uc3m | Universidad Carlos III de Madrid

Master in Telecommunication Engineering

2018-2020

Master Thesis

"Embedded solution for person identification and tracking with a robot"

Ignacio Condés Menchén

Fernando Díaz de María

Eduardo Perdices García

Leganés, 2020

SUMMARY

This project describes the development process of an embedded system capable of performing a reactive following of a person. It makes use of convolutional neural networks and probabilistic tracking for processing the perception acquired by a RGBD camera. This input is processed in a NVIDIA Jetson TX2, an embedded SoM (*System-on-Module*). This device is capable of performing computationally demanding tasks onboard, coping with the complexity required to run a robust tracking and following algorithm. The full design is implemented on a robotic mobile base, which receives velocity commands from the board, intended to move towards the desired person.

Keywords: deep learning, robotics, person following, rgbd

DEDICATION

La escritura de estas palabras cierra este capítulo en mi vida académica. Muchos meses de continua lucha me han traído, madrugada de estudio tras madrugada de estudio, hasta el final del camino, en el que puedo dar por terminada la consecución de este máster. Tengo que agradecer a muchas personas por todo el apoyo, material y emocional, que me han brindado durante este tiempo.

Para empezar, a mis queridos compañeros, colegas y amigos del máster. Cuando todos llegamos aquí desde los rincones más diversos de España, la dificultad, y la calidad y calidez del grupo fomentaron que enseguida fuéramos una piña inquebrantable. Y salimos adelante pese a enfrentarnos a más momentos duros de lo que al principio estimábamos. Con humor, con fuerza y sobre todo con ese saber hacer que finalmente nos ha hecho superar esta etapa tan exigente pero a la vez tan satisfactoria: lo hemos conseguido.

Además, les debo un profundo agradecimiento a mis tutores, Eduardo Perdices y Fernando Díaz, y sobre todo a José María Cañas, por cuánto se han volcado en este proyecto conmigo, en especial en esta última etapa de escritura para ayudarme a mantener el rigor y la calidad necesarios en este proyecto. Gracias por vuestra paciencia e involucración.

También, por supuesto, a mis compañeros y amigos de AFC Ingenieros, que me han aguantado, ayudado y animado con este proyecto cada vez que me veían entrar al laboratorio con mi famoso robot y un puñado de cables para soldar una y otra vez, "que esta era la buena". Por otro lado, gracias a AFC Ingenieros como tal, por la enorme flexibilidad que me ha permitido compatibilizar mi empleo con el seguimiento regular del máster. Sin esa confianza y empatía, me habría sido imposible completarlo.

Asimismo, quiero destacar los ánimos y el cariño que me ha infundido la gente que me soporta cada día. Quiero agradecérselo a mis mejores amigos, Jose y Jorge, a mi banda (qué mal tocamos ¡pero qué bien lo pasamos!), a toda mi familia y en especial a mi abuela, que ha animado durante tantos meses a un nieto pegado a un ordenador, siempre con demasiadas cosas pendientes por terminar.

Finalmente, y al igual que en entregas anteriores, sigo sin tener palabras para condensar el infinito cariño, comprensión, desahogo y empatía que mi pareja ha tenido conmigo. Almudena, aunque sé que te mueres de vergüenza por aparecer aquí, quiero dejar constancia de la pieza fundamental que eres en el rompecabezas de mi vida, de que este proyecto ha salido adelante respaldado por el gran aguante y entereza que has tenido las innumerables veces que nuestra relación se ha visto reducida a un eterno "hoy no puedo, tengo que avanzar con el TFM". Un millón de gracias a ti y a tu familia por haber sido mi fuente de fuerzas tantas veces sin ninguna reserva.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This chapter presents the motivation that led to the development of the proposed work. Later, the general objectives of the developed system are outlined, followed by a summary of the structure of this document.

1.1. Motivation

Last decades, the production prices of digital cameras and high-resolution sensors have been greatly reduced, bringing these devices into the consumer market segment: nowadays, everybody carries at least 2 cameras in their mobile phone, aside of high-quality web cameras, or even driving-assistance cameras in cars. This, beside an increase in the hardware performance, has resulted in a strong drive for the computer vision research (Figure 1.1): there are many possibilities out of industrial environments for applications using cameras, such as fancy image modifications, or autonomous driving, as it can be seen on Figure 1.2.

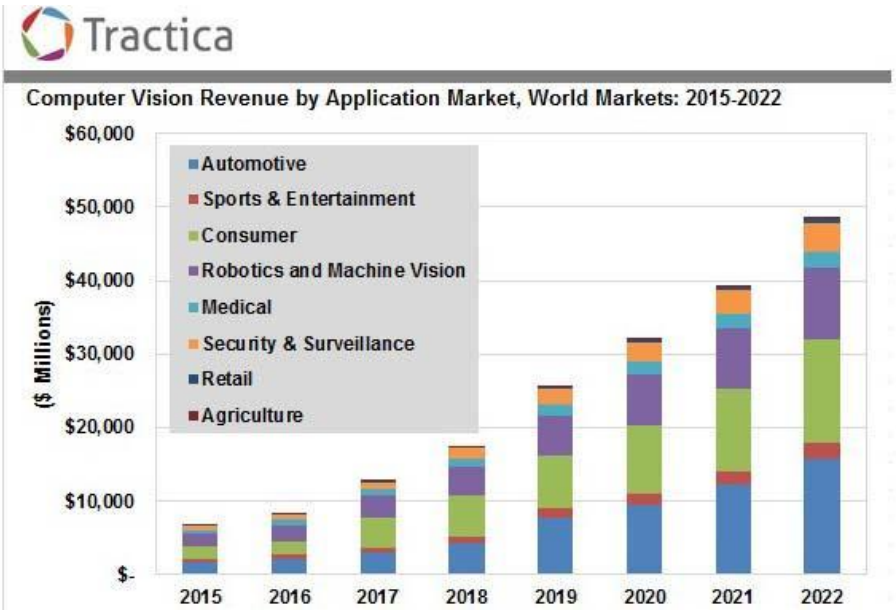


Fig. 1.1. Computer Vision revenues in the last years, and forecast for 2022 (source: [1]).

Specially, the latest times have been notoriously active in this field because of the massive use of *deep learning* for addressing high complexity tasks, such as language understanding [2], speech recognition [3] and computer vision problems, which are linked to the growing interest shown in Figure 1.1. This massive use began in the ImageNet classification contest, where a deep neural network system, AlexNet [4], achieved an overwhelming victory over other approaches [5]. This discovery, along with the significant advances in computing power and parallel computing, has stimulated the

(a) Modifications of a subject on a portrait, such as apparent gender, or age.

(b) Autonomous driving on a Tesla Model X.

Fig. 1.2. Examples of contemporary computer-vision applications.

usage of these technologies, which show an outstanding performance with the available means nowadays [5].

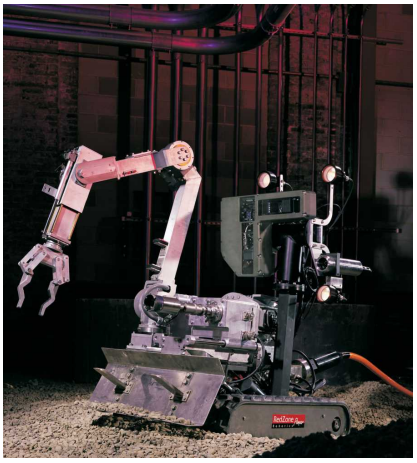
Moreover, robotics applications can be really useful at daily tasks. These tasks are of greater interest when the behavior of a robot tends to emulate the human one, or even pets[1], with the advantage of no people exposed to a significant risk, or, in a less gloomy scenario, without human body physical limitations. This requires a polished (and somehow complex) behavior, which is triggered by a certain input. At this point, two main branches emerge in robotics:

Teleoperated robots: this kind of robots are capable of performing certain actions, which are *remotely controlled by a human operator*. This type is the mostly used one on hazardousness (Figure 1.3a) [6] or high-precision environments [7]. Some advances are made nowadays improving the teleoperation function, implementing *feedback* from the robot, such as haptic feedback [8], or VR (*Virtual Reality*) sensation, to allow that person to sense the environment as if they were in the robot position.

Autonomous robots: these robots are much more complex machines, as they are distinguished for implementing a response by themselves, independently of any kind of remote operator. This is seeked on certain scenarios, where there are some factors (as the time elapsed performing an action, or the cost of a control link with the robot) with a considerable weight in the design [9]. This is the kind of robots that concern us on this work: the state-of-the-art techniques try to emulate *human behavior* (Figure 1.3b), so some actions can begin to be performed autonomously with a certain intelligence, as it will be described below.

The important advances on the last decades on the image processing and audio recognition fields have fostered the development of personal assistants, apart from critical machines as the previously described examples.

[1] https://www.engadget.com/2018/01/08/new-sony-aibo-first-impressions/



(a) Pioneer robot, designed to perform hazardous teleoperated explorations in a deadly radioactive environment.

(b) Pepper, an autonomous humanoid capable of performing on-board processing and reacting to external stimuli intelligently.

Fig. 1.3. Example of a teleoperated (a) and an autonomous (b) robot.

There are outstanding synergies between robotics and computer vision, as it is explored on the system proposed in this work: these fields are combined for obtaining a robust robot capable of following a certain person, navigating towards them on a reactive behavior, and using deep-learning based visual perception. This behavior is composed of two main components: the *perception block*, in charge of processing the images from an embedded RGBD camera, and the *actuation block*, which moves the robotic base accordingly to the relative position of the person to be followed.

This application can be specially interesting on social robots, which are designed to follow a person at home or in a hospital. According to [10]: "*robots that operate around people in the real world need to move in coherent, easily-understood ways, so that they will not startle or harm the people around them. In particular, for robots that operate in hospitals or in nursing homes*".

The work proposed in this thesis improves the system developed on [11], where a neural-network-based following system was run in a standard laptop, with a camera and a robot plugged. In the following dissertation, this work will be revisited, and the points of interest which have allowed to enhance the previous version will be described.

The main contributions of this dissertation may be summarized as follows:

Embedded solution: the final system is mounted on a battery-powered *mobile base*. This robot features a high-performance GPU embedded on a SoM (*System-on-Module*). In contrast to the previous work, this assembly can operate on its own,

without requiring an external computer to perform the deep learning inferences or running algorithms in parallel. A remote monitoring of the behavior is available as well, but it is not required for the system to work. In addition, specific optimization engines allow the system to run faster with 3 neural networks than previously with only 2 networks, on a low-consumption hardware. This will be described in detail in Chapter 4.

Person identification: the proposed system runs 3 neural networks. These networks perform inferences over the images captured by the RGBD sensor, which is attached to the system as the sensing source of the robot. The inferences are devoted to detect the different persons in the scene, as well as to distinguish them by means of a discriminant feature: their face. Unlike the previous development, all the detection and identification tasks are based on neural networks, achieving greater robustness and reliability as it will be discussed in Chapter 4.

Tracking: the full system includes also a person tracker based on optical flow. This tracker aims to guess the trajectories followed by each person that the robot can see. As opposed to the previous work, this tracker allows to roughly follow the persons while the neural network yields a new update, as this tracker takes considerably less time to predict the person displacement. As a result, the robustness of the entire system is improved, compared to a version governed exclusively by the neural inferences, which are sensitive to visual occlusions as well. Trusting just on these inferences could easily result on an unsteady behavior. However, the introduction of the tracker softens the robot movements ensuring a greater robustness in the observable behavior, as it will be explained on Section 3.2.

1.2. Objectives

The main objective of this work is to design and develop an embedded system which allows a low-cost robot with a camera to follow a certain person on a robust way. The result will be an autonomous robot which will follow a specific person, whose face has to be known beforehand (using a *reference face* image). This objective, in turn, can be split into specific subgoals:

1. Implement a real-time person following behavior using embedded low-power hardware and a low-complexity educational robot.

2. Build the inference pipeline using exclusively CNNs (*convolutional neural networks*), as they offer robustness on detection under harsh lighting conditions, such as the ones observed in Figure 1.4.

Fig. 1.4. Poor lighting situations for a low-positioned camera.

3. Combine a neural visual perception with optical tracking to carry out a robust following of the persons in front of the robot. This will provide the system with extra reliability and robustness against detection losses/occlusions.

These subgoals allow to summarize the starting point for the development of this project: the available materials are an educational robot equipped with a battery, an embedded *SoM* and a RGBD sensor.

1.3. Structure of the document

The structure of this work is organized as follows:

- Chapter 1 presents the motivation of this work, as well as summarizing the objectives to be addressed.
- Chapter 2 discusses the state of the art techniques on person detection and robotic following behaviors, placing the work of this document in a technological frame.
- Chapter 3 describes the hardware and software means for developing this work. Later, a full functional description of the implemented system is given, describing the *Perception* and *Actuation* modules that compose the system. Finally, a description of the software architecture that implements the following behavior and makes the robot to follow the person.
- Chapter 4 describes the experiments conducted on the subsystems and modules of this work. The results of each test are shown as well in order to demonstrate the convenience of the design decisions made over the project development. Finally, a global system experiment is shown, where the following behavior of the robot has been studied.
- Chapter 5 discusses the obtained results. Later, conclusions are drawn from the developed work, revisiting the goals and subgoals presented above, and proposing future lines of work that can improve the robot and address its main drawbacks.

The Annexes provide additional tables and results, as it will be mentioned later.

2. STATE OF THE ART

This chapter delivers a review of the state of the art, to provide a general panorama of the problems and methods that this work addresses.

As it was previously introduced, this work is performed to explore the synergies on robotics and deep-learning-based visual perception. In this section, the current approaches and tools will be described in order to outline a general panorama where this work may be framed.

The problem to be addressed is to *get a robot with a camera to follow a person*. This problem can be split into several steps, where different approaches have been previously proposed. These steps will be covered in the following sections.

2.1. Visual person detection

One of the most common approaches is known as the *Viola-Jones* detector [12]. This algorithm relies on a *rigid body model*, which fits a specific shape. On a grayscale image, this shape can be typically distinguished by means of the pixel intensity levels. Although this method was originally designed to detect faces, the rigid body model allows to generalize its usage for detecting different objects, such as persons. With this purpose, several spatial filters called *Haar features* (Figure 2.1) are introduced: these are used across the image looking for the intensity pattern of each template, which should resemble a part of the rigid body. Since this detector provides a weak decision by itself, several filters (previously chosen in a training process) are combined on a *boosted cascade* (Figure 2.2). A person is detected if the weighted combination of several filters are triggered inside a certain area, which is decided to potentially contain a person [13].

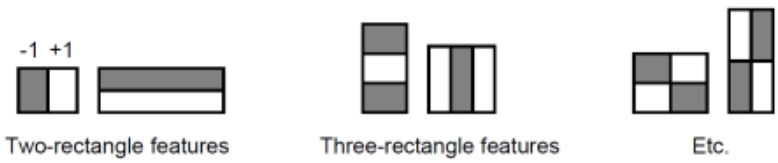


Fig. 2.1. Haar features: some examples [13].

The open-source standard image processing library, OpenCV, includes pre-trained models[2], which can be directly used with their Viola-Jones implementation. Scale invariance can be achieved evaluating the image at multiple scales on runtime.

[2] https://github.com/opencv/opencv/blob/master/data/haarcascades

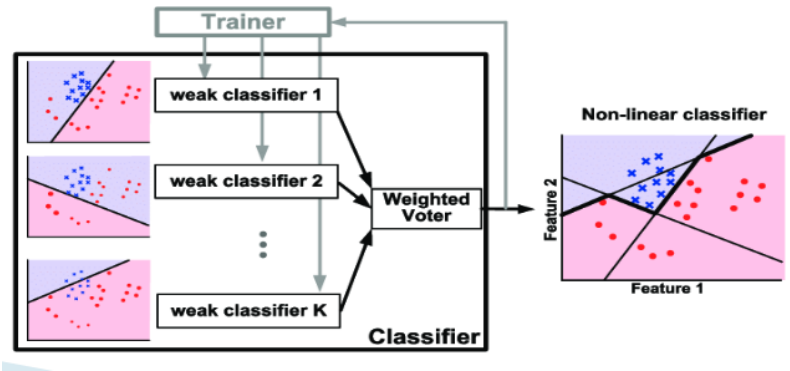


Fig. 2.2. Boosted weak classifiers [13].

