In this context, most of the produced data lack manually assigned metadata needed for their manageability in large-scale scenarios, thus shifting the attention to the automatic understanding of the visual content. Recent developments in Computer Vision and Artificial Intelligence empowered machines with high-level vision perception enabling the automatic extraction of high-quality information from raw visual data. Specifically, Convolutional Neural Networks (CNNs) provided a way to automatically learn effective representations of images and other visual data showing impressive results in vision-based tasks, such as image recognition and retrieval. In this thesis, we investigated and enhanced the usability of CNNs for visual data management. First, we identify three main limitations encountered in the adoption of CNNs and propose general solutions that we experimentally evaluated in the context of image classification. We proposed miniaturized architectures to decrease the usually high computational cost of CNNs and enable edge inference in low-powered embedded devices. We tackled the problem of manually building huge training sets for models by proposing an automatic pipeline for training classifiers based on cross-media learning and Web-scraped weakly-labeled data. We analyzed the robustness of CNNs representations to out-of-distribution data, specifically the vulnerability to adversarial examples, and proposed a detection method to discard spurious classifications provided by the model. Secondly, we focused on the integration of CNN-based Content-based Image Retrieval (CBIR) in the most commonly adopted search paradigm, that is, textual search. We investigated solutions to bridge the gap between image search and highly-developed textual search technologies by reusing both the front-end (text-based queries) and the back-end (distributed and scalable inverted indexes). We proposed a cross-modal image retrieval approach which enables textual-based image search on unlabeled collections by learning a mapping from textual to high-level visual representations. Finally, we formalized, improved, and proposed novel surrogate text representations, i.e., text transcriptions of visual representations that can be indexed and retrieved by available textual search engines enabling CBIR without specialized indexes.* "

4.2 Master Degree Thesis

Hebbian Learning Algorithms for Training Convolutional Neural Networks

Gabriele Lagani, Master of Science in Computer Engineering, University of Pisa, 2019 [78]. Abstract:

" *The concept of Hebbian learning refers to a family of learning rules, inspired by biology, according to which the weight associated with a synapse increases proportionally to the values of the pre-synaptic and post-synaptic stimuli at a given instant of time. Different variants of Hebbian rules can be found in literature. In this thesis, three main Hebbian learning approaches are explored: Winner-Takes-All competition, Self-Organizing Maps and a supervised Hebbian learning solution for training the final classification layer of a network. In literature, applications of Hebbian learning rules to train networks for image classification tasks exist, although they are currently limited to relatively shallow architectures. In this thesis, the possibility of applying Hebbian learning rules to deeper*

network architectures is explored and the results are compared to those achieved with Gradient Descent on the same architectures. ”

Using Virtual Worlds to Train an Object Detector for Personal Protection Equipment

Enrico Meloni, Master of Science in Computer Engineering, University of Pisa, 2019 [79]. Abstract:

“ *Neural Networks are an effective technique in the field of Artificial Intelligence and in the field of Computer Vision. They learn from examples, without programming into them any previous knowledge. Deep Neural Networks saw many successful applications, thanks to the huge amount of data that is available with the growth of the internet. When annotations are not available, images are manually annotated introducing very high costs. In some contexts, gathering valuable images could be impractical for reasons related to privacy, copyright and security. To overcome these limitations the research community has taken interest in creating virtual worlds for the generation of automatically annotated training samples. In previous works, using a graphics engine for augmenting a training is shown to be a valid solution. In this work, we applied the virtual environment to approach to a not yet considered task: the detection of personal protection equipment. We developed V-DAENY, a plugin for GTA-V. With it, we generated over 140,000 automatically annotated images in several locations with different weather conditions. We manually annotated two real datasets for testing. We trained a network with this approach and evaluated its performances. We showed promising results: after training with only virtual data, the network achieves 51.8 mAP on real data and 87.2 mAP on virtual data. After applying Domain Adaptation, the network achieves 76.2 mAP on real data and 73.3 mAP on virtual data.*”

