AIMH Lab 2022 Activities for Vision

Luca Ciampi[1,*], Giuseppe Amato[1], Paolo Bolettieri[1], Fabio Carrara[1], Marco Di Benedetto[1], Fabrizio Falchi[1], Claudio Gennaro[1], Nicola Messina[1], Lucia Vadicamo[1] and Claudio Vairo[1]

[1]ISTI-CNR, via G. Moruzzi, 1, Pisa, 56100, Italy

Abstract

The explosion of smartphones and cameras has led to a vast production of multimedia data. Consequently, Artificial Intelligence-based tools for automatically understanding and exploring these data have recently gained much attention. In this short paper, we report some activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR, tackling some challenges in the field of Computer Vision for the automatic understanding of visual data and for novel interactive tools aimed at multimedia data exploration. Specifically, we provide innovative solutions based on Deep Learning techniques carrying out typical vision tasks such as object detection and visual counting, with particular emphasis on scenarios characterized by scarcity of labeled data needed for the supervised training and on environments with limited power resources imposing miniaturization of the models. Furthermore, we describe VISIONE, our large-scale video search system designed to search extensive multimedia databases in an interactive and user-friendly manner.

Keywords

Computer Vision, Multimedia Understanding, Deep Learning, Large-scale Video Retrieval, Learning with Scarce Data

1. Introduction

The pervasive diffusion of smartphones and cheap cameras leads to an exponential daily production of digital visual data, such as images and videos. In this context, a constant increase of attention to the automatic understanding of this visual content is occurring. Hence, Computer Vision has become one of the hottest fields that make extensive use of Artificial Intelligence (AI), to such a point that some applications are now parts of our everyday lives, and they are making human life easier. Some examples include pedestrian detection and human activity monitoring in surveillance systems or face detection and recognition in smartphones. Furthermore, an important consequence of dealing with these large quantities of data coming from different sources is the need to efficiently and effectively organize them so that also non-expert users can easily manage and browse them.

This paper presents some research topics and applications carried out by the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR, focusing on multimedia understanding and novel interactive software for multimedia data exploration. Specifically, we introduce some innovative solutions relying on Deep Learning (DL) techniques that fulfill some popular and typical Computer Vision tasks, such as object detection, crowd counting, and human activity monitoring, applied in multi-disciplinary areas ranging from agriculture and smart surveillance to biology and smart parking, paying special attention to some interesting and irksome challenges concerning the lack of labeled training data and the adoption of the so-called Edge-AI paradigm that imposes the use of environments with limited power resources and the consequently the miniaturization of the DL models. Furthermore, we propose a large-scale video search system designed to browse massive multimedia databases with an interactive and user-friendly human-machine interface. We pay special attention to some interesting and irksome challenges concerning the lack of labeled training data and the adoption of Deep Learning (DL)-based techniques in environments with limited power resources.

2. Research Areas

2.1. Learning with Scarce Data

Current vision systems powered by data-driven AI methods suffer from strong domain shifts and are not invariant to substantial context variations. Indeed, these AI technologies usually need a massive amount of annotated data required for the supervised learning phase, and they often suffer when applied to unseen data. Consequently, adopting these solutions is often dampened for large-scale contexts, considering that the data annotation procedure requires extraordinary human effort and collecting data for every specific scenario is unfeasible. The

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

luca.ciampi@isti.cnr.it (L. Ciampi)

0000-0002-6985-0439 (L. Ciampi); 0000-0003-0171-4315 (G. Amato); 0000-0002-5225-4278 (P. Bolettieri); 0000-0001-5014-5089 (F. Carrara); 0000-0001-5781-7060 (M. Di Benedetto); 0000-0001-6258-5313 (F. Falchi); 0000-0002-3715-149X (C. Gennaro); 0000-0003-3011-2487 (N. Messina); 0000-0001-7182-7038 (L. Vadicamo); 0000-0000-0003-2740-4331 (C. Vairo)

AIMH Lab is tackling this challenge from several sides, offering different approaches detailed in the following.