Another common approach for person detection is based on HoG (*Histograms of Gradients*) [14]. This method computes local features by means of the intensity gradients across the image, and quantizes them according to their orientations (creating a histogram of oriented gradients for an image block), as it can be seen on Figure 2.3.

These gradients are collected in 64×128 windows, and treated as features. These features are evaluated by a linear SVM (*Support Vector Machine*), which is trained to classify a window as *person/non-person*. Figure 2.3 shows the average gradient patch for a person (the direction of each gradient is not shown). A visual inspection immediately resembles the shape of a person standing up. Thus, this detector will yield the best performance when the person to be detected stands in that specific pose. This template allows as well to retain the gradients placed in the edges of the body (positive gradients), and discard those inside the body (negative gradients), weighting them according to their position inside the mentioned template.

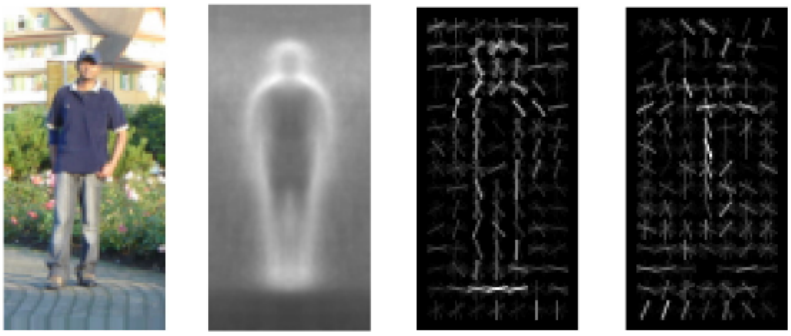
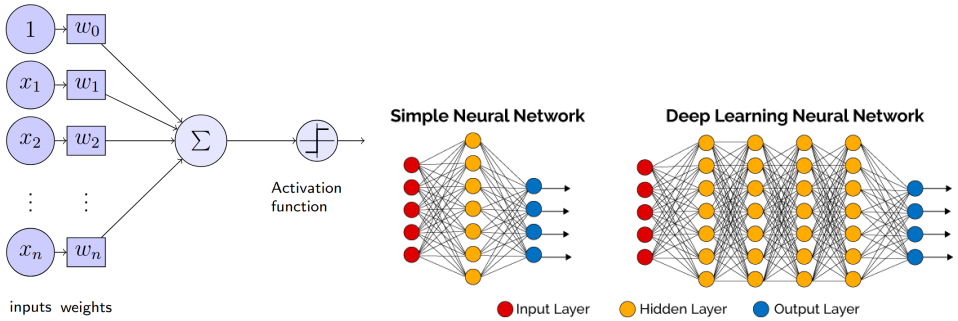


Fig. 2.3. Example of the HoG computed for a person. From left to right: original image, average magnitude of the gradients on a person, directions weighted positive and negative gradients found in the input image. Image from [14].

These methods, among several more, have been the state-of-the-art techniques: the cornerstone are the image gradients, which can be computed with a high efficiency, represented in a compact way by means of a histogram and provide decent performance. Their main drawback is the *generalization* capability, as a successful detection is highly

dependent on the person pose. However, in the latest advances, the detection techniques have moved towards the spreading paradigm: *deep learning*, especially the most salient tools on image processing: CNNs (*convolutional neural networks*).

CNNs are based on standard neural networks, which combine lots of neurons or *perceptrons* organizing them into layers. These perceptrons (Figure 2.4a) implement simple non-linear operations, that allow to extract (after a proper training process) abstract features, which gain in complexity when the number of internal layers increases. When a neural network is composed by many *hidden* layers (in addition to the input/output ones), it is placed into the *deep learning* paradigm (Figure 2.4b), as opposed to *shallow learning*.



(a) Basic unit of a neural network: the perceptron (source: [11]).

(b) Standard neural network vs. deep neural network (source: [11]).

Fig. 2.4. Basis of deep neural networks. (a), schematic of a perceptron. (b), increment on the number of hidden layers on deep learning approaches.

Based on this approach, and taking advantage of the *spatial correlation* when the signal to process is an image, a neural network can be modified to implement a different operation on each perceptron: a *convolution* (Figure 2.5).

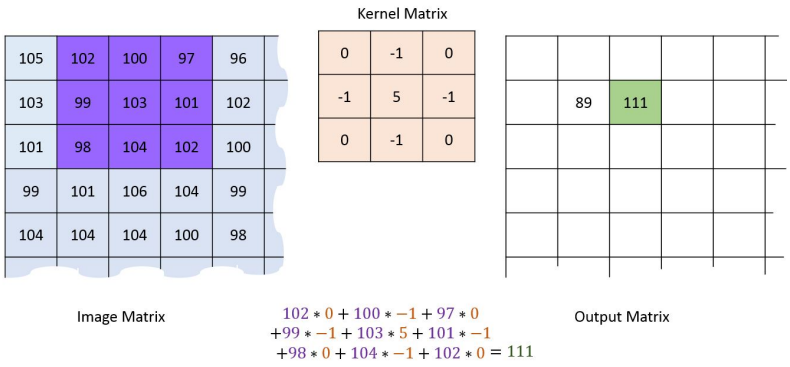


Fig. 2.5. Convolution applied to an image, applying the mask (red) on a region (purple) of the input image, storing the result on the mapping of the central pixel of the region (green). The computation is the sum weighted by the mask values (bottom) (source: [11]).

As it can be seen on Figure 2.6, convolutional units may be arranged conforming a set of layers to build *feature extraction* stages (shown in red in the figure). Several layers

can be concatenated, gaining in depth and obtaining more complex feature maps. These layers are finally followed by a detection/classification ensemble of *dense* layers (shown in blue in the figure): a set of layers with standard perceptrons fully-connected among them, yielding a final output, dependent on the classification structure of the network.

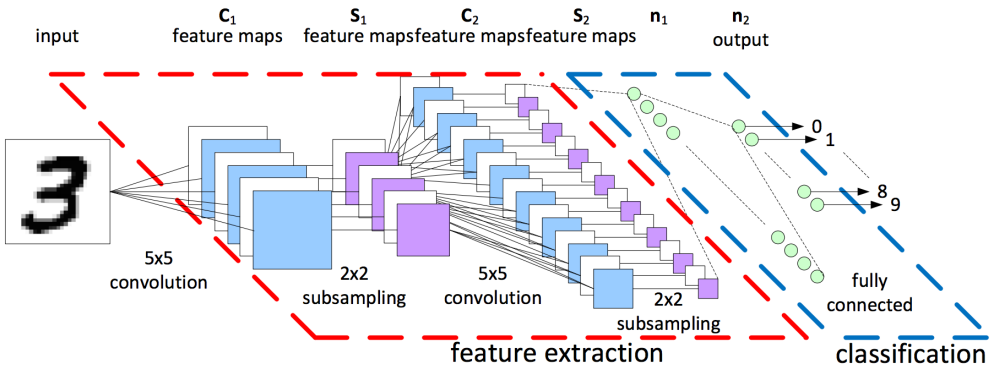


Fig. 2.6. Schematic of a digit classification CNN (source: [11]).

In the case of object detection networks (the ones involved in this work), the output varies depending on the implementation, but it is generally composed of a set of *(location, probability)* tuples, one for each class the network is capable of detecting. Figure 2.7 shows the activation maps of an object detection network, where the map presents higher values in the regions with high probability of containing the object of the class it is designed for. On a convolutional layer, each neuron computes several activation maps across the dimensions of the input data.

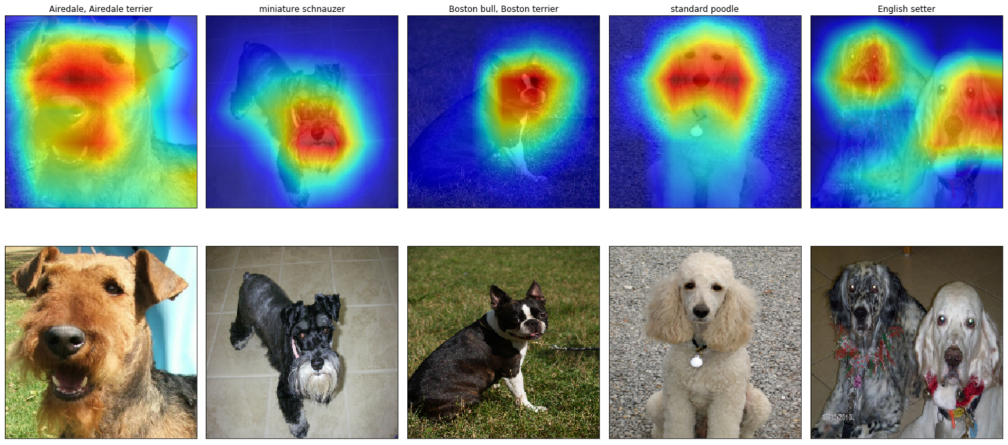


Fig. 2.7. Activation maps of a detection CNN searching for dogs on different images (source: [11]).

One possible application of this concept is focused on what is called *Region-based Convolutional Neural Networks* (R-CNNs) [16], which require a previous step on the image called *region proposal*. This step is devoted to find potential regions on the image to contain an object. This way, the challenge is to label these regions according to the objects contained inside, reducing the problem to a classification task. However, the

process to find these regions and iterate over them makes the process too slow for real-time requirements, which are explicitly considered in our requirements. A notable effort has been made in later works [17] [18] to reduce this computation time.

2.1.1. Single-Shot Multibox Detector (SSD)

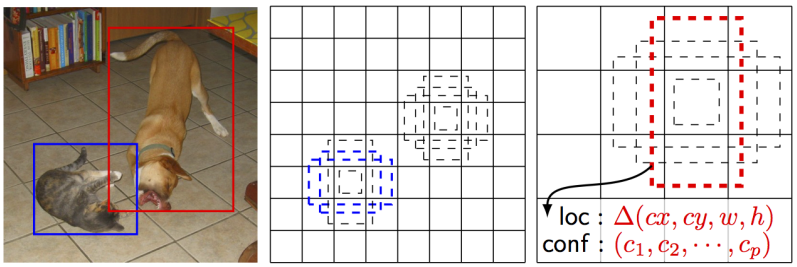
Another outstanding object detection architecture is SSD (*Single-Shot Multibox Detector*) [19]. The main benefit from this architecture is the fact that it embeds all the required computations in a single neural network, reducing the complexity compared to other approaches requiring external region proposals, as it was explained above. This greatly reduces the computational time when the network has to process an image. The architecture can be seen at Figure 2.13, and can be split into several stages [11], namely:

Reshape: the first task to be addressed by the network is to reshape the input image(s) to a fixed size on which the rest of the layers work. In the case of an SSD detector, this shape is $n \times 300 \times 300 \times 3$ (being n the size of the input batch, as n images can be evaluated simultaneously on the neural network). Other image sizes might be used, however this one offers a good trade-off between performance and computational load.

Base network: this first group of layers are reused from a typical image classification model, such as VGG-16 [20]. The first layers of this architecture are utilized in this design, truncated before the first classification layer. This way, the network can leverage the *feature maps* from the classification network, in order to find objects inside the input image. Following the first part of the network, several convolutional layers are appended, decreasing in size. This has the objective of predict detections at multiple scales. One thing to mention at this point is that the base network can be a different one rather than VGG-16, such as a MobileNet [21], which is highly optimized for running on low-end devices. This is interesting as our embedded system will be limited in computing power. It will be revisited in future sections.

Box predictors: for each layer in the base network, an image convolution is performed, generating a small set (typically 3 or 4) of fixed-size *anchors*, with varying aspect ratios for each cell on a grid over the activation map (Figure 2.8). As these maps have different sizes, the system is able to detect objects in different scales. The anchors are then convolved with small filters (one per depth channel), which output confidence scores for each known class, and offsets for the generated bounding box. These scores are passed through a *softmax* operation, that compresses them into a probability vector. Thus, for each detected object (on that scale), the network computes the score on every class and its estimated position inside the feature map (hence, in the image as well).

Postprocessor: as several detections might be triggered in the same area for different classes and scales, a *Non-Maximum-Supression* [22] operation is performed at the output of the network to retain the best boxes, under a combined criteria of detection score and IoU score (*Intersection over Union*), which measures the overlapping quality between two bounding boxes, as it can be seen in Figure 2.9.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Fig. 2.8. A set of boxes are generated centered on each point of every feature map [19].

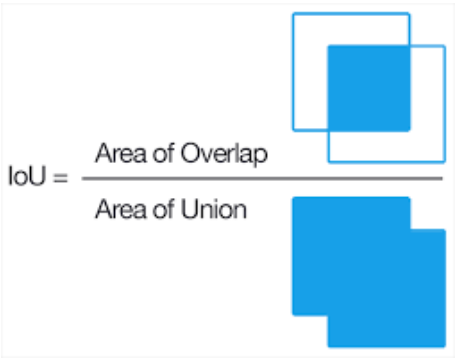


Fig. 2.9. Graphical representation of the IoU score between two bounding boxes.

2.1.2. YOLO (You Only Look Once)

Another interesting approach is the YOLO (*You Only Look Once*) system [23]. Its main advantage is its inference speed, due to the fact that it performs a single analysis on the entire image, dividing it into a grid of cells. Each cell predicts up to 5 boxes, containing an *objectness score* (the predicted IoU of the proposal with an object, regardless its class), the coordinates of the bounding box, and a probability for the object belonging to each class. This design run faster than other methods [23], however it presents a poor performance when detecting small objects.

This design was revisited in YOLO9000 [24], introducing several improvements such as batch normalization at the input of the convolutional layers, or the concept of *anchor boxes*: the box proposals follow a fixed set of aspect ratios, chosen previously using clustering on a training set. As it can be seen on Figure 2.10, limiting the proposal shapes to 5 fixed sizes improves the performance while maintaining a high IoU metric. A visual

inspection shows that the selected anchors seem like a reasonable shape for the majority of the objects the network aims to detect. Additionally, the number of deep layers was increased from 26 layers to 30, and a semantic modeling is performed on the labels across different datasets, allowing the network to be trained in different datasets under a common semantic structure called *WordTree* (Figure 2.11).

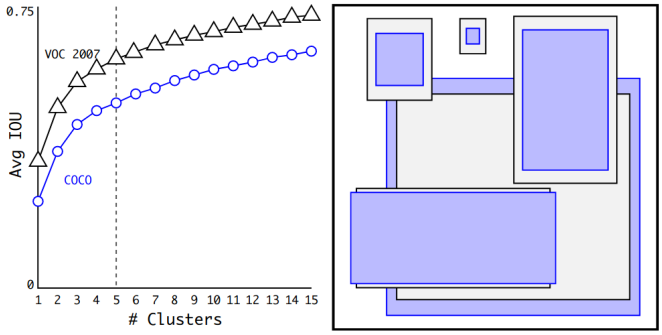


Fig. 2.10. Result of the anchor k-means clustering on VOC and COCO for YOLO9000. Using $k = 5$ anchor sizes on the right yields a good tradeoff between simplicity and improvement on the obtained IoU with respect to using $k - 1$ clusters (source: [24]).

The latest improvement of YOLO, YOLOv3 [25], relies on residual networks [26], which tackle the problem of *vanishing gradients* when the networks become deeper. The stacking of several layers results on gradients diminishing its value up to a point the arithmetical precision of the machine is not able to handle. The gradients are canceled, hindering the training process, as the first layers parameters take a substantially higher time to converge. The residual networks added in this revision of the design add shortcut connections across the layers, focusing the backpropagation gradients on the differences between the input and the output of the layer. As this reference states [25], the combination of these residual layers and convolutional ones allows to train much deeper architectures (53 convolutional layers), capable of yielding a higher generalization. As in the SSD detectors, the YOLO architecture performs multi-scale detections, using 3 scales for splitting the feature maps into cell grids. A similar k-means clustering than in Figure 2.10 is performed on the COCO dataset, selecting 9 anchor sizes instead of 5, and grouping them in 3 scales. Now, on each of the cells, 9 anchor bounding boxes are fit (3 anchor shapes $\times$ 3 scales). This aims to improve the poor performance of the previous version when dealing with small objects, as well as to produce better generalization: in the R-CNN [16] and the SSD [19] the anchor shapes are hand-picked. These changes, with a tuning on the error function, conform the YOLOv3 improvements over the previous versions.