Design and Implementation of a Vehicle Tracking System Based on Deep Learning

Davide Ruisi, Master of Science in Computer Engineering, University of Pisa, 2019 [80]. Abstract: “ *Advanced on deep learning research and the availability of a lot of data to be trained, thanks to the growth of the internet, has allowed progress in many fields of computer vision, such as object detection, object tracking, and object re-identification. Tracking vehicles over multiple cameras placed at different positions is not a single task, but the composition of three distinct research problems: detection, single-camera-tracking, and re-identification. In this thesis work, we realize a system capable of tracking and re-identifying the same vehicle from different cameras using state-of-the-art approaches for detection, tracking, and re-identification. A new vehicle re-identification baseline, V-ReID-KTP-Baseline, that exploits the use of vehicle keypoints, traklets, and license plate information for re-identification, is deployed. In particular, a new re-ranking method based on license plate information is designed specifically for this task. We also present a new labeled dataset, V-ReID-AB-Dataset, created and employed to test the use of license plate information for vehicle re-identification. Test on this new dataset suggests that the availability of license plate information can make a considerable improvement in results for the task of vehicle re-identification.*”

5. Datasets

5.1 ViPeD

ViPeD[4] is a synthetically generated set of pedestrian scenarios extracted from a realistic 3D video game where the labels are automatically generated exploiting 2D pedestrian positions. It extends the JTA (Joint Track Auto) dataset [81], adding real-world camera lens effects and precise bounding box annotations useful for pedestrian detection.

The above dataset was used for our work on pedestrian detection, with a published paper at the International Conference on Image Analysis and Processing (ICIAP) 2019 (see Section 3.2.4).

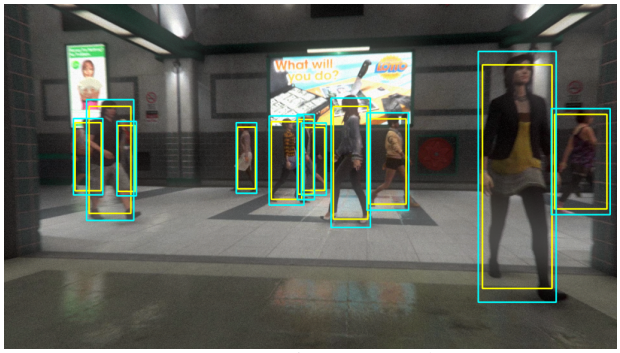


Figure 7. ViPeD example.

5.2 VW-PPE

Virtual World dataset: The virtual world training set was generated using a plugin for GTA V. It is composed of around 140,000 images, automatically annotated. There are 7 object classes: Bare Head, Helmet, Ear Protection, Welding Mask, Bare Chest, High Visibility Vest, Person. The images have been generated in 10 different locations of the game map, with three weather and time variations for each location. Each image contains an average of 12 people with or without personal protection equipment. The archive is ∼131GB, each image is 1088×612. The directory structure is kept as it was generated by the plugin. Inside there are two txt files which divide the dataset in training and validation.

Real World dataset: The real world dataset was manually created by the authors. It is composed of around 220 images, all copyright free and manually annotated. The same object classes are used. The archive is ∼220MB, each image has different size and resolution. Inside there are two directories that separate the images used also for domain adaptation from those used only for testing. You can find also two files, images.txt and pages.txt which provide due attribution to all images.

The above datasets were used for our work on protection equipment detection, with a published paper at the Content-Based Media Indexing (CBMI) 2019 conference (see Section 3.2.1).

[4] http://aimir.isti.cnr.it/viped/

Figure 8. VW-PPE example.

6. Code

6.1 RelationNetworks-CLEVR

A pytorch implementation for "A simple neural network module for relational reasoning", working on the CLEVR dataset. The code, maintained by Nicola Messina, is publicly available on GitHub[5].

6.2 Hebbian Learning

A Pytorch implementation for the thesis work [78] and the paper [68] on Hebbian learning is publicly available on GitHub[6].

6.3 Neural ODE Image Classifiers

The Pytorch code reproducing [6, 9] is publicly available on GitHub[7].

