Spoke 2 use cases

Definition and workflow document

Extended Computer Vision at high rate

Spoke	2
WP	6,5
Use case short name	Extended Computer Vision at high rate
Use case ID	UC2.6.1
Expected Completion	1/8/2025

Approval workflow

Status	Version	Date	Submitter	Note	Signature
Submitted to Spoke Leaders	1.1	27/7/2023	WP Leader	First version	
Comments from Spoke Leaders' team	1.2	04/08/2023	Spoke Leader	First round of comments	
Approved	1.3	11/9/2023	Spoke Leader		

Acronyms

- SP(s): Spoke Leader(s)
- WP(s): Work Package Leader(s)
- UC: Use Case
- KPI: Key Performance Indicator

Use case definition

The last decade has witnessed the uprising of Computer Vision technologies, from real time object detection to semantic segmentation, passing through image restoration, visual tracking, etc. (see, e.g. [0], and references therein). Such rising has deeply affected also the research realm; for example, in the High Energy sector, Deep Learning and Computer Vision techniques have become ubiquitous, as easily readable from the living review of [1], reporting the most relevant reports, lectures, and articles about such applications in the field of High Energy Physics.

For such applications, a huge amount of work has been done to reduce the inference time of the trained deep neural networks, and distributed computing techniques have become relevant when dealing with such applications.

Furthermore, the intersection between fundamental research, physics technologies and deep neural networks experience an additional, relevant intersection: extended spectral computer vision.
Extended spectral computer vision, where images are acquired beyond the human-visible spectrum, comprises a set of ubiquitous techniques in the applied physics domain, ranging its applications from the fields of (extra)galactic observation to nuclear-based medical analysis, passing through earth remote sensing, environmental monitoring, biomedicine, geology, earth sciences, cultural heritage, archeology and materials analysis (see, as a limited example, [2-6]). (We are referring here to either all the spectral-based technologies, in all wavelength range, such as - but not limited to - Multi- and Hyper-spectral imaging, raman

imaging, ion beam analysis, double- and multi-energy CT scans, muography, scanning electron microscopy; but also on Near-Infrared RGB imaging). These techniques may all produce *datacubes,* i.e. images whose point, instead of being a triple R,G,B, are one dimensional vectors (spectrograms).

It is becoming also relevant performing statistical based analysis (based either on unsupervised as well as supervised and deep learning methods) on such datacubes (see, e.g. [7-13]), also in real time, e.g. for industrial production quality control, real time telescopy adaptation, environmental risk alerting, and similar.

One of the active current research topics in computer vision is the image segmentation task, performed to automatically identify relevant regions in images. From the spectral perspective, this task is vastly complicated by the high dimensionality of spectral datacubes. On the other hand, datacubes contain an enormous amount of spectral information which can be exploited to perform any computer vision approach.

The goal is thus to apply statistical learning approach for:

- Perform a deep clustering on data cubes to infer an image segmentation out of a pixel-based spectral analysis; this is also helpful in denoising and data compression of huge data cubes.
- Apply standard computer vision supervised method to perform inference tasks on datacubes of applied physics interest.
- Test those models on the data lake in an ad hoc environment for high rate processing.

The flagship use case proposed in this document focuses on two aspects related to applied research in domains using computer vision, coupled with the need to perform fast ("high rate") analyses on large real / realistic datasets.

From the technological point of view, one of the purposes of the presented use case, is to build a demonstrator to validate the high rate analysis infrastructure provided by WP5 has the capability to support a wide range of different scientific domains.

The proposed infrastructure aims to support the paradigm shift from a batch-based approach to an interactive analysis model based on a parallel processing possibly over geographically distributed backends. While it is born to support HEP data analysis frameworks it is supposed to be generic enough and from WP5 perspectives the goal is to provide scientific domain free setup. The aforementioned high rate infrastructure will offer a user-friendly interface based on Jupyter (multi tenancy JupyterHub) and will adopt open-source industry standards such as *Dask* in order to overcome the memory limitations during the data manipulation and to ensure a high level of parallelization. The technical architecture will provide handles to transparently distribute workloads (i.e. unit of work made of clusters of Dask master plus its workers) in order to access huge amounts of computing capacity. In order to achieve this objective the plan is to test the offloading mechanism which allows to scale out the local system by sending users payloads to be executed on suitable resources.