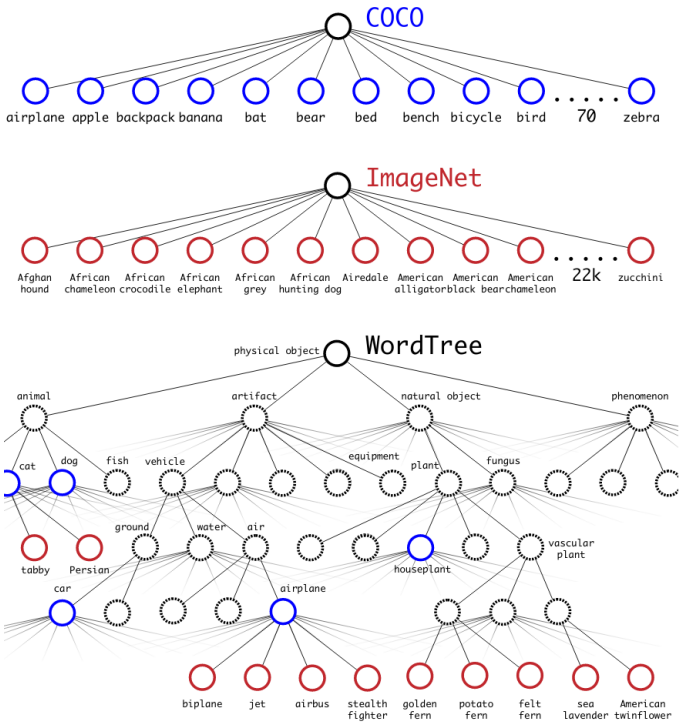


Fig. 2.11. Comparison between simple labeling structures (top) and a WordTree semantic grouping under categories (bottom). This allows to follow a dataset-agnostic training process as the labels can be combined using WordTree. Image from [24].

For each *(anchor, cell, scale)* combination, YOLOv3 predicts:

- The coordinates of the object within the anchor. Details can be visualized on Figure 2.12.
- *objectness* score, which is computed by means of a logistic regression in order to maximizing the probability of overlap with a ground truth bounding box with respect to that of any other prior anchor.
- 80 scores, as the original implementation is trained in the COCO dataset, which contains 80 classes. These classes might be overlapping (e.g. "woman" and "person"). Thus, these scores are computed by independent logistic classifiers and are not passed through a *softmax* operation.

The architecture of a YOLO-based detection network can be compared to that of a SSD-based one in Figure 2.13. This allows to see the fundamental difference in the feature extraction stage of each approach.

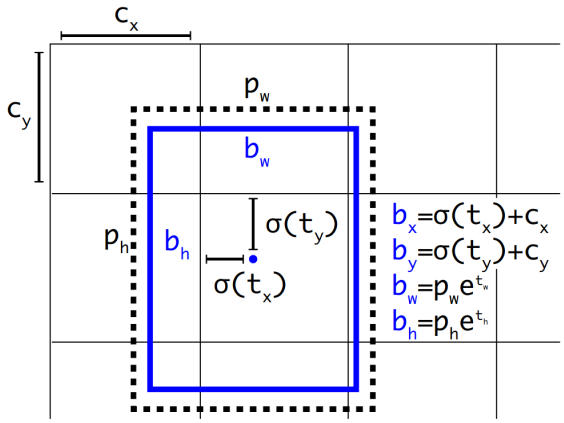


Fig. 2.12. Output on YOLO for each anchor and cell. The dashed line represents the prior anchor, while the blue line represents the detection which corrects that anchor.

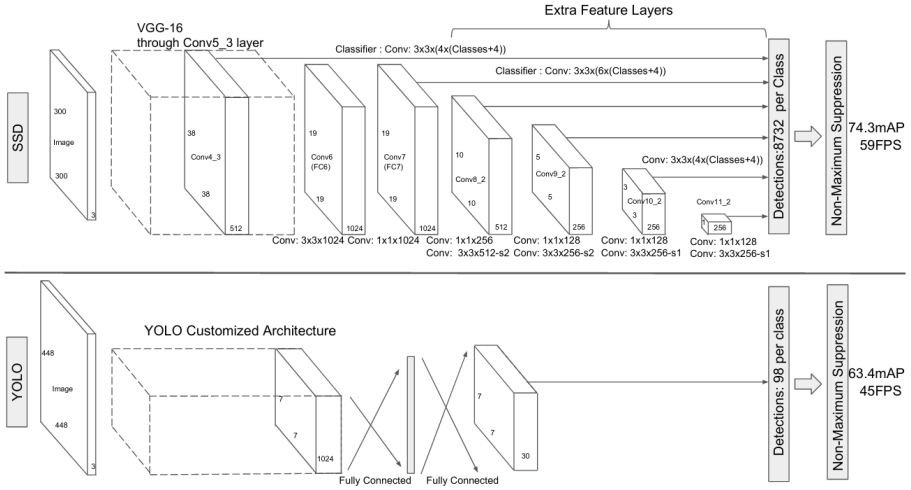


Fig. 2.13. General architecture of a SSD network (top) and a YOLO one (bottom). Image from [19].

2.2. Person identification

On a controlled environment, where the only present person is the one to be followed, a person detection system could be enough for following purposes. However, in a real scenario, there might be several people inside the field of vision of the robot. This problem can be approached by means of a distinguishing feature of the person of interest, provided beforehand. One example is [27], which computes the color distribution of the person of interest, and later compares this distribution with the ones belonging to the different persons using the Bhattacharyya coefficient [28] (a measurement of similarity between two probability distributions). This metric can be applied for computing the similarity between the color histograms of the reference person and the detected one. However, this system can be deceived replicating the color distribution of the person of interest: wearing similar clothes helps to reduce the distance between the histogram, leaving a chance to confound another person with the one to follow.

A more robust approach is to use the *face* of the person as the discriminant feature, as its uniqueness makes it a good reference to identify the detected person. As it is summarized in [29], several applications extract facial *landmarks* from the morphology of a given face (Figure 2.14), and use them to recognize the face, comparing it with a set of known faces and estimating the identity based on the distance to each known face. Some open-source libraries such as `dlib` and `OpenCV` provide algorithms to perform these processes.

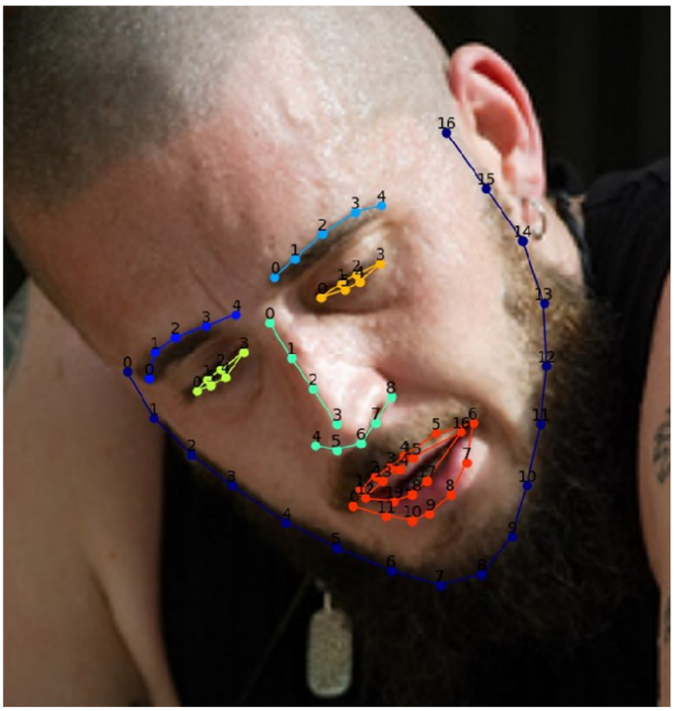


Fig. 2.14. Facial landmarks are dependent of the face shape and morphology (image from [29]).

The intuition behind these methods are to *project* the image of the face into a lower-dimensional space, which allows to extract significant features from each face. These features have to be consistent for the same face across different pose and lighting conditions (Figure 2.15). An useful transformation when a dimensionality reduction is pursued is PCA (*Principal Component Analysis*), a linear transformation that can be implemented to deal with the face recognition problem [30].

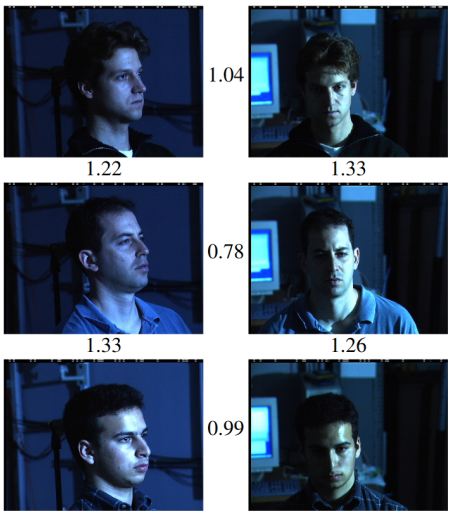


Fig. 2.15. Examples of poses and light conditions across which the face projections are desired to be consistent for the same person (image from [31]).

2.2.1. Deep learning face identification: FaceNet

However, once again neural networks can be leveraged in order to achieve better performance: as the PCA is a linear operation, it could be learned by a single layer neural network. Thus, the introduction of deep networks can yield interesting results. The most relevant approach so far uses deep convolutional networks for performing this process [31], implementing an architecture called *FaceNet*, which is partially based on the Inception [32] module, designed by Google researchers in order to greatly reduce the number of parameters in a neural network. What this network computes is called an *embedding*, a projection of the input face image into a point in a 128-dimensional hyper-sphere. This allows to translate the identification into linear algebra terms, such as *distance* between two faces, as well as clustering and applying unsupervised algorithms. The architecture can be visualized in Figure 2.16. These networks can be trained using a loss function called *triplet loss*, inspired by the work in [33]. Given a training sample (*anchor*), a *positive* example (same class than the anchor) and a *negative* example (different class than the anchor) are chosen, and the network is tuned to maximize the *anchor-negative* embeddings distance, and minimize at the same time the *anchor-positive* one (Figure 2.17).

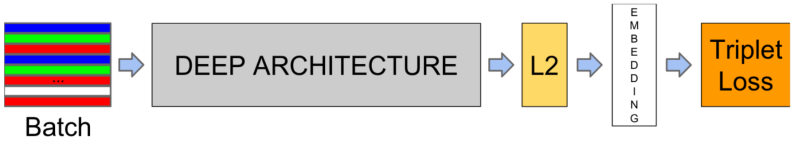


Fig. 2.16. Architecture of the FaceNet system (from [31]).

One thing to mention about the algorithms described above is that they perform the operations on the face image. Thus, a face detection is required for previously cropping the face of the person to be identified. One interesting approach using this technique is

Fig. 2.17. Triplet loss training. It minimizes the distance between an *anchor* (current example) and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity (from [31]).

faced [34]. This is a custom small ensemble of two neural networks, responsible to detect faces and correct the bounding boxes found. The main objective of the system is *speed*, so the main detector architecture is based in YOLO [23], and the second correction stage raises the precision achieved by the detector, achieving better results than a classical Haar approach, as illustrated on Figure 2.18. Further comparisons are performed on Chapter 4 between these two detection methods.

Fig. 2.18. Classical Haar based face detector [12] (left) vs. *faced* (right). Image from [34].

2.3. Embedded deployment

One of the requirements of this work is to be integrated in an autonomous robot. This imposes a power limitation on the algorithms to be deployed. Generally, the robotic systems are deployed using laptops connected to robots, as it was done in [11] (Figure 2.19).

Nowadays, the mentioned increase in the interest into the real-time computer vision applications has fostered the development of specific low-power embedded devices to be integrated in mobile systems. The extending usage of devices such as Arduino or Raspberry Pi has led to embedded robotics systems, such as PiBot [35] (Figure 2.20). These robots are useful in the educational scope, as they are capable of running simple vision and navigation algorithms at a low cost.

Unfortunately, the requirements for running more complex algorithms, such as neural networks, require of the next tier in power terms, keeping the portability nevertheless. The ideal device could be an ASIC (*Application-Specific Integrated Circuit*), as the custom design would lead to a very tight optimization of the performance. However, the objective

(a) Frontal view.

(b) Side view.

Fig. 2.19. Laptop+robot deployment on [11].

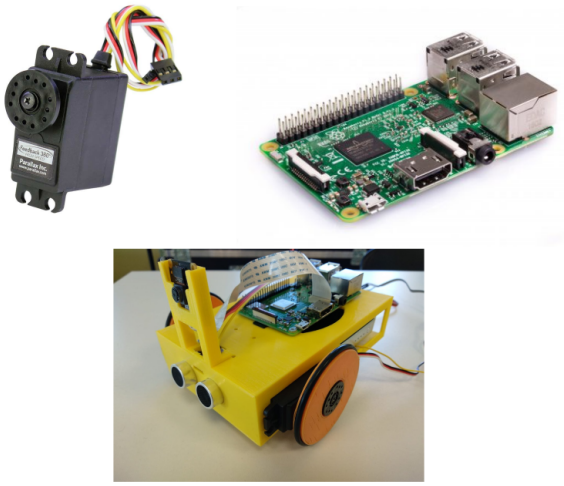


Fig. 2.20. PiBot, an open low-cost robotic platform for education (image from [35]).

is to run the algorithms on existing software frameworks, requiring to use general purpose computers instead. The most remarkable advance in this scope are the Jetson devices manufactured by NVIDIA. These development boards are SoM computers running a tailored version of Linux. The fundamental feature of these systems is that they include a high-performance GPU featuring CUDA, a low-level parallel computation library, as well as several toolkits (such as TensorRT[3]) designed to optimize as much as possible the software implementations for the plethora of possibilities to be designed on this board. As it can be seen in Figure 2.21, its size and power consumption make this system a good choice to be included in an autonomous robot.

[3] https://developer.nvidia.com/tensorrt

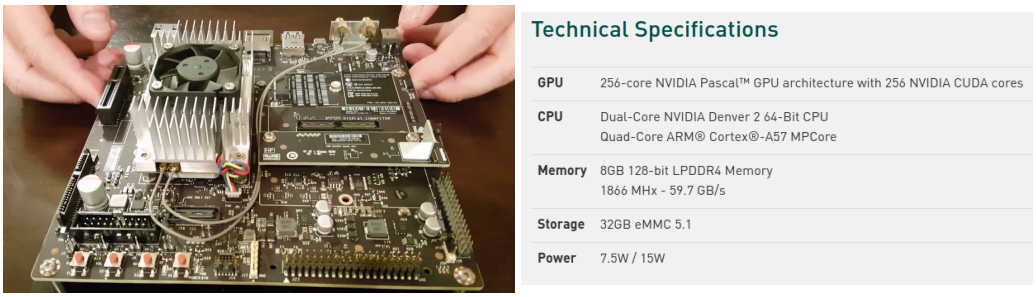
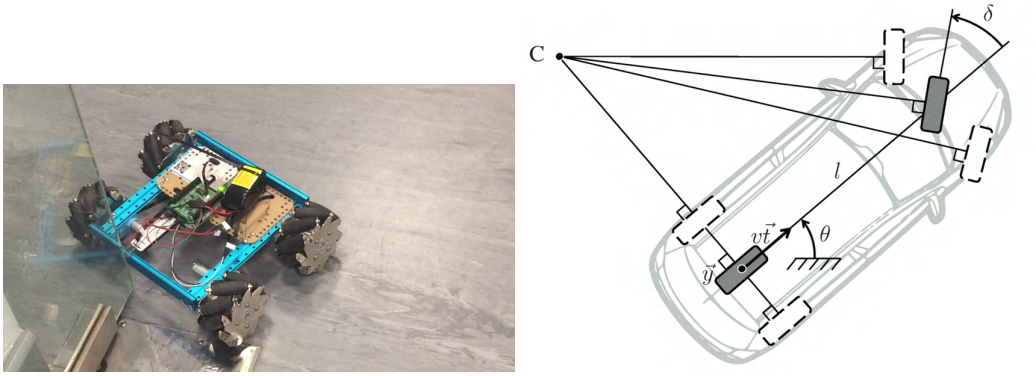


Fig. 2.21. NVIDIA Jetson TX2: an embedded high-performance device including a GPU.

