NAIS: Artificial Intelligence Powered On-Board Sensor Processing for UAVs

Alessandro Palmas – R&D Manager Nurjana Technologies
Via M. Betti 27/29
09067, Elmas (CA)
ITALY

alessandro.palmas@nurjanatech.com

Pietro Andronico– CEO Nurjana Technologies
Via M. Betti 27/29
09067, Elmas (CA)
ITALY

pietro.andronico@nurjanatech.com

ABSTRACT

This paper discusses results of the first year of activities carried out by Nurjana Technologies under the Italian National Plan for Military Research. It describes how advanced Image Processing techniques are applied to enable on-board sensor processing for UAVs. Four use-cases are considered: target detection, identification and localization, semantic segmentation for road/off-road and human body parts, and human action recognition. All algorithms have been developed using state of the art Computer Vision methods based on Deep Neural Networks. Acquisition campaigns have been carried out to collect custom datasets reflecting typical operational scenarios, where the peculiar point of view of a multi-rotor UAV is replicated. Algorithms architectures and trained models performances are reported, demonstrating a high level of both accuracy and inference speed, paving the way to enabling on-board autonomous functions.

1 INTRODUCTION AND BACKGROUND

Recent trends in the UAVs domain see the extensive exploration of drone swarms solutions, especially leveraging the growing availability of small systems, with both rotating and fixed wings architectures, at low costs. These new advancements allow the design and execution of new types of missions, being able to exploit a different configuration of tactical assets, and the increased stream of data made available by the high number of sensors that can be deployed.

In this new setting, additional bottlenecks arise. In fact, every sensor onboard the flying assets requires, in the majority of cases, a dedicated operator to continuously monitor it. This poses an important limitation to the scalability of the system, depending on the available qualified manpower.

In the last decade, another field has seen a relevant growth in terms of technological capabilities, opening major opportunities. The Computer Vision domain underwent a revolution thanks to the breakthroughs achieved in the Deep Learning field. Performances of image processing algorithms overpassed every established state-of-the-art reference, unlocking applications that were not possible before.

This paper presents results achieved during a one-year study carried out under the PNRM program (National Program for Military Research). The PNRM is the ensemble of technological innovation programs aiming at the growth and maturation of technologies for military applications at national level as well as in terms of international cooperation. National industries, small and medium-sized enterprises, research centers and

institutions and universities can participate in the PNRM.

Phase 1 (of 2) of the project, named NAIS - Artificial Intelligence for Drones, had the goal of developing a software library featuring image processing algorithms that can be equipped on-board, in embedded devices and run in real-time to foster vehicle autonomy capabilities. In particular, these algorithms are focused on four verticals: Object Detection, Classification and Localization, Semantic Segmentation for Road/Off-Road and Human Body Segmentation, and Human Action Recognition. For each of them, a specific use case scenario has been identified and the work has been carried out as follows.

The first step has been a thorough literature review of Deep Learning methods for image processing, in order to identify the best performing solution for each of the four domains. The models have been selected focusing on both accuracy and inference speed, given the need of providing real-time inference.

At the same time, with the goal of demonstrating the applicability of these solutions on-board flying vehicles, custom datasets have been collected using multirotor drones. Acquired images have been manually annotated to be used for training and evaluation.

Selected model architectures have been analyzed and trained on custom datasets applying all relevant best practices for Deep Learning training, achieving target performances in terms of both accuracy and frame per second during inference. Chosen technologies can be applied not only on RGB images as demonstrated, but also on grayscale, InfraRed, MultiSpectral and HyperSpectral data. In addition, the system can easily be customized to be trained on custom targets. Each model has been trained to solve a specific use case as described below.

2 MODEL ARCHITECTURE AND IMPLEMENTATION

2.1 Object Detection, Classification and Localization

The algorithm for Object Detection, Classification and Localization has been trained using a single class of people in RGB images acquired from a drone flying 10 meters above the ground with the camera oriented downwards (-90°). This capability can be used, for example, to detect a target and autonomously track it in scenarios like Search & Rescue missions or Restricted Area Monitoring.

For this model the Yolo base Architecture has been selected. It uses a feature-map on multiple scales for bounding box predictions and the DarkNet-53 inspired by the ResNet. The base architecture has been modified according to specific needs related to the problem to be addressed, inference speed required and deployment hardware.