2.1.1. Learning from Synthetic Data

An appealing solution to mitigate the human effort needed for manual annotation is to gather synthetic data from virtual environments resembling the real world, where the labels are *automatically* collected by interacting with the graphical engine. In this context, we released and exploited several synthetic collections of images to build DL solutions that carry out several human-centered tasks. In particular, in [1], we presented an embedded modular AI-assisted Computer Vision-based system that provides many functionalities to help monitor individual and collective human safety rules, ranging from social distance estimation and crowd counting to Personal Protective Equipment (PPE) detection (such as helmets and masks). Our solution consists of multiple modules relying on neural network components, each responsible for specific functionalities that users can easily enable, configure, and combine. One of the main peculiarity is that some of these components have been trained by exploiting synthetic data collected from the GTAV videogame and automatically annotated [2] [3] [4] [5]. Furthermore, we employed the GTAV videogame also for gathering other collections of images and labels to train a DL-based approach for human fall detection [6] and a technique for multi-camera vehicle tracking in urban scenarios. Finally, more recently, we proposed *CrowdSim2*, a new synthetic collection of images suitable for people and vehicle detection and tracking gathered from a simulator based on the *Unity* graphical engine [7] [8] consisting of thousands of images collected from various synthetic scenarios resembling the real world, where we varied some factors of interest, such as the weather conditions and the number of objects in the scenes.

2.1.2. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) is a technique that addresses the Domain Shift problem by taking a source labeled dataset and a target *unlabeled* one. The challenge here is to automatically infer some knowledge from the target data to reduce the gap between the two domains. In [9] and [10], the AIMH Lab introduced an end-to-end CNN-based UDA algorithm for traffic density estimation and counting, based on adversarial learning performed directly on the output space. We validated our approach over different types of domain shifts, i.e., the *Camera2Camera*, the *Day2Night*, and the *Synthetic2Real* domain shifts, demonstrating significant improvement compared to the performance of the model without domain adaptation. Furthermore, very recently, we also proposed a UDA scheme for video violence detection based on single-image classification [11], which can mitigate the domain gap between annotated datasets containing violent/non-violent clips in general contexts and a recently introduced collection of videos specific for detection of violent behaviors in public transport [12].

2.1.3. Learning from multi-rating data

Often, non-trivial patterns produce a non-negligible disagreement between multiple annotators, such as when dealing with biological structures in microscopy images. A possible solution to have more robust labels is to aggregate and average the decisions given by several annotators to the same data. However, the scale of many tasks prevents the creation of large datasets annotated by several experts, i.e., annotators prefer to label new data rather than label the same data more than once, resulting in large, single-labeled weakly labeled datasets and very small multi-labeled data, from which it is crucial to make the most. In [13], we proposed a two-stage counting strategy in a weakly labeled data scenario. In the first stage, we trained state-of-the-art DL-based methodologies to detect and count biological structures exploiting a large set of single-labeled data sure to contain errors; in the second stage, using a small set of multi-labeled data, we refined the predictions, increasing the correlation between the scores assigned to the samples and the agreement of the raters on the annotations, i.e., we improved confidence calibration by taking advantage of the redundant information characterizing the multi-labeled data. Furthermore, we are currently exploring the possibility of exploiting multi-labeled data from annotations automatically generated by several state-of-the-art detectors.

2.2. Smart Parking on the Edge

Traffic-related issues are constantly increasing, and tomorrow's cities can be considered intelligent only if they provide smart mobility applications, such as smart parking and traffic management. In this context, city camera networks represent the perfect tool for monitoring large urban areas while providing visual data to AI systems responsible for extracting relevant information and suggesting/making decisions helpful for intelligent mobility applications. However, implementing these solutions is often hampered by the massive flow of data that must be sent to central servers or the cloud for processing. On the other hand, the recent paradigm of edge computing promotes the decentralization of data processing to the border, i.e., where the data are gathered, thus reducing the traffic on the network and the pressure on central servers. Nonetheless, this promising standard brings along with it also some new challenges related to the limited computational resources on the disposable edge

devices and also concerning security inside IoT networks.