2.4. Person following

Several approaches have been developed pursuing this challenge of *following a person*. Once the visual perception algorithms are established, the final output of the pipeline has to be a movement command for the robot to move towards the desired point. Mobile robots can be classified according to their locomotion capabilities. A robot is *holonomic* if the number of its controllable degrees of freedom is equal to its total degrees of freedom. If the controllable degrees of freedom are lower than the total degrees of freedom, the robot is *non-holonomic*. This difference can be observed on Figure 2.22. In the case of a holonomic robot, the navigation process is simplified, as the robot can instantaneously move to a desired target. However, a non-holonomic robot needs to perform maneuvers in order to move towards a point.



(a) Holonomic robot.

(b) Schematic of the degrees of freedom of a non-holonomic vehicle (a standard car).

Fig. 2.22. Comparison of a holonomic system with a non-holonomic one.

The summary on [36] shows an interesting classification of some existing person following algorithms and their applications (Figure 2.23).

Some approaches leverage the detected objects in order to estimate the relative homography of the orthogonal planes, which allows to partially know the environment of the robot and trace a safe path towards the person, as it can be seen on Figure 2.24a.

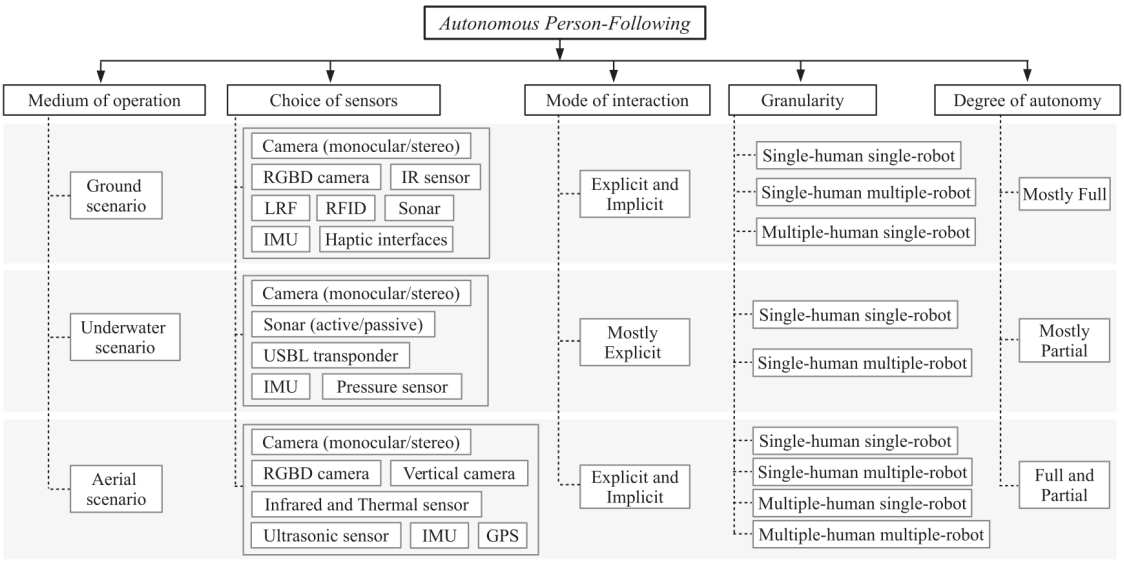
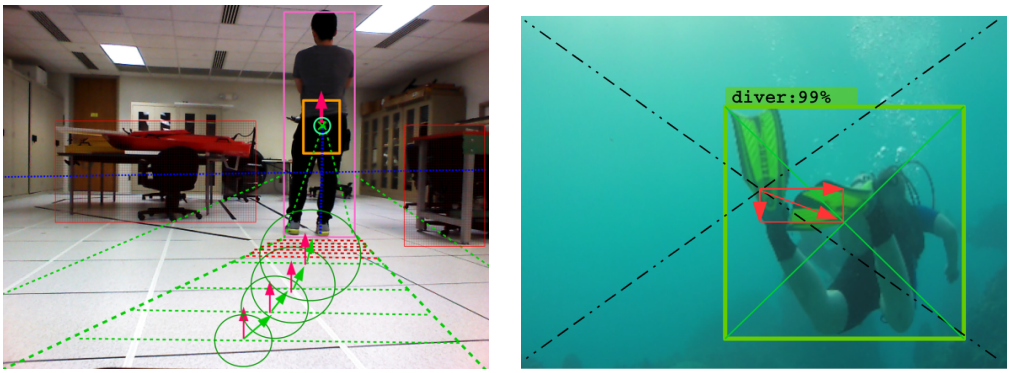


Fig. 2.23. In-depth classification of the existing person following algorithms (image from [36]).



(a) Following with path computation using homographies (image from [36]).

(b) Example of underwater reactive following (image from [37]).

Fig. 2.24. Examples of robotic following behavior.

Other approaches act without a path planning component, implementing what is called a *reactive* behavior [37], similar to the proposed solution on this work. On these approaches, the vector between the center of the image and the center of the person is used to command movements on the robot, as it can be seen on Figure 2.24b.

3. MATERIALS AND METHODS

This chapter is devoted to describe the developed system. The development strategy was based in splitting up the functionality into different modules, which have been tackled sequentially. The next sections will cover each one of the modules, and will describe the achieved solution. Finally, the full ensemble will be described and tested.

3.1. Available materials

3.1.1. Hardware

Base board

As it was described in Chapter 2, typical following behaviors work on a personal computer attached to a robot. However, our solution is developed using a devoted SoM: the NVIDIA Jetson TX2, similar to the one described in Figure 2.21. This system features a high-performance GPU, and low-level optimization engines, which greatly reduce the time required to perform the operations required for deep learning applications, such as tensor convolutions. The low power consumption of this board (15W at full power) makes it suitable to be embedded in a portable robot equipped with a battery. One drawback of this system is the scarce storage space. However, this can be immediately solved by installing an external storage device using its integrated SATA connector. In this project, a 120 GB Kingston SSD (*Solid State Drive*) was used for this purpose, leveraging as well on the high transference throughput this device can achieve. It features a 64-bit ARM processor, and it mounts a fully functional Linux system. As it is equipped with two WiFi antennas, a remote control interface can be easily set using SSH connections. Regarding the available RAM in the board, it is limited to 8 GB, to be shared by the GPU and the CPU. This jeopardizes the execution of the deployed software and the neural networks, which have to be controlled in every moment in order to save the maximum amount of RAM possible. The resulting board can be visualized on Figure 3.1.

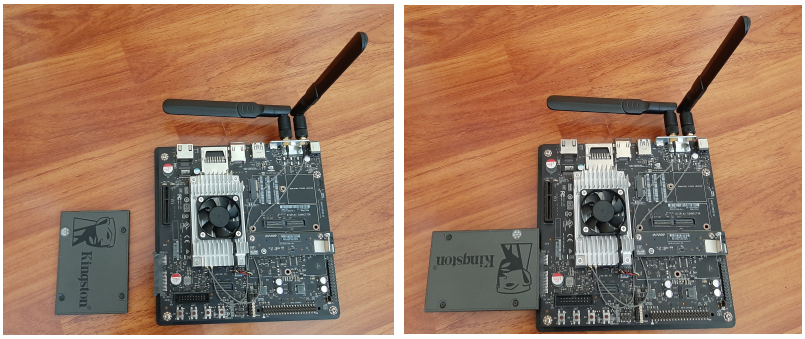


Fig. 3.1. Resulting system: Jetson TX2 board and the installed SSD drive, plugged into the SATA connector.

RGBD sensor

The vision system used in this work, the ASUS Xtion Pro Live (Figure 3.2), is a USB device composed by a RGB camera and an IR (*Infra-Red*) emitter + sensor system, capable of retrieving depth data for each pixel on the image. This is achieved by emitting a known light pattern (Figure 3.3), which reflects in the present surfaces on the scene. These reflections are captured by the IR sensor, inferring the position of the surfaces from the received distribution of the IR pattern.

Fig. 3.2. ASUS Xtion Pro Live

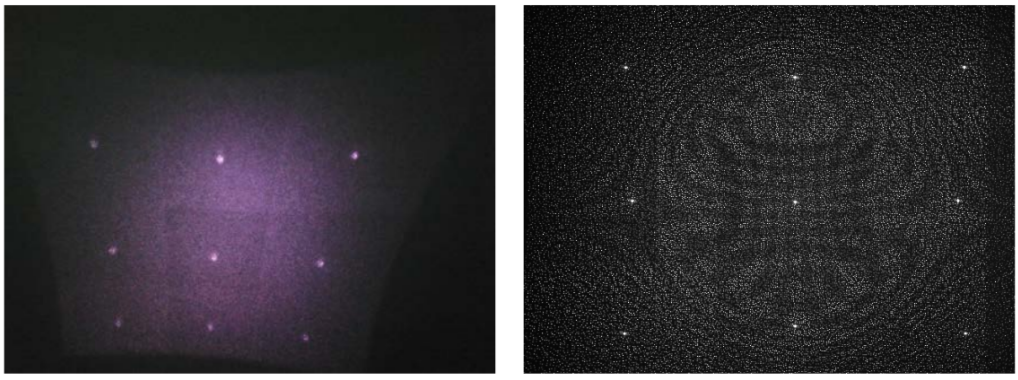


Fig. 3.3. Infrared pattern emitted by the Xtion (images from [38]).

The last problem to be tackled by this device is the discrepancy caused by the different points of view of the RGB and depth sensor. However, as the distance between these two sensors is fixed and known, a *registration* process can be carried on inside the device, projecting the depth data into the RGB pixels [39].

The systems which implement the described design are called RGBD sensors. These are suitable for robotics, as the yielded result is a point cloud, reflecting the distance from the camera for each pixel in the image. Using this, the device is capable of projecting the 2-dimensional RGB image into the 3D space by means of the depth data (Figure 3.5).

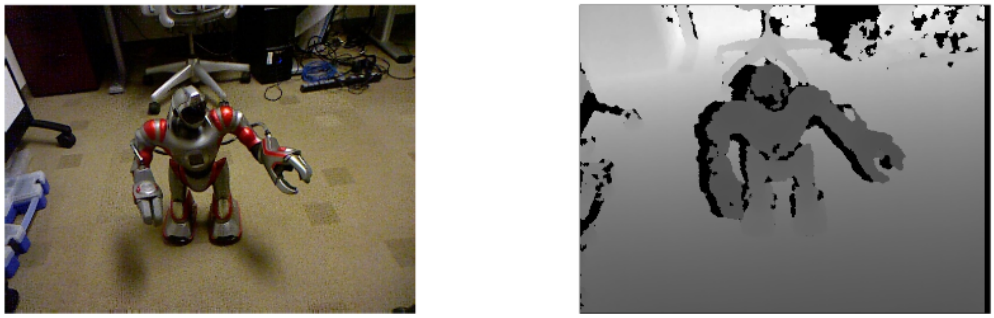


Fig. 3.4. Discrepancy between the RGB and depth images (image from [11]).

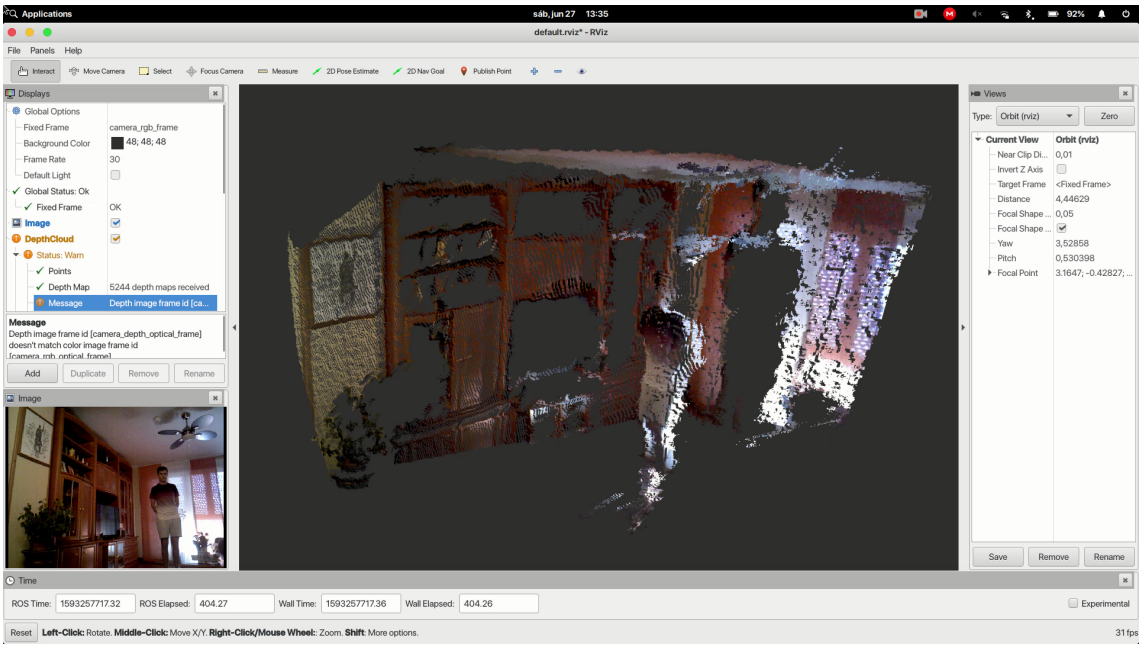


Fig. 3.5. Visualization of the RGB image (bottom left) and the resulting point cloud projected into the 3D space (right).

Robotic platform

On the other hand, the robot used in this work is the Turtlebot2 educational set. It is based on a Yujinn Robotics Kobuki mobile base (Figure 3.6), which is a non-holonomic robot with 2 degrees of freedom: *linear speed* and *angular speed*.

In the Turtlebot2 set, the mobile base has an attached structure, carrying the RGBD sensor and a platform where typically a computer can be placed. This platform is useful to mount the NVIDIA Jetson device on. Additionally, as it can be seen in Figure 3.6b, the Kobuki panel is equipped with a 12V output, yielding up to 1.5A, which in power terms can be translated to a maximum power of 18W. Since the TX2 board peak consumption is 15W, this connector is suitable to power the system up, with an additional power margin of 3W. A lookup in the Kobuki user guide [40] allows to find the suitable Molex connector, which can then be attached to two-wire cable and a rounded connector. This provides the

(a) Appearance of the mobile base robot.

(b) Schematic of the connections panel of the Kobuki.

Fig. 3.6. Kobuki mobile base, which carries the rest of the structure.

NVIDIA Jetson of a 12V DC supply, similar to what it would obtain from a power outlet with a transformer. As the power input is equipped with a DC voltage regulator, it accepts voltages from 5.5V to 19V (table 59 in [41]).

Hence, this is a successful approach to build an *autonomous* system: powering the computing board from the batteries of the robot, with enough autonomy to be powered on for several hours. The amount of time strongly depends on the usage of the motors of the mobile base, which are the most consuming component of the ensemble.

The final hardware setup is displayed on Figure 3.7, where the described components are combined to build the autonomous setup capable of running high-complexity person following algorithms.

(a) Front view.

(b) Rear view.

Fig. 3.7. Autonomous setup: Turtlebot2 + Jetson TX2 + ASUS Xtion Pro Live.

3.1.2. Software

NVIDIA JetPack

The development of our person following behavior has been tackled using exclusively open-source software. The Jetson computing board follows a tightly optimized embedded design guidelines. A tailored version of Ubuntu Linux, named NVIDIA JetPack, is developed and maintained by NVIDIA, and it is available for download and install as the board firmware. For the developed system, the version used is JetPack 4.2.2 (R32.2)[4]. This custom implementation includes low-level interfaces for implementing parallel computing operations (CUDA), and several optimizations SDKs (*Software Development Kits*), such as TensorRT. This engine is of special interest for us, as it allows to optimize the low-level implementation of a neural network, swapping certain modules (such as a convolution operation, a ReLU, or an Inception block), for a low-level optimized version of that module, allowing to greatly increase the inference speed without losing precision. More details about these optimizations will be explained later.