7. Awards

7.1 CBMI 2019 Best Paper

The paper "Learning Safety Equipment Detection using Virtual Worlds", M. Di Benedetto, E. Meloni, G. Amato, F. Falchi, C. Gennaro, received the best paper awards in 2019 International Conference on Content-Based Multimedia Indexing (CBMI).

Figure 9. CBMI Best Paper Awards for [64].

[5] https://github.com/mesnico/RelationNetworks-CLEVR
[6] https://github.com/GabrieleLagani/HebbianLearningThesis
[7] https://github.com/fabiocarrara/neural-ode-features

Acknowledgments

This work was partially funded by: "AI4EU", project, funded by EC (H2020 - Contract n. 825619); "ADA, Automatic Data and documents Analysis to enhance human-based processes", CUP CIPE D55F17000290009; "VIDEMO, Visual Deep Engines for Monitoring" and "VISECH, Visual Engines for Cultural Heritage" parts of the ARCO-CNR, cofounded by the Tuscany region under POR FSE 2014-2020, CUP CIPE B56J17001330004; "Smart News, Social sensing for breakingnews", cofounded by the Tuscany region under the FAR FAS 2014 program, CUP CIPE D58C15000270008; We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for this research.

References

[1] Fabio Carrara, Fabrizio Falchi, Roberto Caldelli, Giuseppe Amato, Roberta Fumarola, and Rudy Becarelli. Detecting adversarial example attacks to deep neural networks. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, CBMI '17, pages 38:1–38:7, New York, NY, USA, 2017. ACM.

[2] Fabio Carrara, Rudy Becarelli, Roberto Caldelli, Fabrizio Falchi, and Giuseppe Amato. Adversarial examples detection in features distance spaces. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 313–327, Cham, 2019. Springer International Publishing.

[3] Fabio Carrara, Fabrizio Falchi, Roberto Caldelli, Giuseppe Amato, and Rudy Becarelli. Adversarial image detection in deep neural networks. *Multimedia Tools and Applications*, Mar 2018.

[4] Fabio Carrara, Fabrizio Falchi, Giuseppe Amato, Rudy Becarelli, and Roberto Caldelli. Detecting adversarial inputs by looking in the black box. *ERCIM News*, January 2019.

[5] R. Caldelli, R. Becarelli, F. Carrara, F. Falchi, and G. Amato. Exploiting cnn layer activations to improve adversarial image classification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2289–2293, Sep. 2019.

[6] Fabio Carrara, Roberto Caldelli, Giuseppe Amato, and Fabrizio Falchi. On the robustness to adversarial examples of neural ode image classifiers. In *2019 IEEE International Workshop on Information Forensics and Security*, 2019. to appear.

[7] Fabio Valerio Massoli, Fabio Carrara, Giuseppe Amato, and Fabrizio Falchi. Detection of face recognition adversarial attacks, 2019.

[8] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[9] Fabio Carrara, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. Evaluation of continuous image features learned by ode nets. In Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 432–442, Cham, 2019. Springer International Publishing.

[10] Manuela Piazza, Andrea Mechelli, Brian Butterworth, and Cathy J. Price. Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuroImage*, 15(2):435 – 446, 2002.

[11] Daniel Hyde. Two systems of non-symbolic numerical cognition. *Frontiers in Human Neuroscience*, 5:150, 2011.

[12] Luca Ciampi, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Rabitti Fausto. Counting vehicles with cameras. In *2018 Italian Symposium on Advanced Database Systems*, 2018.

[13] Giuseppe Amato, Paolo Bolettieri, Davide Moroni, Fabio Carrara, Luca Ciampi, Gabriele Pieri, Claudio Gennaro, Giuseppe Riccardo Leone, and Claudio Vairo. A wireless smart camera network for parking monitoring. In *3rd International Workshop on Cooperative Sensing for Smart MObility (COSSMO)*, pages 1–6. IEEE, 2018.