The system will also include the integration with Data Management services for data movement between providers of the ICSC Data Lake, as well as to keep track of things like data locality. WP5 will provide services to be integrated with the storage endpoint in order to demonstrate, also in this case, the genericity of the adopted solutions and technologies.

It will initially focus on the technical part, preparing and validating the tools, while at the same time identifying a (some) proper dataset(s) for the second Proof-Of-Concept part, where the tools will be validated on a realistic scenario.

References:

[0] Chai, J., Zeng, H., Li, A., Ngai, E.W.T., (2021), Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Machine Learning with Applications, Volume 6, 2021, 100134, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2021.100134.

[1] Feickert, M., & Nachman, B.P. (2021). A Living Review of Machine Learning for Particle Physics. *ArXiv, abs/2102.02770*. https://doi.org/10.48550/arXiv.2102.02770

[2] Risaliti, G., Elvis, M. (2004). A Panchromatic View of AGN. In: Barger, A.J. (eds) Supermassive Black Holes in the Distant Universe. Astrophysics and Space Science Library, vol 308. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-2471-9_6

[3] Fornaro J, Leschka S, Hibbeln D, Butler A, Anderson N, Pache G, Scheffel H, Wildermuth S, Alkadhi H, Stolzmann P. (2011) Dual- and multi-energy CT: approach to functional imaging. Insights Imaging. 2011 Apr;2(2):149-159. doi: 10.1007/s13244-010-0057-0. Epub 2011 Jan 19. PMID: 22347944; PMCID: PMC3259372.

[4] Im, J. and Jensen, J.R. (2008), Hyperspectral Remote Sensing of Vegetation. Geography Compass, 2: 1943-1961. https://doi.org/10.1111/j.1749-8198.2008.00182.x

[5] Qian, S.-E. (2022). Overview of Hyperspectral Imaging Remote Sensing from Satellites. In Advances in Hyperspectral Image Processing Techniques, C.-I. Chang (Ed.). https://doi.org/10.1002/9781119687788.ch2

[6] Mandò, P.A. and Przybyłowicz, W.J. (2016). Particle-Induced X-Ray Emission (PIXE). In Encyclopedia of Analytical Chemistry, R.A. Meyers (Ed.). https://doi.org/10.1002/9780470027318.a6210.pub3

[7] Christopher J. Shallue, & Andrew Vanderburg (2018). Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90*. The Astronomical Journal, 155(2), 94.*

[8] Kevin Schawinski and others, Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit, *Monthly Notices of the Royal Astronomical Society: Letters*, Volume 467, Issue 1, May 2017, Pages L110–L114, https://doi.org/10.1093/mnrasl/slx008

[9] Khan, A., Munir, M.T., Yu, W. *et al.* A Review Towards Hyperspectral Imaging for Real-Time Quality Control of Food Products with an Illustrative Case Study of Milk Powder Production. *Food Bioprocess Technol* **13**, 739–752 (2020). https://doi.org/10.1007/s11947-020-02433-w

[10] Wang, Q., Zhang, X., Tang, B. *et al.* Lithology identification technology using BP neural network based on XRF. *Acta Geophys.* **69**, 2231–2240 (2021). https://doi.org/10.1007/s11600-021-00665-8

[11] Kleynhans, T., Schmidt Patterson, C.M., Dooley, K.A. *et al.* An alternative approach to mapping pigments in paintings with hyperspectral reflectance image cubes using artificial intelligence. *Herit Sci* **8**, 84 (2020). https://doi.org/10.1186/s40494-020-00427-7

[12] Bombini, A., Anderlini, L., dell'Agnello, L., Giacomini, F., Ruberto, C., Taccetti, F. (2022). The AIRES-CH Project: Artificial Intelligence for Digital REStoration of Cultural Heritages Using Nuclear Imaging and Multidimensional Adversarial Neural Networks. In: Sclaroff, S., Distante, C., Leo, M., Farinella, G.M., Tombari, F. (eds) Image Analysis and Processing – ICIAP 2022. ICIAP 2022. Lecture Notes in Computer Science, vol 13231. Springer, Cham. https://doi.org/10.1007/978-3-031-06427-2_57