Python

This work has been developed using the Python programming language. In the previous work [11], the used version of the language was Python 2.7. However, as of today, that version has reached its EOL (*End of Life*) date, remaining unsupported. For avoiding the obsolescence, all the code base was migrated to Python 3.6, a currently supported release, before making any change or improvement in the functionality.

NumPy

NumPy[5] (*Numeric Python*) is a library for Python (written in C++), born to extend the numerical capabilities of this language. It provides a powerful `ndarray` class, which allows to keep an N-dimensional collection of values/objects in a really handy way (in comparison with Python's standard *lists*). It also provides a rich set of methods to manage arrays (such as advanced indexing, shaping, data formatting, etc.).

These capabilities immediately turn this library into an excellent framework for data processing in a lower level. It allows to store and handle images and tensors on an intuitive way, providing methods to perform typical tasks such as row-wise/column-wise averaging, transposing, type conversion, or matrix slicing. The majority of these

[4]Details available on: https://docs.nvidia.com/jetson/archives/jetpack-archived/jetpack-422/release-notes/index.html

[5]http://www.numpy.org/

structures and methods are implemented using the C++ language, which provides a higher speed than a Python implementation.

ROS

This project requires hardware-software interaction, as the development board needs to read the images captured by the Xtion sensor, as well as sending the final velocity commands to the robot. For this purpose, the ROS middleware is used. ROS (*Robot Operating System*) is "*an open-source, meta-operating system for your robot*", maintained by the *OSRF* (*Open Source Robotics Foundation*) [42]. It is a framework that provides a distributed, easily-scalable environment of *nodes*. These nodes are programs which run independently on the computer (or distributed over a network), so they can perform individual tasks. However, they can communicate between themselves on a synchronous way (over *services*, implementing a client-server role system between nodes), or on an asynchronous way, via *topics*. These topics, which rely on a standard TCP/UDP communication between sockets, are intended for an unidirectional, streaming communication, where a node can take roles: *publisher* (if it is writing data inside the topic), or *subscriber* (if it is reading the data that publishers are broadcasting into the topic). The data stream through the topic is not unrestricted, it must follow a ROS specific syntax, a *Message* type, which is strictly defined for the communication purpose (geometry, sensoring, etc.).

For this project, the packages `cv_bridge` and `openni2_camera` have been used for handling the RGBD data. The robot can be controlled with the package `kobuki_node`. All the software architecture is controlled by `rospy`, the interface for Python to communicate with the described ROS infrastructure.

Another useful feature of ROS middleware is the *ROSBag* storage system. Recording a ROSBag allows to save in a single file the messages read from several topics for the time it is recorded. Later, the ROSBag can be played again to recover the messages from the topics, in the same order they were recorded. This is useful for recording video sequences from the RGBD camera, saving simultaneously the image and depth information, allowing the user to perform testing of different parameters using the exact same image source.

As well as in the Python case, the version of ROS used on [11] reached its EOL date. Thus, the ROS version has been migrated as well to the currently supported release: *Melodic Morenia*, which firstly provided the compatibility with Python 3. As the Jetson TX2 board is based on an ARM architecture, this upgrade has required several tweaks on the software compiling and implementation processes, which have been properly

documented in the project repository[6] for the sake of repeatability.

OpenCV

For general image processing, OpenCV (*Open Source Computer Vision*) is a C++/Python/Java open-source library (natively written in C++) for Computer Vision purposes. Among the classic/*state-of-the-art* methods it bundles, several functions can be found suitable for face recognition, image stitching, eye tracking, computing homographies, establishing markers for augmented reality, etc.

OpenCV focuses on *efficiency and real-time functionality*, due to the low-level optimizations at hardware level (i.e., integration with NVIDIA CUDA and OpenCL GPU processing libraries). Thus, the excellent performance achieved by this open source library has turned it into the *de facto* standard for every kind of users (from researchers to big companies or even governmental bodies, as their website stands[7]).

This library has been used across the entire project, on its version number 4.2. It has been useful for diverse tasks, such as image normalization, drawing, computing local features or optical flow approximations.

TensorFlow

The deep learning framework used is TensorFlow. This is a high-performance numerical computation library, strongly focused on parallel computing, typically carried on by GPUs or processing clusters. This library is a state-of-the-art tool to deploy deep neural networks because of its efficiency. Besides of training/running a neural model, this library allows to load a pretrained model from a storage device, by means of a *frozen graph* file. This file contains both the network definition and the weights of its nodes.

Additionally, a binding component called `TensorFlowRT/TRT` have been used to implement the low-level optimizations on the TensorFlow neural engines, as it will be described later.

[6] https://github.com/RoboticsLabURJC/2017-tfg-nacho_condes
[7] https://opencv.org/

3.2. Design

The software implemented in this work has been divided into two main components or modules, namely the *Perception* module and the *Actuation module*, which can be observed in Figure 3.8.

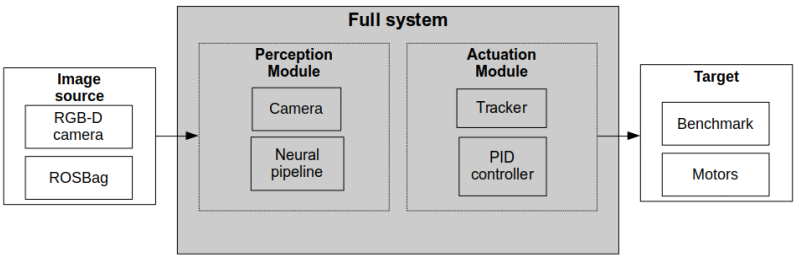


Fig. 3.8. Functional architecture of the developed work, showing the two main blocks.

These two modules cope with specific tasks on an independent manner, as it will be described in the following subsections.

3.2.1. Perception Module

This module encompasses what the robot perceives from its sensors (the camera, in this case), and the subsequent processing of the images in order to determine the location of the person to be followed.

Camera

As it was described before, the Xtion device yields two simultaneous images: an RGB image and a depth image. The ROS controller for the camera, OpenNI2[8], fetches the image and registered depth map from the camera, making this information available through several ROS topics. As ROS follows a *publisher-subscriber* semantics, once the driver is up and running, any application may subscribe to the topics in order to receive all the published messages. In our *Camera* module, two subscribers are deployed to retrieve the latest *(RGB, depth)* pair on an asynchronous way. These images are then converted into the standard image format in the OpenCV library, and they are ready to be used by other components. Additionally, in order to be able to perform objective testing and benchmarks, the Camera module is able to retrieve the images from a recorded ROSBag instead of the online camera. This is useful to obtain objective metrics of another components of the software on unit tests, as the ROSBag ensures that exactly the same images are used regardless the tested system.

[8] https://structure.io/openni

The implemented *Camera* module abstracts this condition, allowing to apply the system to an *online* source (camera) or an *offline* source (recorded ROSBag), with a transparent adaptation to the rest of the system. Whenever a new (RGB, depth) pair is required, the *Camera* module will serve the latest available image from the specified source.

Neural pipeline

The captured images are passed through an ensemble of neural networks, which provide the capability of detecting the persons in the scene, as well as identifying which one is the one to be followed. As it was studied in Chapter 2, the most powerful and robust approaches are achieved nowadays using deep learning. Thus, the complex problem of determining the identity and location of the person of interest has been decomposed into three tasks, which are all addressed using the corresponding deep learning techniques:

1. **Person detection:** the *object detection* task (Figure 3.9) is a common one in computer vision. The existing solutions use object detectors similar to those explained in Chapter 2, which are typically trained with large image datasets. The classes these models are capable to detect contain the *person* class. Thus, as it was demonstrated in [11], a deep object detector can be readily used for detecting persons. In this work, several models have been tested, varying the base network architecture and its depth. Since one of the objectives of the system is to work on a portable (low-power) system, only the architectures which yield a good performance with a sufficiently low inference time are considered. The two most suitable models for this purpose are SSD [19] using a MobileNet [21] for feature extraction, and the *tiny* version[9] of YOLOv3 [25]. These models are already trained and publicly available on the TensorFlow Model Zoo [43] and on repositories hosted on GitHub[10]. In-depth tests have been conducted to compare the performance of these two models, which can be found in Chapter 4. The previously developed work [11] only supported SSD-based detectors, however, the object detection component of the program has been upgraded and it features YOLOv3 support as well, making it available through the configuration file specified on launch.

2. **Face detection:** as the previous task, this problem can be addressed using an object detection neural network. However, the previously described models are not suitable for detecting faces, as that object class was not included among the labels on the datasets used for training the networks. In this case, the adopted

[9] The usage of the tiny version of YOLOv3 is due to issues with the limited memory on the Jetson TX2 board. The full model was tried unsuccessfully, as it requires more memory than the available one on a typical execution.

[10] https://github.com/mystic123/tensorflow-yolo-v3

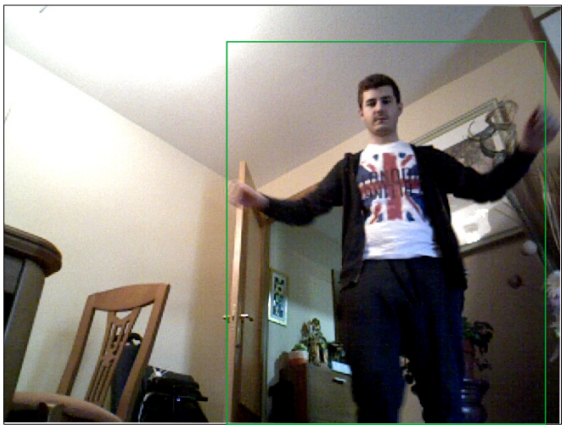


Fig. 3.9. Example of a person detection task.

solution is a single-class detection system. The network trained in [34] implements a two-stage neural network capable of detecting faces. As it was explained in Chapter 2, this detector is based on YOLO, which ensures a high-speed and efficient detection based on a class-specific neural network, which is lighter than a multi-class detection system. The repository where the project is hosted[11] contains a video sequence comparison comparing the accuracy of this system against a classical Haar cascade approach [12]. Chapter 4 contains data and captions of this sequence to show the superior performance for the face detection issue.

3. **Face identification:** Once the face of a person has been detected, it can be used as a discriminant feature for determining their identity. As the basis of this work is to take advantage of deep learning power, a neural system has been selected to perform this task too. For this purpose, *FaceNet* (described on Section 2.2.1) has been used to perform identification, using a publicly available implementation in TensorFlow[12]. As a result, the image of a face is transformed into a 128-dimensional vector, known as projection or *embedding*. This transformation is learned after a triplet-loss training process, which separates different faces as much as possible, while projecting similar faces as close as possible. As it can be seen on Figure 2.15, it produces similar projections when two images of the same face are evaluated, despite different lighting conditions (as a channel-wise normalization step is performed before passing the image through the network).

To sum up, this ensemble of 3 neural networks provides a sequential pipeline to obtain *person locations, face detections and face projections* from a single image, taking advantage of the flexibility and robustness that deep learning methods offer, in order to address three different problems ion an efficient way. Its functionality has been depicted in Figure 3.10.

[11] https://github.com/iitzco/faced
[12] https://github.com/davidsandberg/facenet

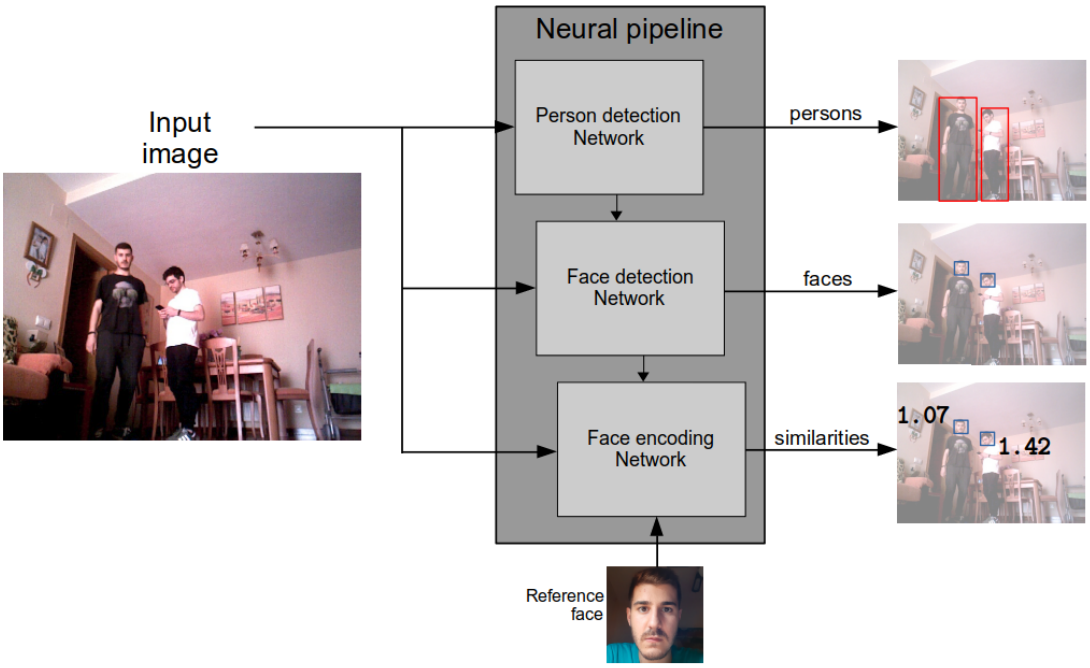


Fig. 3.10. Neural pipeline, showing the cascade of the three neural networks used to output persons, faces and similarities with the reference face.

Once the inference pipeline has been designed and implemented, it can take advantage of the optimization libraries of the Jetson TX2 board, using TensorRT for this purpose. Using this library, several segments from the architecture of a given network can be modified according to certain parameters, as explained next.

MSS (*Minimum Segment Size*): the threshold above which a segment is selected to be replaced by the TensorRT optimization. Increasing this value makes the optimizer more selective, in order to optimize only the heaviest segments of the network. A low value aims to optimize smaller segments, although this may cause an excessively high overhead, causing the resulting graph to run slower than the original one.

MCE (*Maximum Cached Engines*): TensorRT keeps a cache of engines on runtime, with the purpose of reducing the time spent for loading them into the GPU. This parameter modulates the amount of engines kept in that cache, as the available memory to establish the cache is limited.

Precision mode: typically, the weights and parameters of the trained neural networks are handled as 64-bit floating point numbers. A reduction in the precision to 32-bit or 16-bit achieves very similar results (as it will be demonstrated later on Subsection 4.4.2), making the operations much lighter as the precision mode is reduced to the half or the quarter. A more daring approach reduces the precision up to 8-bit integers, performing an additional *quantization* step since the range will be limited

to 256 values. The quantization step analyzes the segment, computing the numeric range of its weights. This range is typically narrow enough to perform a 8-bit quantization, mapping the high-precision weights into a range composed of 256 steps between the minimum and maximum values of the weights.

An experimental tuning of these parameters has been performed in Chapter 4, looking for an optimization of the inference time and taking into account that the enhanced models of the three neural networks have to share the limited available memory on board. Thus, special attention has been payed to the memory footprint that an excessive runtime optimization might cause, as it would lead to a strong penalization if the system cache is utilized to store the models.

The *Camera* and *Neural* components form the *Perception* module, responsible of capturing the external image and extracting pertinent information from the image: position and identity of the person to be followed. This information serves as input to the *Actuation* module, explained below.

3.2.2. Actuation Module

The second module of the system addresses the actuation task: once the external stimuli have been acquired and processed, an action has to be performed in order to move the robot towards its goal. As the final objective of the system is to follow a person, these movements have to be reactive, happening as soon as possible whenever the person changes their position.