[14] Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, and Claudio Gennaro. Counting vehicles with deep learning in onboard uav imagery. In *2019 IEEE International Symposium on Computers and Communications*, 2019. to appear.

[15] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006.

[16] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[17] Giuseppe Amato, Paolo Barsocchi, Fabrizio Falchi, Erina Ferro, Claudio Gennaro, Giuseppe Leone, Davide Moroni, Ovidio Salvetti, and Claudio Vairo. Towards multimodal surveillance for smart building security. In *6th International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, Multidisciplinary Digital Publishing Institute Proceedings, vol. 2, no. 2, pages 95:1–95:8, 2018.

[18] Paolo Barsocchi, Antonello Calabrò, Erina Ferro, Claudio Gennaro, Eda Marchetti, and Claudio Vairo. Boosting a low-cost smart home environment with usage and access control rules. *Sensors*, 18(6):1886:1– 1886:22, 2018.

[19] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. Facial-based intrusion detection system with deep learning in embedded devices. In *1st International Conference on Sensors, Signal and Image Processing (SSIP)*, pages 64–68. ACM, 2018.

[20] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. A comparison of face verification with facial landmarks and deep features. In *10th International Conference on Advances in Multimedia (MMEDIA)*, pages 1–6, 2018.

[21] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. Car parking occupancy detection using smart camera networks and Deep Learning. In *21st IEEE Symposium on Computers and Communication (ISCC)*, pages 1212–1217. IEEE, 2016.

[22] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo. Deep Learning for decentralized parking lot occupancy detection. *Expert Systems with Applications*, 72:327–334, 2017.

[23] Alessandro Artusi, Francesco Banterle, Fabio Carra, and Alejandro Moreno. Efficient evaluation of image quality via deep-learning approximation of perceptual metrics. *IEEE Transactions on Image Processing*, 29:1843–1855, 2019.

[24] Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Fausto Rabitti. Similarity caching in large-scale image retrieval. *Information Processing & Management*, 48(5):803 – 818, 2012.

[25] Giuseppe Amato, Andrea Esuli, and Fabrizio Falchi. A comparison of pivot selection techniques for permutation-based indexing. *Information Systems*, 52:176 – 188, 2015. Special Issue on Selected papers from SISAP 2013.

[26] Giuseppe Amato, Franca Debole, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. *Large Scale Indexing and Searching Deep Convolutional Neural Network Features*, pages 213–224. Springer International Publishing, Cham, 2016.

[27] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. *Deep Permutations: Deep Convolutional Neural Networks and Permutation-Based Indexing*, pages 93–106. Springer International Publishing, Cham, 2016.

[28] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. *YFCC100M-HNfc6: A Large-Scale Deep Features Benchmark for Similarity Search*, pages 196–209. Springer International Publishing, Cham, 2016.

[29] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, and Claudio Gennaro. Large-scale image retrieval with elasticsearch. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 925–928, New York, NY, USA, 2018. ACM.

[30] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. Visione at vbs2019. In *MultiMedia Modeling*, pages 591–596, Cham, 2019. Springer International Publishing.

[31] Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. An image retrieval system for video. In *Similarity Search and Applications*, pages 332–339, Cham, 2019. Springer International Publishing.

[32] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad. On influential trends in interactive video retrieval: Video browser showdown 2015-2017. *IEEE Transactions on Multimedia*, pages 1–1, 2018.

[33] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. Searching and Annotating 100M Images with YFCC100M-HNfc6 and MI-File. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, CBMI '17, pages 26:1–26:4. ACM, 2017.

[34] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[35] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[36] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[37] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[38] George Awad, Cees G. M. Snoek, Alan F. Smeaton, and Georges Quénot. Trecvid semantic indexing of video : A 6-year retrospective. *ITE Transactions on Media Technology and Applications*, 4(3):187–208, 2016.

[39] G. Amato, C. Gennaro, and P. Savino. MI-File: Using inverted files for scalable approximate similarity search. *Multimedia Tools and Applications*, (3):1333–1362, 2014.