The AIMH Lab proposed and is actively researching DL-based solutions for intelligent parking monitoring matching the edge AI idea, i.e., that can run directly onboard embedded vision systems equipped with limited computational capabilities able to capture images, process them, and eventually communicate with other devices sending the elaborated information. Specifically, in [14] and [15], we introduced a decentralized and efficient solution for visual parking lot occupancy detection, which exploits a miniaturized CNN to classify parking space occupancy. It runs directly onboard smart cameras built using the Raspberry Pi platform equipped with a camera module. On the other hand, in [16] and [17], we extended this application by proposing a DL-based method that can instead estimate the number of vehicles present in the Field Of View of the smart cameras. Such a task is more flexible than the previous one since it does not rely on meta-information regarding the monitored scene, such as the position of the parking lots. More, in [18], we proposed a DL solution to automatically detect and count vehicles in images taken from a camera-equipped drone directly onboard the UAV. More recently, we proposed a multi-camera system capable of automatically estimating the number of cars in an *entire* parking lot directly on board the edge devices [19]. The peculiarity of this solution is that, unlike most of the works in the literature which focus on the analysis of single images, it uses multiple visual sources to monitor a wider parking area from different perspectives. More in detail, it comprises an on-device DL-based detector that locates and counts the vehicles from the captured images of a single smart camera together with a decentralized geometric-based approach that can analyze the inter-camera shared areas and merge the data acquired by all the devices.

2.3. Object Detection and Visual Counting in Multi-disciplinary Areas

The AIMH Lab is currently researching the field of object detection and crowd counting, proposing novel solutions in multi-disciplinary areas.

In [20], we collected and publicly released *MOBDrone*, a large-scale dataset of aerial footage of people who, being in the water, simulated the need to be rescued. This data includes 66 video clips with 126,170 frames manually annotated having more than 180K bounding boxes (of which more than 113K belong to the *person* category). We provide a sample of our dataset in Figure 1. Furthermore, we presented an in-depth experimental analysis of the performance of several state-of-the-art object detectors over this newly established scenario. In [21], we introduced a Computer Vision tool in the Smart Agriculture area aimed at automatically counting pests in pictures of sticky paper traps; controlling pest population is crucial in agriculture, and an effective Integrated Pest Management tool can prevent crop damage and suggest corrective measures to keep pests from causing significant problems. On the other hand, in [22], we tackled the problem of counting cells in microscopy images, a fundamental step in diagnosing several diseases in biology and medicine. Finally, in [23], we proposed a smart-surveillance application for crowd counting in videos gathered from city cameras. Here, we also considered the temporal context relying on the evidence that when instances of particular objects are moving – like persons or cars – it is usually easier to spot their presence and consequently count them with greater accuracy. Specifically, we introduced a transformer-based attentive mechanism where the movement flows are initially estimated through a network trained by enforcing person-corservation laws – no persons can suddenly disappear between consecutive frames in the video except at the frame borders. Then, the person flow is integrated to get the actual people count. This solution obtained state-of-the-art results, surpassing frame-based visual counting networks.

Figure 1: A sample of our MobDrone dataset for man overboard detection from drones.

2.4. The VISIONE Search System

With the increasing diffusion of multimedia databases, there is today, as never before, the need to analyze, organize, and index all the produced data so that they can be quickly and efficiently retrieved. The development of tools to automatically analyze and index all these contents constitutes a significant achievement in the automatic content-based organization and browsing of large multimedia databases. Exciting applications of these technologies are, for example, the organization of raw multimedia data scraped from the web or the browsing of audiovisual archives owned by national televisions.

To fulfill these needs, the AIMH Lab is developing VISIONE [24, 25, 26, 27], a large-scale video search system designed to search extensive multimedia databases in an interactive and user-friendly manner. VISIONE