Motion Tracker

The previously depicted *Neural* component outputs reliable inferences with a certain refresh rate, namely k frames, which can reach a relatively high value depending on the current load and power profile in the development board. If k is too high, the system may be affected by an important delay when the movement is performed. This may lead to unsteady movements, increasing the probability of losing the reference person. To avoid this, a *Tracker* component is added to the system. Its functionality is to be able to *estimate* the person movement along k frames, while the neural pipeline is performing the next detection. This way, currently detected persons can be tracked along the image while they wander, until the neural ensemble outputs the latest predictions, which determine the true new position of the persons. To fulfill this requirement, the tracking method has to be able to run at a higher rate than k, preferably with a considerably lower inference time. This way, the system counts on a slow, reliable detection system backed-up by a fast tracking system, devoted to guess the movements between detections. This tracker

has been situated in the *Actuation* module. This is because it is focused on keeping the position of the person updated, in order to move towards them as fast as possible. This task is performed without a detection algorithm behind, just moving the box using the estimated optical flow, which is a completely different task than that of the *Perception* module one. For this reason, it has been separated from the neural pipeline and placed in the *Actuation* module.

The method chosen for this purpose is a *Lucas-Kanade* visual tracker [44]. This technique estimates the *motion field* between the images taken in two time instants, addressing the problem using a differential approach [45].

This algorithm relies on the fact that in a video sequence, for small changes in space and time, the intensity remains almost constant within a certain pixel neighborhood:

$$\mathbb{I}(\mathbf{x}, t) \approx \mathbb{I}(\mathbf{x} + \Delta\mathbf{x}, t + \Delta t)$$

Using a 1st order Taylor series approximation and algebra, the *optical flow equation* can be found[46]:

$$f_x u + f_y v + f_t = 0$$

where

$$f_x = \frac{\partial f}{\partial x}; f_y = \frac{\partial f}{\partial y}$$
$$u = \frac{dx}{dt}; v = \frac{dy}{dt}$$

i.e., f_x and f_y represent the image gradients with respect to the space, f_t with respect to time, and (u, v) represents the movement vector over the scene.

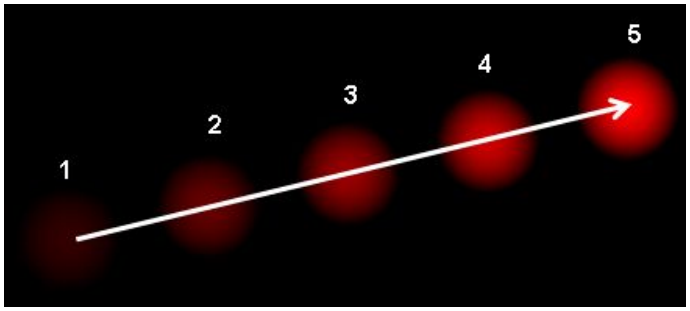


Fig. 3.11. Optical flow for different time instants. Image from [46].

At this point, the resulting system is under-determined as the problem presents 1 equation with 2 unknown variables. Lucas-Kanade algorithm addresses this problem taking advantage of the previously mentioned assumption: in a pixel neighborhood, one can expect the same movement. All the contained pixels will share a common (u, v) movement vector (typically, a small square or circular neighborhood is assumed).

Assembling together those equations results in an over-determined system, where a *Least-Squares* solution yields the best-fitting motion vector (u, v) for that neighborhood, allowing to have a local estimation for the movement in that area:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i {f_{x_i}}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i {f_{y_i}}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix} \tag{3.1}$$

The solution of Equation 3.1 can be efficiently obtained with high-performance libraries, such as *NumPy* or *TBB*, which ensure a fast execution. This makes Lucas-Kanade estimation an efficient approach to compute the optical flow in tasks such as image registration, video stabilization or depth computation in stereo vision systems. This technique is implemented in the *OpenCV* library through the method `cv2.calcOpticalFlowPyrLK`, which iteratively evaluates regions of the image on a pyramid of scales to improve the robustness. This method offers a set of tunable parameters to detect the new position of the corners:

`winSize` size of the window to solve the LS problem.

`maxLevel` number of additional scales to evaluate the image on a scale pyramid.

`criteria` flags to determine the stop condition on the iterations of the algorithm.

However, in the case of study of this work, the objective is not to compute the entire optical flow (it would be an unnecessary consume of computational resources, which are scarce). The estimation can be limited to the pixels inside and surrounding the persons in the scene. Furthermore, one can notice the existence of more informative regions inside the person than others, given its texture: typically object *corners* will be the best choice to be tracked [15], given their easiness to be identified and the fact that they provide more motion information than another areas (aperture problem) [44]. In order to detect these corners, a Harris corner detector can be used. A *corner response* can be computed, yielding a score depending on the eigenvalues and their ratio:

$$R = \det M - c(\text{trace}(M))^2$$

with c being an empirical constant $c = 0.04 - 0.06$, and M being the diagonal matrix resulting of the singular value decomposition of the current window.

The value of R determines the decision taken in the window containing a corner.

A modification of this algorithm, known as the *Shi-Tomasi* corner detector, was published on [47], improving the performance of the corner detector by changing the corner response computation to:

$$R = \min(\lambda_1, \lambda_2)$$

taking the window as a corner if R is greater than a given threshold. The scoring diagrams for determining the corner response on the described methods can be observed in Figure 3.12. One advantage of this methodology is its invariance to rotation, as it works using the eigenvalues, that automatically align to the highest variation directions. However, one important thing to mention as a flaw is the variance to scale: the relative size of the corner with respect to the window size has influence on the eigenvalues, as illustrated on Figure 3.13.

Other methods for corner detection are widely used in state-of-the-art developments, such as SIFT [48] or FAST [49]. However, according to the evaluation among several corner detectors in [50], the Harris/Shi-Tomasi approach yields a more reliable result for this purpose, while taking a low time to execute: it takes around 25 ms to evaluate the 640×480 image from the Asus Xtion, which makes the tracking module to run $5\times$ faster than the neural pipeline.

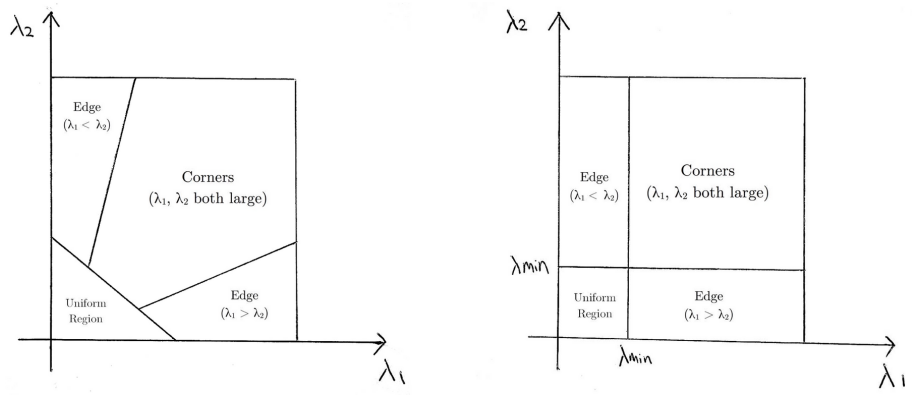


Fig. 3.12. Corner response R scoring functions on $\lambda_1 - \lambda_2$ on the Harris (left) and Shi-Tomasi (right) detectors (source:[51]).

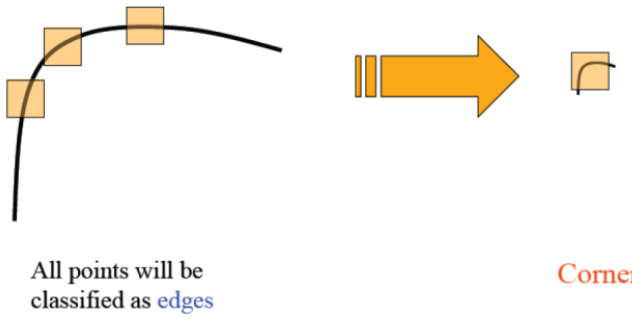


Fig. 3.13. Scale variance of the Harris/Shi-Tomasi methods. It can be seen that the size of the corner with respect to the `winSize` jeopardizes the eigenvalues. Image from [15].

Using this method returns what the authors call the *good features to track*, namely, the best N corners of the image or region provided.

This method is implemented in the *OpenCV* library through the method

`cv2.goodFeaturesToTrack`, which offers a set of tunable parameters to extract corners from a given image:

`maxCorners` maximum number of corners to be found.

`qualityLevel` multiplicative factor for the R of the best corner. A corner response below qualityLevel $\cdot R_{max}$ will be discarded.

`minDistance` minimum euclidean distance between the selected corners.

`blockSize` size of the pixel block to compute the eigenvalues.

The combination of these two methods provides a fast methodology to estimate the movement of a region using exclusively algebraic calculations on the pixel intensities. As these computations are bounded in complexity, the iteration time is around 5x faster than the neural pipeline. Thus, the simultaneous combination of both algorithms allows to track the movements of the persons during k frames, until the next neural update arrives. This is shown in Figure 3.14.

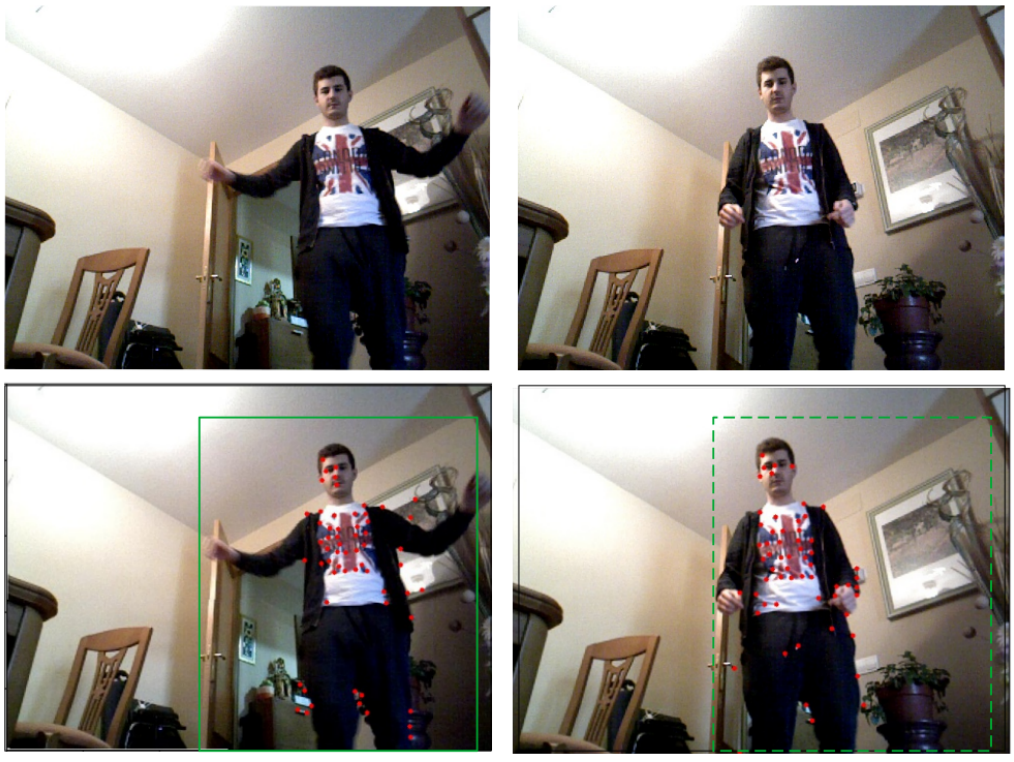


Fig. 3.14. Operation of the tracking module: the last detection (green) determines the person position. The keypoints (red) are tracked during k frames until the next neural update.

As the *OpenCV* implementation of Lucas-Kanade identifies the points that have been found in both frames, the average displacement of all the points can be computed. This allows to shift the bounding box of that person using the computed displacement vector.

This is required, as the bounding box changes its position and size when the person moves, as Figure 3.14 shows. Additionally, it can be rescaled in case the person moves closer or further from the camera, using the distribution of the points in the previous and current frame. As it can be seen on Figure 3.15, the Shi-Tomasi corner detector finds a set of corners (keypoints) in the frame t. These points are distributed with a given mean: the centroid of the cloud, represented with an "x", besides of a standard deviation pair (σ_x^t, σ_y^t). On the next frame, some new keypoints are found (yellow), whereas other keypoints from the previous frame are successfully identified (green). These points are useful for computing the new centroid $(\mu_x^{t+1}, \mu_y^{t+1})$ and deviations pair $(\sigma_x^{t+1}, \sigma_y^{t+1})$. The remaining points from t (red) are not used since they could not be located on $t+1$. With this information, the person box can be updated accordingly:

$$\text{person\_coordinates}(t) = \left[\mu_x^t, \mu_y^t, w, h\right]$$

$$\text{person\_coordinates}(t+1) = \left[\mu_x^{t+1}, \mu_y^{t+1}, w \cdot \frac{\sigma_x^{t+1}}{\sigma_x^t}, h \cdot \frac{\sigma_y^{t+1}}{\sigma_y^t}\right]$$

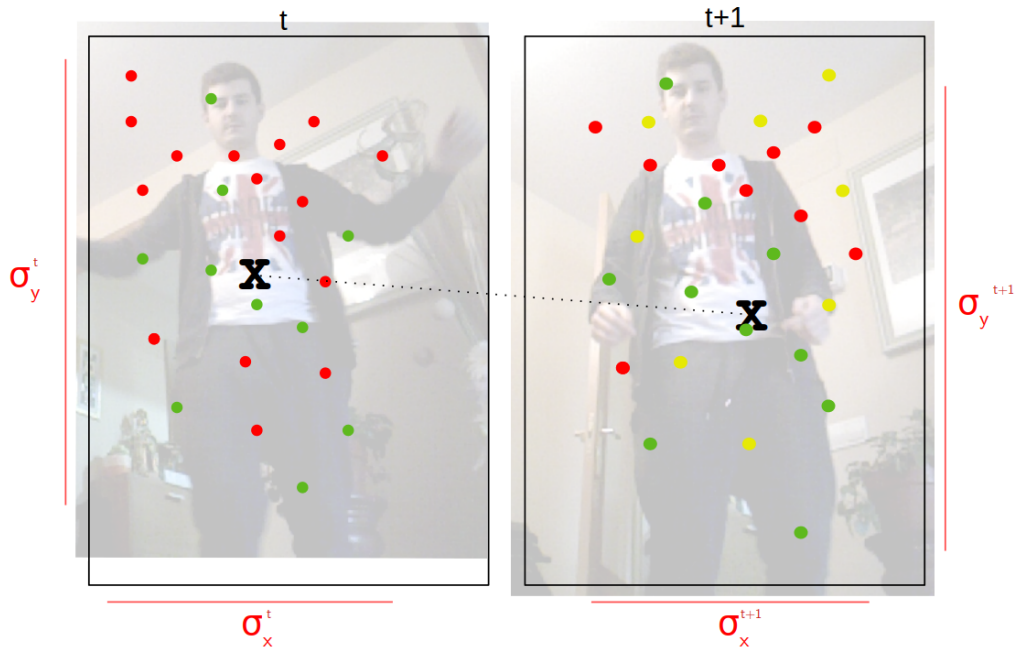


Fig. 3.15. Update of the Lucas-Kanade tracker from frame t to frame $t+1$. The green points are the correctly detected in both frames, while red and yellow points are only detected in t and $t+1$, respectively. The green points determine the new centroid and the size deformation of the box.

This way, the update is sensitive to displacements and scale changes in both directions, in case the person changes their linear distance to the camera.

The incorporation of this Motion Tracker enhances the robustness since the output of the system will not depend only on the neural detections. This improves the performance as partial occlusions might cause some detections to be discarded momentarily. The introduction of the tracker can alleviate this effect, as the person will be kept as *detected* for a number of frames even if it is not detected by the neural pipeline, and its position will be tracked using Lucas-Kanade. This number of frames is called *patience*, P, and introduces a hysteresis in the tracker, as a person has to be lost for P frames in a row to be discarded.

On the same way, a detection has to be maintained during P frames to be joined to the tracked persons. The patience component is introduced in pursuit of stability in complicated scenarios. In such cases a detection flickering is observable, and this could lead to an erratic movement on the robot. The introduction of the patience solves this problem successfully.

PID Controllers

The combination of the described systems results in a efficient way to detect and identify the person to be followed, and additionally, track their movements on a fast way between slower neural detections.

The last block of the system is responsible of translating this location information of the reference person into velocity commands that move the robot towards an *acceptable position* with respect to the person, where certain conditions are fulfilled.