[13] Bombini, A., Anderlini, L., dell'Agnello, L., Giacomini, F., Ruberto, C., Taccetti, F. (2022). Hyperparameter Optimisation of Artificial Intelligence for Digital REStoration of Cultural Heritages (AIRES-CH) Models. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds) Computational Science and Its Applications – ICCSA 2022 Workshops. ICCSA 2022. Lecture Notes in Computer Science, vol 13377. Springer, Cham. https://doi.org/10.1007/978-3-031-10536-4_7

Participating Institutions

- Leader: Alessandro Bombini (INFN-FI)
- Participants: INFN-PG and INFN-NA for the deployment on distributed computing infrastructures; additional institutions depending on the use case selection phase.

Industrial involvement

The analysis of large datasets of (extended spectra) images is of interest in many productive and industrial domains, including medical analysis, space economy and quality assurance. It is expected that the activities in this flagship project will intersect with those in other initiatives related to Spoke2, in conjunction with industrial realities:

- The Innovation Grant Agri@Intesa, lead by Intesa Sanpaolo, aiming at a financial analysis of agricultural enterprises from satellite images;
- The Innovation Grant Hammon, lead by SOGEI, aiming at a risk assessment (financial and civil) of urban areas from satellite / aerial images;
- The FlagShip "AI algorithms for (satellite) imaging reconstruction" of Spoke2/WP6, focusing on analysis techniques on aerial / drone / satellite images.

Use Case Expected Activities

The activities are planned over a period of 24 months, with a broad-brush division into a preparation and an execution phase.
The first part includes preparation of the technical tools and the selection of at least one dataset to be processed. The second focuses on the Proof-Of-Concept realization, and the wrapping-up of results

1. <u>ICSC 12-24 m (Sept-2023 - Aug 2024, synced with ICSC MS9)</u>: Survey of the State-of-the-Art; tracking of R&D technologies to be used; selection of datasets for use cases (at least one).
   - ➢ <u>D1</u>: report on technologies to be used, selection of at least one test dataset.
   - ➢ Intermediate report MS8
2. <u>ICSC 25-36 m (Sept 2024 - Aug 2025, synced with ICSC MS10)</u>: Implementation of the selected technology(ies); test and validation on selected dataset(s). Proof-of-Concept deployment.
   - ➢ <u>D2</u>: Report on the work carried out; release of the developed code on public repository.

KPIs

KPI ID	Description	Acceptance threshold
KPI2.6.1.1	Publications submitted (including conference proceedings)	1
KPI2.6.1.2	Presentation at conferences	1
KPI2.6.1.3	Public Code repositories	1
KPI2.6.1.4	Use case Test Dataset	1

Risk Analysis

Identifier	Description	Mitigation
R1	The CN is unable to provide the needed resources	The program of work and the intermediate and final tests are scaled to what is available, for example from INFNCloud
R2	Impossible to select a proper dataset for the final test	The Spoke2 will have available multiple adequate datasets from other initiatives (for example, the Hammon Innovation Grant). On top of that, INFN is a stakeholder in projects handling heritage

		datasets, which can be used as a test case.
R3	Impossible to setup a distributed environment to test the high rate part, due to technical features of the ICSC Datalake	The setup as described is already operational (on a smaller scale) on non-ICSC resources. A difference preventing the system to work on ICSC is not probable. In any case, the Proof of Concept can be eventually, if needed, executed on other Cloud-like resources.

Resource Needs

Computing resources

Projected Needs

- R&D Phase (ICSC 12-24 m):
    - one R&D node to test algorithms and external tools. Optimal: 1 node with at least 16 cores, 256 GB RAM, 1-2 NVidia A100 or equivalent, 5 TB of attached disk
    - One small cloud-instantiated cluster to test the distributed part. Optimal: 10 VM nodes with O(8 cores) + a fraction of GPU (MIG is ok) for 10% of the time.
    - Access to the same disk area in both systems.
- PoC Phase (ICSC 13-24 m):
    - Scale out to 100 VM with O(8 cores) + a fraction of GPU (MIG is ok), for 10% of the time.
    - 10 TB disk for final tests

Resources already secured

- A single VM node equipped with GPU available under ML_INFN with 1 TB of attached disk;

- For short time ranges, O(10) VM from INFN_Cloud.