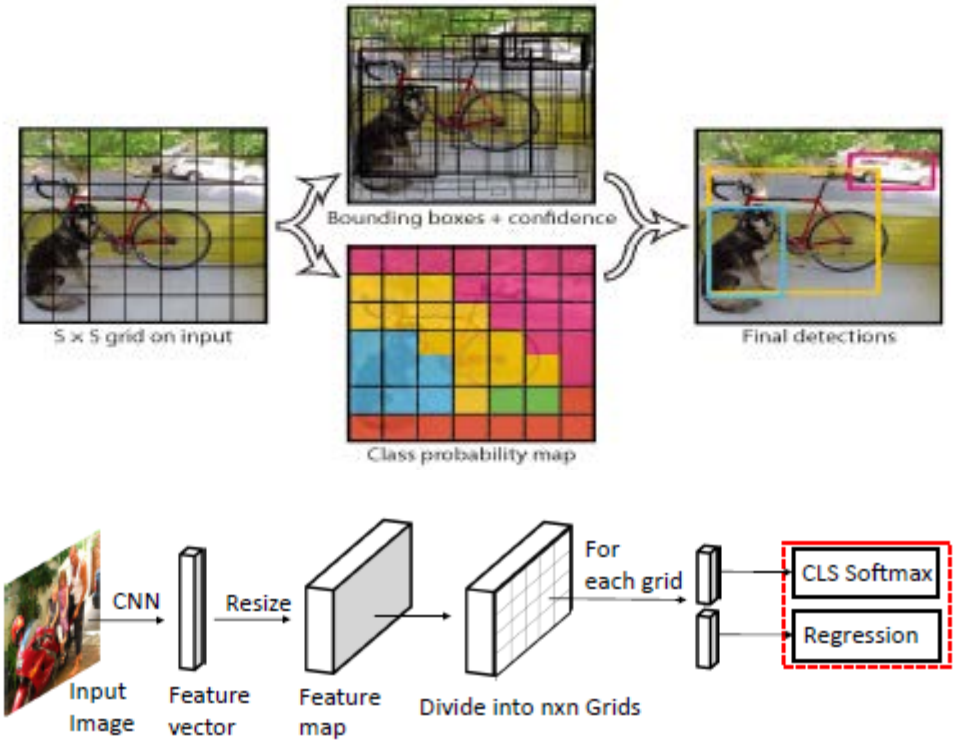

Figure 1: Yolo base architecture schemes

The image below shows an example of the model's output.

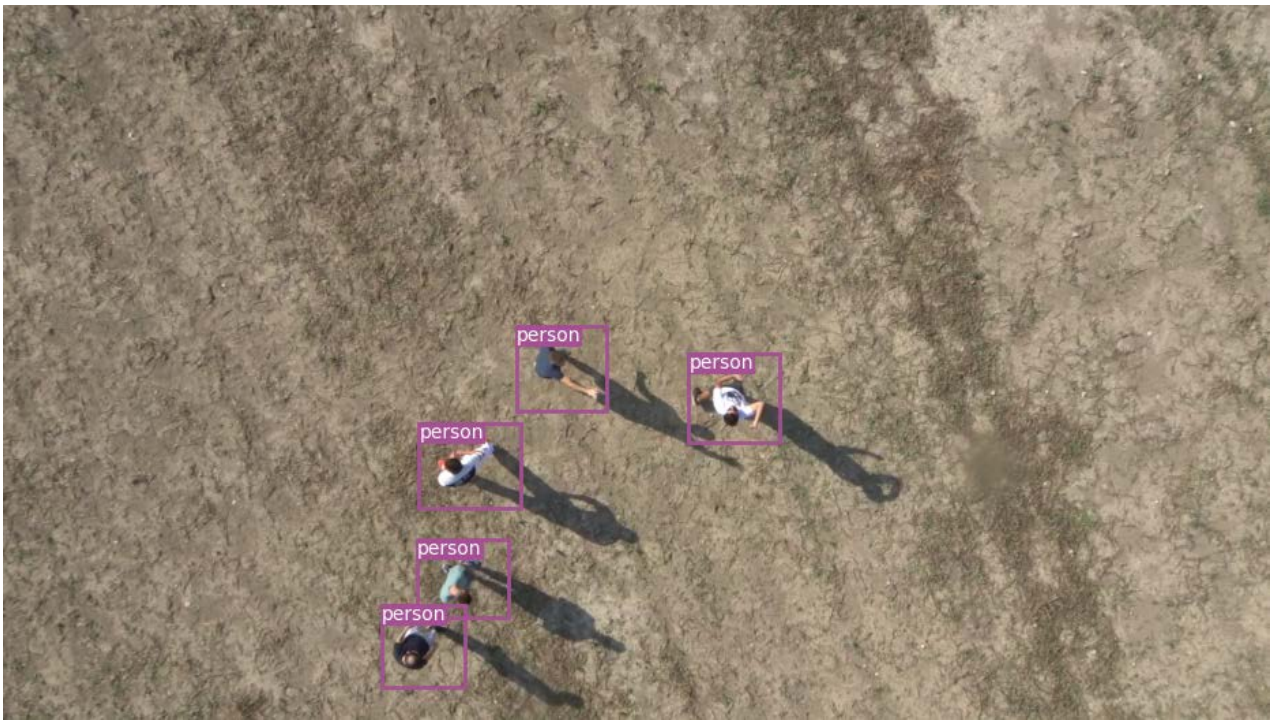

Figure 2: Example of the Object Detection, Classification and Localization model output

2.2 Semantic Segmentation

The model for Semantic Segmentation has been applied in two different contexts. The first aims at distinguishing between road and off-road parts in RGB images acquired from a drone flying 10 meters above the ground with the camera oriented downwards (-90°). This functionality can be used to implement autonomous navigation, particularly useful in GNSS-denied zones, to build for example a system able to autonomously follow a road for patrolling operations.