[40] Giuseppe Amato, Fabrizio Falchi, Fausto Rabitti, and Lucia Vadicamo. Some theoretical and experimental observations on permutation spaces and similarity search. In *Similarity Search and Applications*, pages 37–49, Cham, 2014. Springer International Publishing.

[41] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. Deep permutations: deep convolutional neural networks and permutation-based indexing. In *International Conference on Similarity Search and Applications*, pages 93–106. Springer, 2016.

[42] Giuseppe Amato, Edgar Chávez, Richard Connor, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. Reranking permutation-based candidate sets with the n-simplex projection. In *Similarity Search and Applications*, pages 3–17, Cham, 2018. Springer International Publishing.

[43] Richard Connor, Lucia Vadicamo, Franco Alberto Cardillo, and Fausto Rabitti. Supermetric search with the four-point property. In *Similarity Search and Applications*, pages 51–64, Cham, 2016. Springer International Publishing.

[44] Richard Connor, Lucia Vadicamo, and Fausto Rabitti. High-dimensional simplexes for supermetric search. In *Similarity Search and Applications*, pages 96–109, Cham, 2017. Springer International Publishing.

[45] Vladimir Mic, David Novak, Lucia Vadicamo, and Pavel Zezula. Selecting sketches for similarity search. In *European Conference on Advances in Databases and Information Systems*, pages 127–141. Springer, 2018.

[46] L. M. Blumenthal. *Theory and applications of distance geometry*. Clarendon Press, 1953.

[47] Richard Connor, Franco Alberto Cardillo, Lucia Vadicamo, and Fausto Rabitti. Hilbert exclusion: improved metric search through finite isometric embeddings. *ACM Transactions on Information Systems (TOIS)*, 35(3):17, 2017.

[48] Richard Connor, Lucia Vadicamo, Franco Alberto Cardillo, and Fausto Rabitti. Supermetric search. *Information Systems*, 80:108 – 123, 2019.

[49] Edgar Chávez, Richard Connor, and Lucia Vadicamo. Query filtering with low-dimensional local embeddings. In *Similarity Search and Applications*, pages 233–246, Cham, 2019. Springer International Publishing.

[50] Lucia Vadicamo, Richard Connor, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. Splx-perm: A novel permutation-based representation for approximate metric search. In *Similarity Search and Applications*, pages 40–48, Cham, 2019. Springer International Publishing.

[51] Richard Connor, Al Dearle, and Lucia Vadicamo. Modelling string structure in vector spaces. In *Proceedings of the 27th Italian Symposium on Advanced Database Systems*, 2019.

[52] Lucia Vadicamo, Vladimir Mic, Fabrizio Falchi, and Pavel Zezula. Metric embedding into the hamming space with the n-simplex projection. In *Similarity Search and Applications*, pages 265–272, Cham, 2019. Springer International Publishing.

[53] Nicola Messina, Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, and Claudio Gennaro. Learning relationship-aware visual features. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[54] Nicola Messina, Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, and Claudio Gennaro. Learning visual features for relational cbir. *International Journal of Multimedia Information Retrieval*, Sep 2019.

[55] Fabio Carrara, Franca Debole, Claudio Gennaro, and Giuseppe Amato. Image analysis in technical documentation (discussion paper). 2019.

[56] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. Large-scale instance-level image retrieval. *Information Processing & Management*, page 102100, 2019.

[57] Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. Combining local and global visual feature similarity using a text search engine. In *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 49–54. IEEE, 2011.

[58] Hanna Kavalionak, Claudio Gennaro, Giuseppe Amato, Claudio Vairo, Costantino Perciante, Carlo Meghini, and Fabrizio Falchi. Distributed video surveillance using smart cameras. *Journal of Grid Computing*, 17(1):59–77, 2019.

[59] Claudio Vairo, Giuseppe Amato, Stefano Chessa, and Paolo Valleri. Modeling detection and tracking of complex events in Wireless Sensor Networks. In *2010 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 235–242. IEEE, 2010.