As it was described on Section 3.1, the robot offers 2 degrees of freedom: rotation speed and linear speed. Thus, this *acceptable position* can be described in those 2 dimensions:

Angular position: the reference person has to be placed at a side angle of 0º with respect to the robot front.

Linear position: the reference person has to be placed at a distance of 1 m with respect to the robot front.

Due to the sensors uncertainty, the prediction and tracking estimation, and the movements of the person, these positions have to be extended to *safe areas*, inside of which the robot will not trigger a velocity command for that dimension. This is achieved introducing a *margin/tolerance* on each dimension. Additionally, these geometric criteria have to be translated to measurable discrepancies. This way, the safe zones can be defined as:

Angular zone: the reference person has to be placed at the horizontal center of the image, with a margin of ±50 pixels on the sides.

Linear zone: the reference person has to be placed at a distance of 1 m with respect to the robot front, with a distance margin of ±30 cm[13].

These regions, which are completely tunable using the configuration file, can be visualized on Figure 3.16.

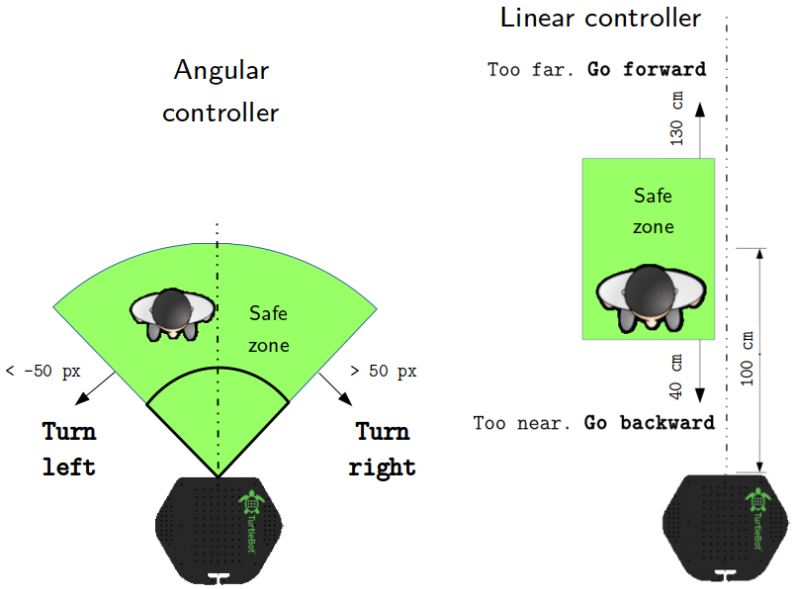


Fig. 3.16. Safe zones for each controller. Image from [11].

To place the person inside these safe zones, the robot has to move on certain directions. For determining a movement, an *error* vector (e_x, e_w) is computed, using the tracked person coordinates:

e_x**:** the linear error or *range* is computed using the depth image, estimating the distance from the robot to the person. As the Xtion sensor registers the depth image into the RGB one, the person coordinates can be used in the depth image in order to find the distance of each pixel inside the bounding box of the reference person: the *person depth map*. As it is feasible that the box contains an important region of the background (specially if the person opens their arms, as the neural detection will encompass the entire body), the edges of the depth map are trimmed. Later, a 10x10 grid is computed to have 100 uniformly distributed samples of the depth of the person. In order to ensure that the background does not affect the range measurement, the median value is computed, as even if some outlier points belong

[13]This criterion can be maintained in metric distance, as the depth sensor specifically yields that information. In the angular case, the image is a 2D projection on the camera plane, which does not allow to infer the relative angle with the person without extra computations using the relative distance.

to the background, they would have to make up the 50% of the sampled set to deviate the measurement from the true range.

e_w: the angular error can be computed taking into account that if the robot and the person are aligned, its bounding box will be horizontally placed near the center of the image. Therefore, an error metric can be extracted computing the difference on the horizontal coordinate between the image center and the center of the bounding box of the reference person.

These computations can be visualized on Figure 3.17.

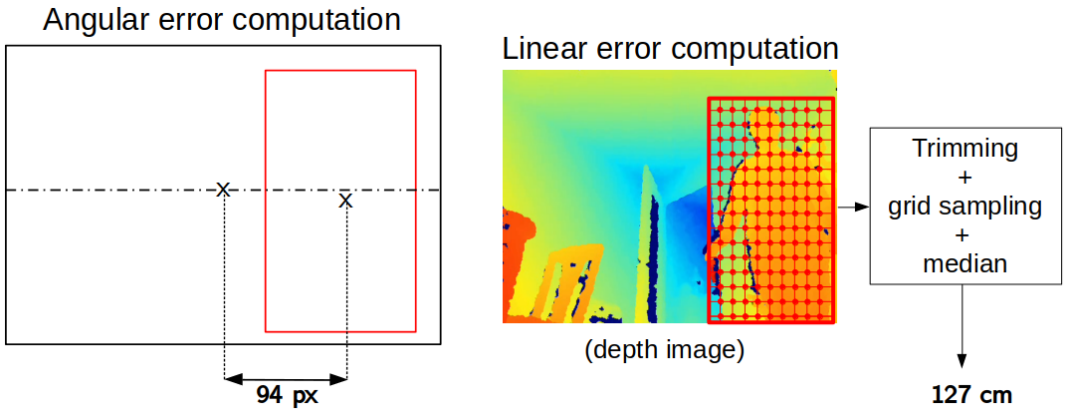


Fig. 3.17. Error computation on each controller.

The last step of the controller takes care of computing two proper responses (linear and angular) for the robot. If these responses depended only on the error readouts, the robot might receive unsteady commands, that might cause a total loss of the person from the field of view. This can be solved introducing a slightly more complex system: a PID controller [52], which is a closed-loop control system that outputs a response taking into account the previously sent responses.

The *PID* acronym stands for *Proportional, Integral and Derivative*, as that is the methodology followed to output a response. The output in the time instant t, $u[t]$ depends on the currently measured error, $e[n]$, and it is computed as it can be seen on Figure 3.18:

This can be expressed by means of the following equation:

$$u[n] = k_p e[n] \quad + \quad k_i \sum_{i=0}^{n} e[i] \quad + \quad k_d(e[n] - e[n-1]) \tag{3.2}$$

This equation can be split into the three components:

***Proportional*:** $k_p e[n]$. This is the basic component, that computes a response directly proportional to the measured error.

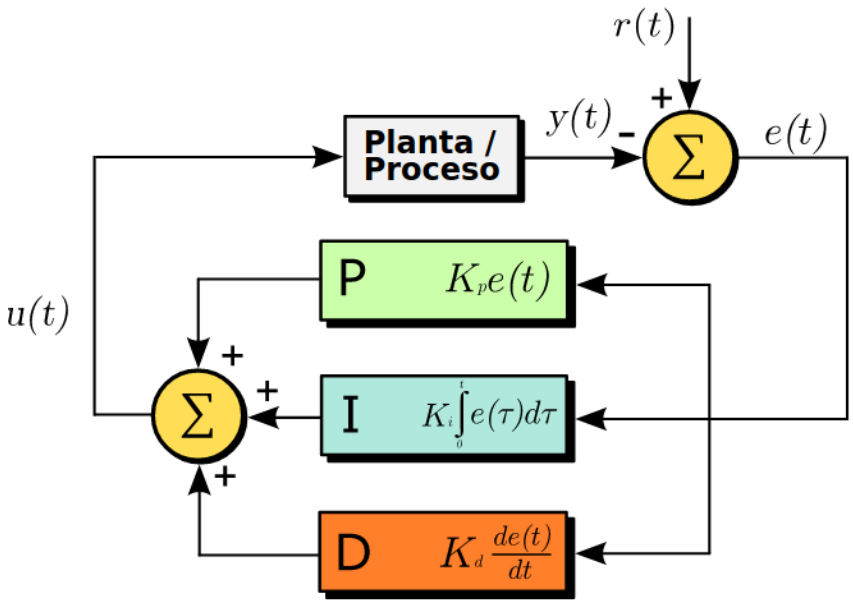


Fig. 3.18. Schematic of a generic PID controller.

***Integral*:** $k_i \sum_{i=0}^{n} e[i]$. An additional response, equivalent to the sum of the total error until the current instant. This way, although a proportional response is not enough and the error gets stabilized in a non-zero value, the system will accumulate that error, increasing the response magnitude in order to close the existing gap between the error and the desired readout[14].

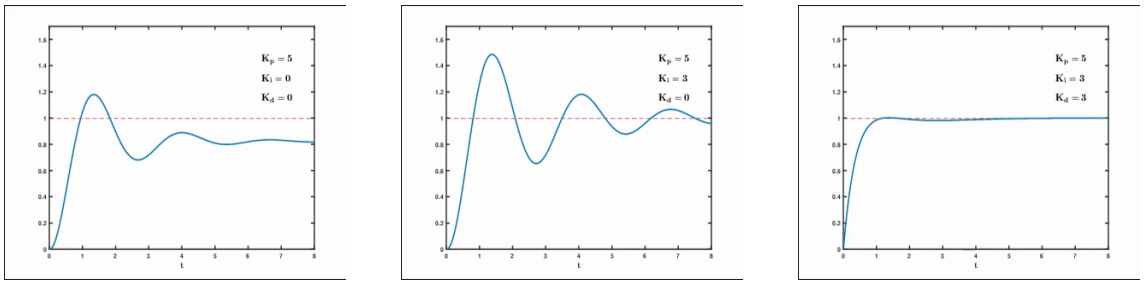
***Derivative*:** $k_d(e[n] - e[n-1])$. This part stands for the *difference* between the last measured error and the current one, and it quantifies how well is the system responding[15]. If the difference is positive, that means that the system is on a further state/position with respect to the last iteration. So, in order to eliminate the *inertia* the system could have acquired (which might bring oscillations and overshooting), the derivative part acts, braking or accelerating the robot depending on the value of the derivative.

Figure 3.19 shows that the combination of the three sub-responses can achieve a fast and steady response (Figure 3.19), bringing back the system under control on an efficient way.

Each contribution is parameterized by its corresponding constant (k_p, k_i, k_d), so a task to perform is to find the optimum value for each one of them. Visual assessments of the robot stability under different combinations lead to the values present in Table 3.1, which yielded a steady behavior of the robot when it is subject to typical indoor conditions of following a wandering person. As for previous parameters, all these values can be

[14]When the monitored variable goes into the tolerated zone again, the total error has to be reset, as it is not required from now on.

[15]On systems without inertia, this contribution is generally ignored, having a simple PI control loop instead.



(a) Proportional. (b) PI. (c) Full PID.

Fig. 3.19. Different controllers response along time.

changed using the configuration file.

	Linear	**Angular**
k_p	0.4	0.005
k_d	0.04	0.0003
k_i	0.05	0.006

Table 3.1. Optimal found values for the parameters in each PID controller.

Finally, when the speed is computed, it is adapted to a ROS `Twist` message, and it is published to the topic devoted to velocity commands to the robot. On the other side of the topic, the driver reads these messages and moves the robot accordingly with the commands received.

This last block completes the design of the full proposed person following behavior.

3.3. Software architecture

The developed software puts all the previous components together, offering two application modes:

`followperson` mode: this is the default mode of the system. When running on this mode, the program feeds the tracker and the neural pipeline with images from the ASUS Xtion, and sends the velocity commands to the robot, writing them into the specified ROS topic.

`benchmark` mode: this mode is designed to test the entire infrastructure, with the purpose of tuning parameters or extracting objective metrics for comparisons, such as precision, or inference time. The images are read from a previously recorded ROSBag, emulating the Xtion sensor and providing always the same RGBD sequence to be fed in different implementations, allowing to compare the

performance of different configurations under identical conditions. On this mode, the velocity commands are not sent to the robot, just drawn in the output image (Figure 3.21), which is also saved into an output video for later visualization. Aside of the video, execution graphs and YAML[16] files are stored containing information about the tracked persons and times for each frame processed by the `Main` thread.

This mode, and other parameters, can be configured on the program execution without modifying the source code. The program receives a YAML configuration file specifying all the required parameters in order to run the system:

```yaml
NodeName: "followperson"
Benchmark: true # true for benchmark, false for followperson
RosbagFile: "resources/bag1.bag"  # path to the ROSBag if benchmark
LogDir: "resources/benchmarks" # where to write the results

Networks:
  # Parameters for the neural pipeline
  Arch: ssd # detection architecture [ssd, yolov3, yolov3tiny]
  DetectionModel: "models/ssd_mobilenet_v1_0.75_depth_coco.pb"
  DetectionWidth: 416 # usually 300 for SSD, 416 for YOLOv3tiny
  DetectionHeight: 416 # usually 300 for SSD, 416 for YOLOv3tiny
  FaceEncoderModel: "models/facenet_inception_resnet_vggface2.pb"

RefFace: "resources/ref_face.jpg" # Image of the reference face

Topics:
  RGB: "/camera/rgb/image_raw" # topic publishing the RGB images
  Depth: "/camera/depth_registered/image_raw" # topic publishing the
      depth images

# Parameters for the speed controllers
XController:
  Kp: 0.4
  Ki: 0.04
  Kd: 0.05
  Min: 0.7
  Max: 1

WController:
  Kp: 0.005
  Ki: 0.0003
  Kd: 0.006
  Min: -50
  Max: 50
# Parameters for the people tracker
PeopleTracker:
```

[16] YAML is a plain-text data serialization format. It has been chosen as a standard format on this project as it offers a good tradeoff between serialization (allowing the data to be converted back into data structures in Python) and readability of the file without processing it.

```
    Patience: 5
    RefSimThr: 1.0
    SamePersonThr: 60
```

The previously depicted structure can be implemented on the Jetson board using the programming language Python. As the tracking module has to run asynchronously, the `threading` library is used, deploying the following threads:

Main thread: the purpose of this thread is to continuously draw the output image (shown in Figure 3.21 and explained below), and compute the errors and suitable responses, as well as sending them to the robot. One thing to notice about this thread is that it does not process all the frames in the sequence, as its rate depends on the drawing time and the computation time of the response. It works asynchronously, fetching the latest frame from the `tracker` thread.

`networks_controller` thread: this controller handles the 3 described neural networks, running sequential inferences on them. In the Jetson platform, these neural networks are deployed in the GPU of the board. Therefore, this thread can be seen as the one which interacts with the GPU in order to pass, retrieve and transform tensors from the networks.

`tracker` thread: as it was shown before, the tracker must inherently iterate at a higher rate than the neural infrastructure. However, including it in the main thread would be bad for its performance, as the speed would be limited by the image drawing and responses publication in the speed topics. Therefore, it is extracted to an specific thread. The simplicity of the Lucas-Kanade tracker makes it fast to execute, however it would be pointless to track a person several times before a new image arrives from the camera. To avoid this, the thread has a rate limitation of 30 Hz, equal to the framerate of the Xtion sensor.

As this is the fastest thread to execute, and it is crucial that the tracker has access to each and every image from the camera, this is the first component to receive the images from the source, on a 30 Hz synchronous manner. The rest of components can fetch the images asynchronously from the tracker whenever they need them.

ROSCam: this component, responsible of fetching the images from the source (a ROSBag or the Xtion camera, as explained before), is not deployed as a thread. However, as it works by means of subscribers when a synchronous mode is required (thus, when the source is the Xtion camera), the ROS API for Python, `rospy` automatically deploys these subscribers on independent threads.

This software architecture can be seen in Figure 3.20, where the interaction between the threads can be visualized. The `Main` thread varies its behavior depending on

the configured mode (`followperson/benchmark`), whereas the rest of threads behave similarly in both configurations.