The second application of the Semantic Segmentation aimed at segmenting the human silhouette in 19 different body parts (classes) in RGB frames acquired from a drone flying at 5 meters above the ground with the camera oriented at -40°. Some of the human body parts classes that have been used are: head, hair, arms, legs, torso, feet, etc.

Among the different architectures explored, the one that has been selected as a base architecture is the DDRNet, representing a good tradeoff between inference speed and accuracy.

The network has two branches, which apply different processing to the input frames. The first branch generates feature maps at higher resolution while the other extracts context information using down sampling. The two branches are also connected to assure an efficient information concatenation. Also in this case, the base architecture has been modified according to specific needs related to the problem to be addressed, inference speed required and deployment hardware.

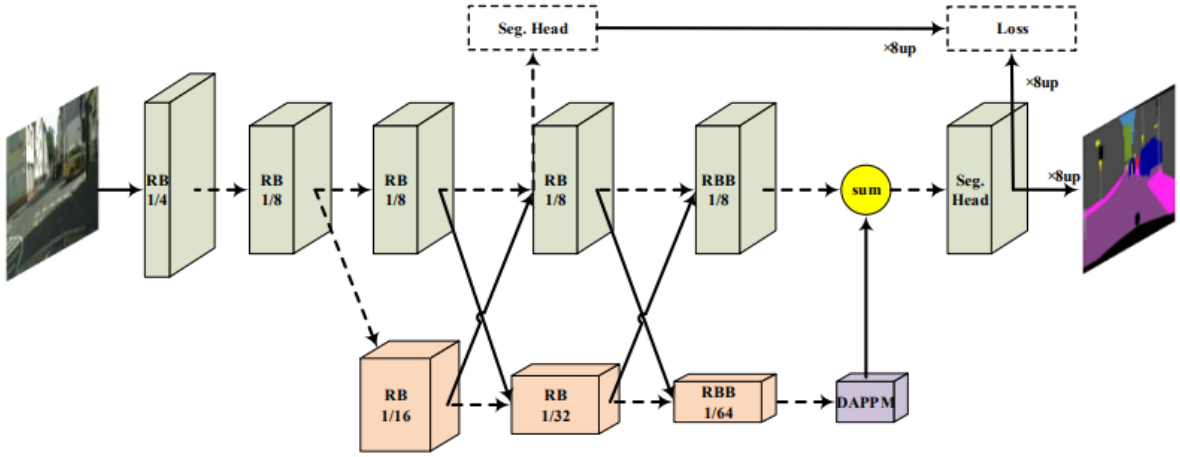

Figure 3: DDRNet base architecture scheme

The images below show two examples of the model's output.

Figure 4: Examples of Semantic Segmentation model output, for Road Segmentation (left) and Human Body Segmentation (right)

2.3 Human Action Recognition

The model for Human Action Recognition has been applied to classify six different actions in RGB frames sequences acquired from a drone flying 10 meters above the ground with the camera oriented downwards (-90°). Actions performed were: Standing Idle, Walk, Run, Crouch, Aim and Throw. This feature can be used, for example, in crowded areas monitoring applications to automatically identify threatening behavior.

Among the different architectures explored, the one that has been selected as a base architecture is the Two-Stream ConvNet. This architecture exploits spatial and temporal information, running two different branches in parallel and exchanging information between the two at the level of convolutional layers. As seen before for the other models, the base architecture has been modified according to specific needs related to the problem to be addressed, inference speed required and deployment hardware.