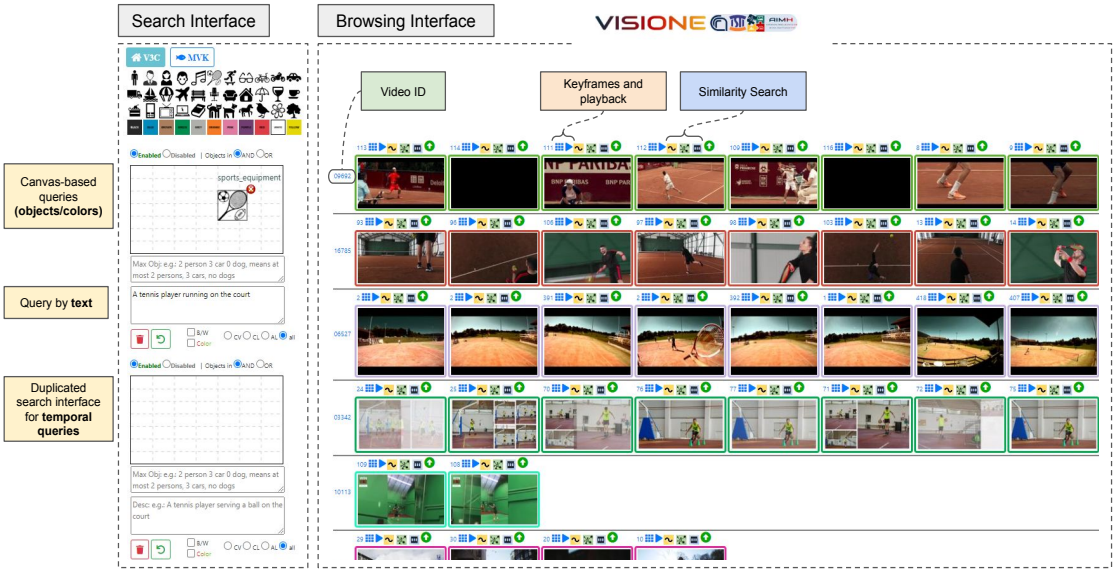

Figure 2: User Interface of the VISIONE system.

employs various content-based analysis tools to extract knowledge from raw shots, and it uses reliable indexing techniques for achieving good scalability. The system offers advanced search functionalities powered by employing publicly-available state-of-the-art image analysis models and technologies developed internally for advanced multimedia representation and large-scale indexing. Specifically, VISIONE enables the search for video shots based on specific object classes, employing a canvas-oriented interface that enables users to specify objects and colors in particular positions within the frame. The latest versions of VISIONE feature state-of-the-art cross-modal models able to search keyframes and videos using natural language prompts. Specifically, VISIONE integrates some CLIP-based models[28, 29], as well as a novel cross-modal retrieval deep neural network, called ALADIN (ALign And DIstill Network) [30]. This network generates fixed-length features in a common visual-textual space by distilling the relevance scores from large pre-trained vision-language transformers, enabling quick and accurate keyframe retrieval using detailed natural language prompts. VISIONE also provides visual similarity techniques to help users browse results and find keyframes similar to the selected shots based on instance or semantic similarities. An overview of the interface is shown in Figure 2. Finally, VISIONE supports temporal queries, allowing the user to specify two independent queries used for searching videos containing two keyframes satisfying the two queries but having a temporal distance smaller than 12 seconds. This enables users to easily search for long-lasting specific actions or videos containing particular scene cuts.

The system employs two different indexes to store and perform similarity search on the extracted visual features and the detected objects and colors, Apache Lucene[1]and FAISS[2]. The need for two indexes is motivated by their different functionalities and implementations. Lucene, in particular, is disk-based and designed to handle billions of documents, scaling better than in-memory indexes like FAISS. Lucene is commonly used for text-based search in collections of long unstructured text documents. To encode image features for similarity search, we developed a family of techniques called Surrogate Text Representations (STRs) [31, 32] to enable dense features to be transformed into sparse term frequencies from an appropriate codebook.

The system participated in the 12th Video Browser Showdown competition [33], where it ranked second in the overall leaderboard and performed well in several subtasks, achieving first place in visual known item search.

3. Conclusions

In this short paper, we reported some activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR concerning Computer Vision approaches for multimedia understanding as well as solutions for interactive tools aimed at exploring multimedia

[1]https://lucene.apache.org/
[2]https://github.com/facebookresearch/faiss

data. The widely spread of multimedia data, such as images and videos gathered from smartphones or smart cameras, is driving the research of these tools. Indeed, this deluge of visual data needs more and more AI-based techniques able to automatically understand and browse it. Specifically, we described some particularly interesting challenges we are tackling, such as the Deep Learning methods operating in contexts of scarce data or in limited-powered environments, as well as systems designed to search extremely large video databases with interactive and user-friendly interfaces.

Acknowledgments

This work was partially supported by: the AI4Media project, funded by the EC (H2020 - Contract n. 951911); PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] M. D. Benedetto, F. Carrara, L. Ciampi, F. Falchi, C. Gennaro, G. Amato, An embedded toolset for human activity monitoring in critical environments, Expert Systems with Applications 199 (2022) 117125. doi:10.1016/j.eswa.2022.117125.