[60] Giuseppe Amato, Stefano Chessa, Claudio Gennaro, and Claudio Vairo. Efficient detection of composite events in Wireless Sensor Networks: design and evaluation. In *16th IEEE Symposium on Computers and Communications (ISCC)*, pages 821–823. IEEE, 2011.

[61] Giuseppe Amato, Stefano Chessa, Claudio Gennaro, and Claudio Vairo. Dynamic tracking of composite events in Wireless Sensor Networks. In *4th International Conference on Ad Hoc Networks (ADHOCNETS)*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST), volume 111, pages 72–86. Springer, 2012.

[62] Giuseppe Amato, Stefano Chessa, Claudio Gennaro, and Claudio Vairo. Querying moving events in Wireless Sensor Networks. *Pervasive and Mobile Computing*, 16:51–75, 2015.

[63] Fabio Carrara, Petr Elias, Jan Sedmidubsky, and Pavel Zezula. Lstm-based real-time action detection and prediction in human motion streams. *Multimedia Tools and Applications*, pages 1–23, 2019.

[64] M. Di Benedetto, E. Meloni, G. Amato, F. Falchi, and C. Gennaro. Learning safety equipment detection using virtual worlds. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 8–13, Sep. 2019.

[65] N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro. Testing deep neural networks on the same-different task. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 38–43, Sep. 2019.

[66] Fabio Valerio Massoli, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. Improving multi-scale face recognition using vggface2. In Marco Cristani, Andrea Prati, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *New Trends in Image Analysis and Processing – ICIAP 2019*, pages 21–29, Cham, 2019. Springer International Publishing.

[67] Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Nicola Messina. Learning pedestrian detection from virtual worlds. In Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 302–312, Cham, 2019. Springer International Publishing.

[68] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Gabriele Lagani. Hebbian learning meets deep convolutional neural networks. In Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 324–334, Cham, 2019. Springer International Publishing.

[69] G. Amato, F. Falchi, C. Gennaro, F. V. Massoli, N. Passalis, A. Tefas, A. Trivilini, and C. Vairo. Face verification and recognition for digital forensics and information security. In *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6, June 2019.

[70] Fabio Valerio Massoli, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. CNN-based system for low resolution face recognition. In *to appear in 27th Italian Symposium on Advanced Database Systems (SEBD)*, 2019.

[71] Fabio Carrara, Giuseppe Amato, Franca Debole, Marco Di Benedetto, Fabrizio Falchi, Claudio Gennaro, and Nicola Messina. Intelligenza artificiale per ricerca in big multimedia data. In *Ital-IA*. CINI, 2019.

[72] Lucia Vadicamo, Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. Intelligenza artificiale, retrieval e beni culturali. In *Ital-IA*. CINI, 2019.

[73] Claudio Vairo, Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Fabio Valerio Massoli. Intelligenza artificiale e analisi visuale per la cyber security. In *Ital-IA*. CINI, 2019.

[74] Fabrizio Falchi. About deep learning, intuition and thinking. *ERCIM News*, January 2019.

[75] Baptiste Caramiaux, Fabien Lotte, Joost Geurts, Giuseppe Amato, Malte Behrmann, Frédéric Bimbot, Fabrizio Falchi, Ander Garcia, Jaume Gibert, Guillaume Gravier, et al. Ai in the media and creative industries. 2019.

[76] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition, 2019.

[77] Fabio Carrara. *Deep Learning for Image Classification and Retrieval: Analysis and Solutions to Cur-*

rent Limitations. PhD thesis, Dottorato in Ingegneria dell'informazione,, University of Pisa, Italy, 2019.

[78] Gabriele Lagani. Hebbian learning algorithms for training convolutional neural networks. Master's thesis, Computer Engineering, University of Pisa, Italy, 2019.

[79] Enrico Meloni. Using virtual worlds to train an object detector for personal protection equipment. Master's thesis, Computer Engineering, University of Pisa, Italy, 2019.

[80] Davide Ruisi. Design and implementation of a vehicle tracking system based on deep learning. Master's thesis, Computer Engineering, University of Pisa, Italy, 2019.

[81] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*, 2018.