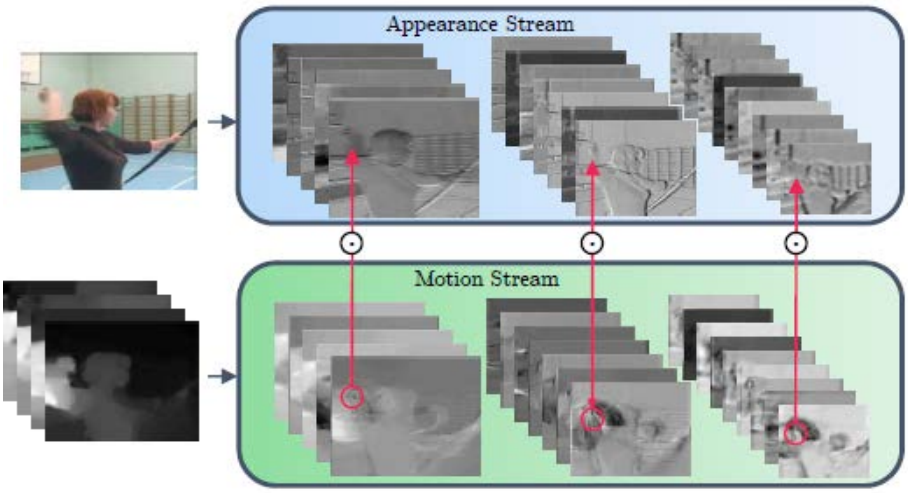

Figure 5: Two-Stream ConvNet base architecture

The image below shows an example of the model's output.

Figure 6: Example of Human Action Recognition model output

The software library has been packaged and delivered to the Italian Armed Forces with a detailed user manual and the dataset to make sure results can be easily reproduced.

3 MODELS PERFORMANCES AND DEPLOYMENT DETAILS

The following table summarizes performances of implemented algorithms. For each algorithm it shows two metrics, one for the inference speed and one for the accuracy. The former is measured for every algorithm in terms of frames per second, while the latter depends on the specific application and varies to present the most meaningful metric typically used with it. For each performance measure, both the requirement and the result obtained are shown, confirming that all targets have been met.

Table 1: Algorithms performances

Algorithm	Performance Measure	Requirement	Result
Automatic target detection, identification and localization	Frames Per Second (FPS)	15-20 FPS	**50 FPS**
	Mean Average Precision (mAP)	55-60 %	**75.5 %**
Context Semantic Segmentation	Frames Per Second (FPS)	10 FPS	**50 FPS**
	Intersection over Union	50 %	**98.86 %**
Human Body Segmentation	Frames Per Second (FPS)	25 FPS	**35 FPS**
	Mean Average Precision (mAP)	50 %	**52.84 %**
Entity Action Recognition	Frames Per Second (FPS)	10 FPS	**15 FPS**
	Mean Average Precision (mAP)	50 %	**55.8 %**

All the algorithms have been implemented using state of the art Deep Learning Frameworks (TensorFlow / Pytorch). For this first phase of the PNRM contract, they have been deployed on a desktop computer powered by a NVIDIA RTX 3090 GPU. They have been developed to be easily portable to embedded, GPU-powered, devices, such as boards of the NVIDIA Jetson family (Nano, TX, Xavier). This hardware is particularly well suited to be equipped onboard, due to its low weight and power consumption, as well as very flexible hardware and software interfaces.

4 MODELS APPLICABILITY

All selected models have been chosen to assure a broad applicability, working on both input and output ends, favoring the easiest possible integration with third party elements. The adopted technology and the implementation choices have been driven by the goal of creating a software library able to handle different types of input sources and to provide outputs that contain all the information extracted from the frames.

4.1 Inputs

Each model, here applied on RGB images coming from a visible camera, can work on the following type of inputs, and easily be extended to additional similar ones

- RGB (3-channels)/Grayscale (single-channel) images coming from visible camera
- Grayscale (single-channel) images coming from InfraRed cameras
- Multi-Spectral and Hyper-Spectral images coming from ground, air or space vehicles

4.2 Outputs

The following list describes outputs provided by each model

- Automatic target detection, identification and localization
    - Array of detections having the whole list of targets objects identified in every frame;
    - Bounding box for each detection, defining target position inside the frame;
    - Classification of each detection, representing the class in which the identified and localized target belongs;
    - Classification confidence for each detection, measuring the probability associated with the label assigned to the target identified in the frame
- Context Semantic Segmentation
    - Pixel map with classification, where each frame pixel is associated with a classification label. For example, it allows to identify pixels belonging to the “Road” class and those belonging to the “Off-Road” one.
- Entity Action Recognition
    - Action classification of each frame sequence, representing the class in which the target action belongs;
    - Classification confidence for each action detection, measuring the probability associated with the label assigned to the target action identified in the frame
- Human Body Segmentation
    - Pixel map with classification, where each frame pixel is associated with a classification label. For example, it allows to identify pixels belonging to the “Head” class and those belonging to the “Torso” one.

5 CONCLUSIONS

Results discussed demonstrate the ability of the system to provide the required functionalities with satisfying results in terms of both accuracy and inference time. Leveraging Deep Learning methods for Computer Vision proved to be a very powerful way to pursue scalability in UAVs applications, building functionalities that can be directly applied in reaching fully autonomous systems.