[2] L. Ciampi, N. Messina, F. Falchi, C. Gennaro, G. Amato, Virtual to real adaptation of pedestrian detectors, Sensors 20 (2020) 5250. doi:10.3390/s20185250.

[3] G. Amato, L. Ciampi, F. Falchi, C. Gennaro, N. Messina, Learning pedestrian detection from virtual worlds, in: Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 302–312. doi:10.1007/978-3-030-30642-7_27.

[4] M. di Benedetto, E. Meloni, G. Amato, F. Falchi, C. Gennaro, Learning safety equipment detection using virtual worlds, in: 2019 International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, 2019. doi:10.1109/cbmi.2019.8877466.

[5] M. D. Benedetto, F. Carrara, E. Meloni, G. Amato, F. Falchi, C. Gennaro, Learning accurate personal protective equipment detection from virtual worlds, Multimedia Tools and Applications 80 (2020) 23241–23253. doi:10.1007/s11042-020-09597-9.

[6] F. Carrara, L. Pasco, C. Gennaro, F. Falchi, Learning to detect fallen people in virtual worlds, in: International Conference on Content-based Multimedia Indexing, ACM, 2022. doi:10.1145/3549555.3549573.

[7] P. Foszner, A. Szczęsna, L. Ciampi, N. Messina, A. Cygan, B. Bizoń, M. Cogiel, D. Golba, E. Macioszek, M. Staniszewski, CrowdSim2: An open synthetic benchmark for object detectors, in: Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SCITEPRESS - Science and Technology Publications, 2023. doi:10.5220/0011692500003417.

[8] P. Foszner, A. Szczęsna, L. Ciampi, N. Messina, A. Cygan, B. Bizoń, M. Cogiel, D. Golba, E. Macioszek, M. Staniszewski, Development of a realistic crowd simulation environment for fine-grained validation of people tracking methods, in: Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SCITEPRESS - Science and Technology Publications, 2023. doi:10.5220/0011691500003417.

[9] L. Ciampi, C. Santiago, J. Costeira, C. Gennaro, G. Amato, Domain adaptation for traffic density estimation, in: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SCITEPRESS - Science and Technology Publications, 2021. doi:10.5220/0010303401850195.

[10] L. Ciampi, C. Santiago, J. P. Costeira, C. Gennaro, G. Amato, Unsupervised vehicle counting via multiple camera domain adaptation, in: Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020, volume 2659 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 82–85.

[11] L. Ciampi, C. Santiago, J. Costeira, F. Falchi, C. Gennaro, G. Amato, Unsupervised domain adaptation for video violence detection in the wild, in: Accepted at the 3th International Conference on Image Processing and Vision Engineering (IMPROVE), SCITEPRESS - Science and Technology Publications, 2023.

[12] L. Ciampi, P. Foszner, N. Messina, M. Staniszewski, C. Gennaro, F. Falchi, G. Serao, M. Cogiel, D. Golba, A. Szczęsna, G. Amato, Bus violence: An open benchmark for video violence detection on public transport, Sensors 22 (2022) 8345. doi:10.3390/s22218345.

[13] L. Ciampi, F. Carrara, V. Totaro, R. Mazziotti, L. Lupori, C. Santiago, G. Amato, T. Pizzorusso, C. Gennaro, Learning to count biological structures with raters' uncertainty, Medical Image Analysis 80 (2022) 102500. doi:10.1016/j.media.2022.102500.

[14] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Vairo, Car parking occupancy detection using smart camera networks and deep learning, in: 2016 IEEE Symposium on Computers and Communication (ISCC),

IEEE, 2016. doi:10.1109/iscc.2016.7543901.

[15] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, C. Vairo, Deep learning for decentralized parking lot occupancy detection, Expert Systems with Applications 72 (2017) 327–334. doi:10.1016/j.eswa.2016.10.055.

[16] G. Amato, P. Bolettieri, D. Moroni, F. Carrara, L. Ciampi, G. Pieri, C. Gennaro, G. R. Leone, C. Vairo, A wireless smart camera network for parking monitoring, in: 2018 IEEE Globecom Workshops (GC Wkshps), IEEE, 2018. doi:10.1109/glocomw.2018.8644226.