External Resources Needed

At the moment we do not expect the need for resources external to what ICSC should be able to provide.

Expected synergies with other WPs

The activities will be carried on with the synergy and the support of Spoke2/WP5, in charge of the deployment, operations and reporting of the distributed high rate processing system. The system to be deployed is similar to the one described in the Flagship of Spoke2/WP2 “Quasi interactive analysis of big data with high throughput”.

Periodic reports during implementation

Intermediate report End of October 2023

During the period September - October 2023 we:

- Started the overview of the technologies needed to carry out the distributed computing inference
- Started the overview of the possible application of the technology at hand; in particular
  - Defined a pipeline to build a synthetic dataset of X-ray imaging on pictorial artworks
  - Defined a pipeline to build a synthetic dataset of spectral imaging of sky regions
  - Found and exploring the adaptability of open dataset of Astrophysical imaging
- Developed the code for a preliminary analysis with Deep learning technologies.
- Seek collaboration with soon-to-be affiliated researchers from partner institutions to validate the method.

MS7 final report

The Flagship UC2.6.1 does not have an explicit Milestone or target for MS7; an intermediate report describing the activities so far is included.

During the period November 2023 - February 2024 we:

- Selected a set of technologies to be used for the distributed computing inference
- Conducted a set of applications of aforementioned technologies:
  - Defined and implemented a pipeline to build a synthetic dataset of X-ray imaging on pictorial artworks
  - Defined and implemented a pipeline to build a synthetic dataset of spectral imaging of sky regions

  This effort culminated in the Open Access publication of the two dataset [1].
- Designed, Developed, trained and benchmarked a set of Deep learning models for Deep Clustering, which culminated in the Open Source publication of two trained models [2].
- The publication of the first version of the developed code for the points above [3] in the ICSC Spoke 2 GitHub repository.
- The work done has been reported in a pre-print published on ArXiv [4].

[1] Alessandro Bombini, Fernando García-Avello Bofías, Caterina Bracci, Michele Ginolfi, & Chiara Ruberto. (2024). Synthetic Datasets for ICSC Flagship 2.6.1. "Extended Computer Vision at high rate" paper #1 "Datacube segmentation via Deep Spectral Clustering" (Version 20240105) [Data set]. http://doi.org/10.15161/oar.it/143545
[2] Alessandro Bombini, Fernando García-Avello Bofías, Caterina Bracci, Michele Ginolfi, & Chiara Ruberto. (2024, January 5). Trained Models for ICSC Flagship 2.6.1. "Extended Computer Vision at high rate" paper #1 "Datacube segmentation via Deep Spectral Clustering" (Version 20240105). http://doi.org/10.15161/oar.it/143543
[3] Bombini, A., Bofías, F. G., Bracci, C., Ginolfi, M., & Ruberto, C. (2024). FastExtendedVision-DeepCluster. GitHub repository (Version 2024-01-05) [Computer software]. https://doi.org/10.15161/oar.it/143544
[4] Bombini, A., Bof'ias, F.G., Bracci, C., Ginolfi, M., & Ruberto, C. (2024). Datacube segmentation via Deep Spectral Clustering. ArXiv, abs/2401.17695.

MS8 final report

The Flagship UC2.6.1 does not have an explicit Milestone or target for MS7; an intermediate report describing the activities so far is included.

In addition to the work reported in the previous reports,dDuring the period February 2024 - June 2024 we:

- Conducted a set of applications of technologies to be used for the distributed computing inference;
    - Defined a virtual recoloring pipeline, starting from the synthetic dataset pipeline generation;
        - Created many synthetic datasets of MA-XRF images, each of which is O(300k) images; the difference among the various data set is the palette (i.e., the seed couples rgb - tabulated XRF signal)
        - Used the dataset to train a Variational Deep Embedding model to reduce the on-disk size of the dataset(s)
        - Trained a Vision Transformer to map (embedded) MA-XRF images to RGB images
    - We are currently validating the aforementioned pipeline
    - We are testing domain adaptation algorithms to enforce extendibility of the pipeline to real use cases scenario
    - We are conducting an XRF measurement campaign to have a real dataset.
- Conducted a set of tests for distributed computing inference of models.