[17] L. Ciampi, G. Amato, F. Falchi, C. Gennaro, F. Rabitti, Counting vehicles with cameras, in: Proceedings of the 26th Italian Symposium on Advanced Database Systems, Castellaneta Marina (Taranto), Italy, June 24-27, 2018, volume 2161 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

[18] G. Amato, L. Ciampi, F. Falchi, C. Gennaro, Counting vehicles with deep learning in onboard UAV imagery, in: 2019 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2019. doi:10.1109/iscc47284.2019.8969620.

[19] L. Ciampi, C. Gennaro, F. Carrara, F. Falchi, C. Vairo, G. Amato, Multi-camera vehicle counting using edge-AI, Expert Systems with Applications 207 (2022) 117929. doi:10.1016/j.eswa.2022.117929.

[20] D. Cafarelli, L. Ciampi, L. Vadicamo, C. Gennaro, A. Berton, M. Paterni, C. Benvenuti, M. Passera, F. Falchi, MOBDrone: A drone video dataset for man OverBoard rescue, in: Image Analysis and Processing – ICIAP 2022, Springer International Publishing, 2022, pp. 633–644. doi:10.1007/978-3-031-06430-2_53.

[21] V. Zeni, G. Benelli, L. Incrocci, A. Canale, L. Ciampi, G. Amato, S. Chessa, Precision agriculture to improve the monitoring and management of tomato insect pests, Agrochimica 65 (2022) 107–112. doi:10.12871/000218572022013.

[22] L. Ciampi, F. Carrara, G. Amato, C. Gennaro, Counting or localizing? evaluating cell counting and detection in microscopy images, in: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SCITEPRESS - Science and Technology Publications, 2022. doi:10.5220/0010923000003124.

[23] M. Avvenuti, M. Bongiovanni, L. Ciampi, F. Falchi, C. Gennaro, N. Messina, A spatio- temporal attentive network for video-based crowd counting, in: 2022 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2022. doi:10.1109/iscc55528.2022.9913019.

[24] G. Amato, P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, C. Vairo, VISIONE at VBS2019, in: MultiMedia Modeling, Springer International Publishing, 2018, pp. 591–596. doi:10.1007/978-3-030-05716-9_51.

[25] G. Amato, P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, C. Vairo, The VISIONE video search system: Exploiting off-the-shelf text search engines for large-scale video retrieval, Journal of Imaging 7 (2021) 76. doi:10.3390/jimaging7050076.

[26] G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, C. Vairo, VISIONE at video browser showdown 2021, in: MultiMedia Modeling, Springer International Publishing, 2021, pp. 473–478. doi:10.1007/978-3-030-67835-7_47.

[27] G. Amato, P. Bolettieri, F. Carrara, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, C. Vairo, VISIONE at video browser showdown 2022, in: MultiMedia Modeling, Springer International Publishing, 2022, pp. 543–548. doi:10.1007/978-3-030-98355-0_52.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[29] H. Fang, P. Xiong, L. Xu, W. Luo, Transferring image-CLIP to video-text retrieval via temporal relations, IEEE Transactions on Multimedia (2022) 1–14. doi:10.1109/tmm.2022.3227416.

[30] N. Messina, M. Stefanini, M. Cornia, L. Baraldi, F. Falchi, G. Amato, R. Cucchiara, ALADIN: Distilling fine-grained alignment scores for efficient image-text matching and retrieval, in: International Conference on Content-based Multimedia Indexing, ACM, 2022. doi:10.1145/3549555.3549576.

[31] G. Amato, F. Carrara, F. Falchi, C. Gennaro, L. Vadicamo, Large-scale instance-level image retrieval, Information Processing & Management 57 (2020) 102100. doi:10.1016/j.ipm.2019.102100.

[32] F. Carrara, L. Vadicamo, C. Gennaro, G. Amato, Approximate nearest neighbor search on standard search engines, in: Similarity Search and Applications, Springer International Publishing, 2022, pp. 214–221. doi:10.1007/978-3-031-17849-8_17.

[33] S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. Þ. Jónsson, J. Lokoč, A. Leibetseder, F. Mejzlík, L. Peška, L. Rossetto, K. Schall, K. Schoeffmann, H. Schuldt, F. Spiess, L.-D. Tran, L. Vadicamo, P. Veselý, S. Vrochidis, J. Wu, Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown, International Journal of Multimedia Information Retrieval 11 (2022) 1–18. doi:10.1007/s13735-021-00225-2.