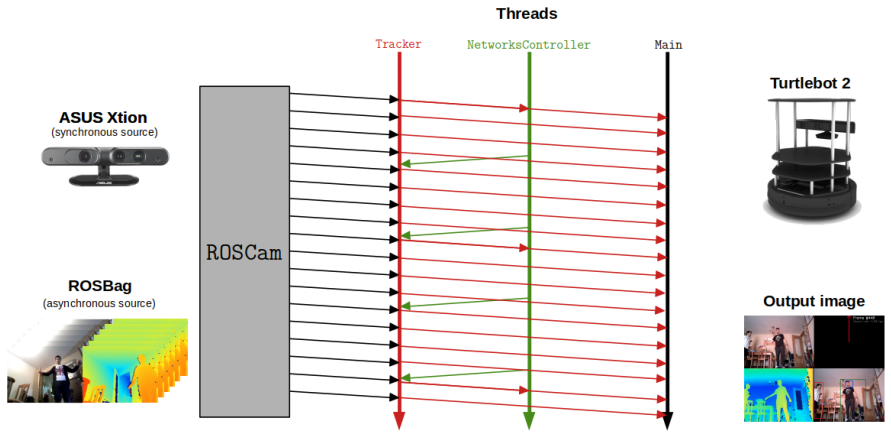


Fig. 3.20. Software architecture for the system.

The visible output of the system is the image shown in Figure 3.21. This image is drawn by the main thread, when the position errors are computed and the responses have been sent to the robot, and it serves for monitoring the execution, showing the images, the tracked persons and the sent commands. If the benchmark mode is enabled, these image are appended to a output video, which serves for posterior visualizations or assessments of the performance.

Fig. 3.21. Output image drawn by the program. Upper left: input RGB image. Bottom left: input depth image. Upper right: velocity commands sent to the robot, and information about the neural rate and number of current frame. Bottom right: tracked persons (green if it is reference, red otherwise) and their faces

4. RESULTS

This chapter describes the different experiments and benchmarks applied to the proposed system and its subsystems. These tests have the purpose of taking design or implementation decisions, selecting the best choices to improve performance, accuracy and robustness of the final system and subsystems. For this purpose, several video sequences were recorded with the ASUS Xtion inside ROSBag files. This way, the same video can be used to assess the performance of different configurations, ensuring that the results will not be affected by external variability due to different environmental conditions on the test data.

The majority of the tests described below for the neural pipeline measure the IoU score (Figure 2.9), which determines the overlapping quality between two bounding boxes. Thus, it is required to label the video sequences, specifying on each frame the location of the ground truth labels for every video. For this purpose, the tool LabelMe [53] was used to provide the labels to the video, creating a JSON file for each frame of the video sequence. A screenshot of this tool is shown on Figure 4.1.

The source code of the experiments conducted below can be found in a separate `experiments` branch of the source repository on GitHub[17], hosting both the testing and plotting source files, as well as the CSV files containing the data plotted in the figures below.

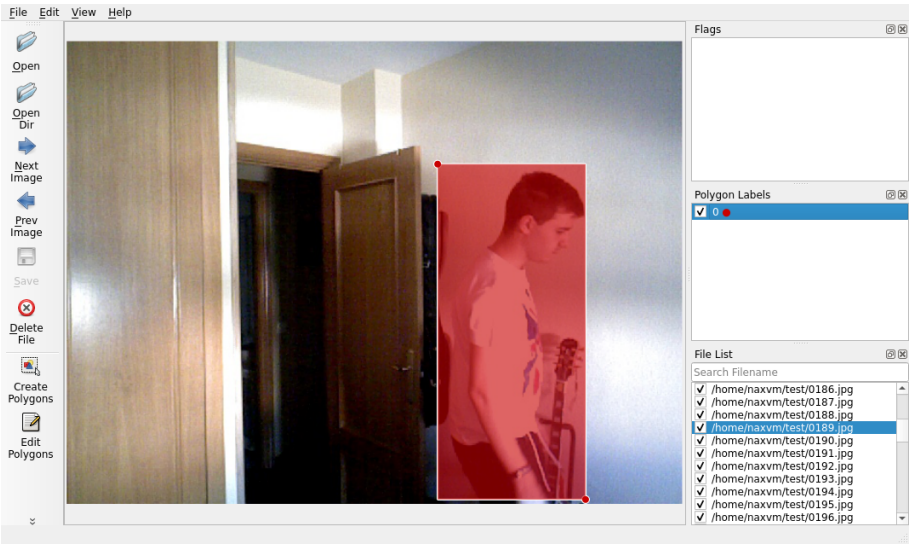


Fig. 4.1. Interface of the LabelMe annotation tool [53].

[17] https://github.com/RoboticsLabURJC/2017-tfg-nacho_condes/tree/experiments

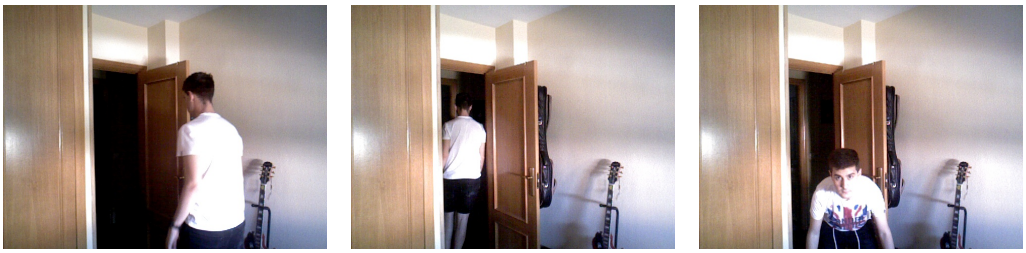
4.1. Person detection experiments

This experiment compares the two detection architectures implemented on this system: YOLO [25] and SSD[19].

In the case of YOLO, the implemented architecture is YOLOv3, in its *tiny* version. This is due to the memory constraints of the Jetson board where the models are loaded. The available memory (8 GB) has to be shared among TensorFlow and the rest of processes, causing the more memory-intensive models to fail on loading. The YOLOv3 full model demands too much memory, making impossible to use it properly on the Jetson TX2 board. Thus, the chosen architecture is a lighter one, publicly available on the YOLO website[18]: the Tiny YOLOv3 model.

On the other hand, as it was explained in Chapter 2, on a real-time application the most convenient variant of the SSD-based detectors is the one that uses a MobileNet [21] as a feature extraction network. The TensorFlow Model Zoo [43] offers several pre-trained models implementing this network, along which a selection has been carried out (as it will be described in other tests). The chosen model integrates a MobileNetv1 whose weights have been quantized [54] in order to reduce the computational cost without reducing the accuracy.

In order to quantify the different accuracy vs. inference time tradeoffs that these architectures offer, a specific test has been designed. A specific video sequence of 721 frames long has been recorded, containing a person wandering across the field of view of the camera. Several extracted frames from this sequence can be observed on Figure 4.2. For every frame of the sequence, the persons are detected using YOLO and SSD respectively, and the IoU and the inference time have been measured, as it can be seen on Figure 4.3. Some gaps can be noticed on the detections, corresponding to the frames where the person was out of the sight of the camera.



(a) Frame 322. (b) Frame 479. (c) Frame 648.

Fig. 4.2. 3 frames from the test video sequence.

The two outstanding object detection architectures have been compared, using both to extract inferences on the same video sequence. The results can be visualized on Figure 4.3 and summarized on Table 4.1. The YOLO-based detector offers a slightly minor IoU than

[18] https://pjreddie.com/darknet/yolo/

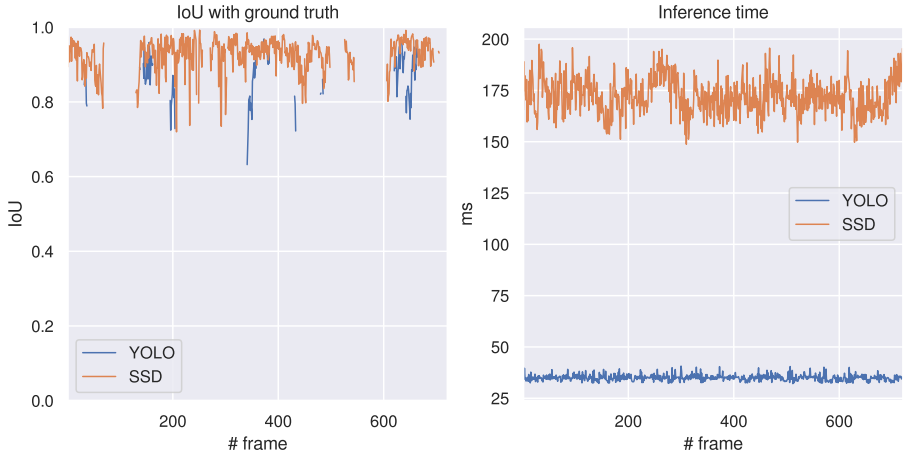


Fig. 4.3. Results of the person detection test: IoU score with ground truth (left) and inference time per frame (right). A discontinuity represents absence of detections.

	YOLO	**SSD**
IoU	0.858 ± 0.068	0.926 ± 0.044
Inf. time (ms)	35.003 ± 1.503	172.237 ± 8.791
Frames with detection	123 (17.06%)	533 (73.93%)

Table 4.1. Numeric summary (average ± standard deviation) for the person detection experiment.

the SSD-based one (around 0.858 and 0.926 respectively), while taking 5 times less time to make inferences (35 ms vs. 172 ms). On these terms, the YOLO-based detector seems much more efficient. However, Table 4.1 shows as well a very unstable detection in the YOLO case, being able to detect the person only in 17% of the frames, whereas SSD detects the person successfully in 74% of the cases. In fact, as there are several frames where the person is not seen, SSD is successful practically in all the cases.
This shows that the YOLO detector is too dependent on pose and lighting conditions for the detections to be successful. On the other hand, the SSD detector yields steady predictions, only cutting on the periods where the person was truly out of the field of view. Hence, this system is much more robust for our application scenario.

One fundamental requirement of the system is the real-time behavior, which makes inference time an important factor to be taken into account. However, as the system includes the described optical tracker, the YOLO detector can be discarded in favor of the SSD-based one, given that the YOLO version has a much lower detection rate[19] and this can not be palliated by the motion tracker.

[19] As it was described before, the implemented version of the YOLO detector is *Tiny YOLOv3*, due to the memory requirements for deploying the full YOLOv3 model, which are higher than what the Jetson TX2 can handle. Thus, it is probable to expect a better performance on the full model in a different computer capable of handling it.

4.2. Face detection experiments

One of the improvements of the proposed system over the previous work [11] is the utilization of a fully neural detection pipeline, as it was described on Chapter 3. This requires the replacement of the face detection Haar cascade classifier explained on Chapter 2 by a neural alternative: *faced*.

This experiment is devoted to compare the performance of both face detection systems. Its design is similar to the previous experiment, using the same video sequence (Figure 4.2) with the ground truth faces labeled using LabelMe. For each frame in the sequence, the faces are extracted using each one of the described methods, and the IoU score is computed with the ground truth face bounding box. The result can be visualized in Figure 4.4.

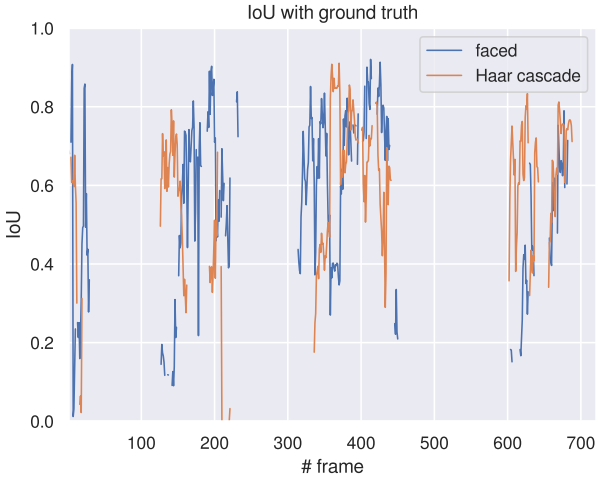


Fig. 4.4. IoU score with the ground truth for each one of the face detection systems.

	haar	**faced**
IoU	0.579 ± 0.202	0.559 ± 0.221
Frames with detection	248 (34.40%)	266 (36.89%)

Table 4.2. Numeric summary (average ± standard deviation) for the face detection experiment.

Figure 4.4 and Table 4.2 show the detection scores for the two mentioned systems on the same video sequence. It can be seen that both yield similar IoU scores and drop at the same time when the person turns their back to the camera. However, the `faced` implementation (which uses deep learning to predict the face positions) is capable of keeping a non-zero IoU at several instants where the Haar performance drops to zero. This

is due to pose variances of the person, as the main drawback of the Haar cascade classifier is that it is only capable of detecting frontal faces, dropping the performance whenever the person turns the face towards a side. This effect is observable in Table 4.2, since both methods yield similar IoU on average, but the deep-learning approach, `faced`, detects a face in 36.89% of the frames, whereas the Haar cascade slightly drops the detection rate to 34.40%.

Hence, this test validates the improvement of the face detection performance when using a specific neural network trained for that purpose.

4.3. Face recognition experiments

The last component of the neural pipeline is a *face recognition* neural network, devoted to confirm the identity of the reference person. This is useful for discerning whether that person has to be followed even if they turns back later, as their position is tracked with the described means. This subsystem is based on a FaceNet [31] network, which projects a face into a 128-dimensional space. These projections are used by the proposed system, as their euclidean distance to the projection of a reference face is used to determine if the input face belongs to the reference person.

This experiment is designed to assess the quality of the projection system, which should yield far points for a different face and near points for a matching face. For this proposal, a video sequence was recorded containing two persons wandering in front of the robot. The faces of each frame are labeled, separating the faces of the two persons in two different classes. A caption of the video with the labels can be seen on Figure 4.5. Figure 4.6 shows several frames from the sequence as well, where some occlusions on the faces can be observed.

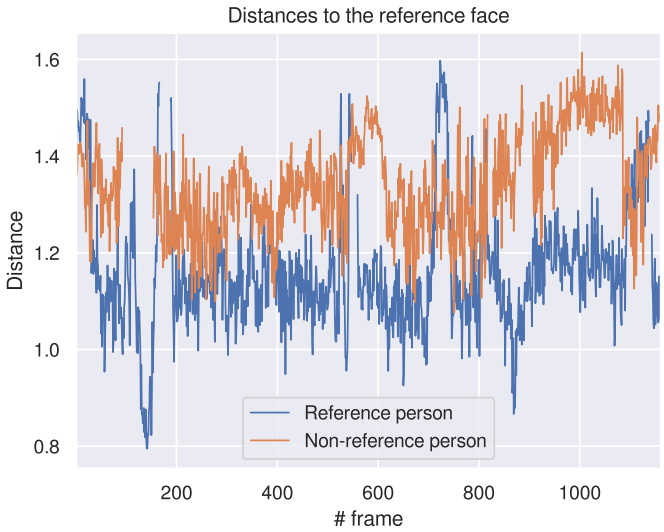
Fig. 4.5. A frame of the test sequence showing the labels on the faces.

(a) Frame 74. (b) Frame 735. (c) Frame 1136.

Fig. 4.6. 3 frames from the test video sequence.

For computing the distance, the reference face was set using the image on Figure 4.7a, and the distance to the reference face of each one of the faces in the video was stored. The result can be observed on Figure 4.7b.

(a) Image of the reference face.



(b) Resulting distance for the two faces present in the video.

Fig. 4.7. Results of the face recognition experiment. (a): reference face used for the test. (b): distance of each face to the reference projection of (a).

	Ref. person	**Non-ref. person**
IoU	1.160 ± 0.128	1.344 ± 0.102

Table 4.3. Numeric summary (average ± standard deviation) for the face recognition experiment.

The results obtained on Figure 4.7 and Table 4.3 allow to extract two conclusions about the quality of the projections of the faces:

- The encodings of the reference person (the person with the same face than the reference one) have an overall remarkable stability. In average, the obtained projections for every frame are located at an approximate distance of 1.16 (threshold

chosen for accepting a person as the reference one). Exceptional rises in the distance can be found as well, but they are due to changes in the pose of the face and occlusions, that reduce the quality of the projection.

- The encodings of a person different than the reference one have an overall higher distance from the reference face. This is convenient for avoiding false positives while determining that a face is the reference one.

This allows to conclude a correct performance of the triplet loss (Figure 2.17) on which a FaceNet is trained [31]. This yields an efficient separation between the encodings of different persons, as well as close encodings for faces belonging to the same person, making this system a robust approach to perform person recognition tasks, since the distance of a projection to the reference face has to be below the threshold for being labeled as the reference face.

4.4. TensorRT experiments

4.4.1. Performance tuning the optimization parameters

In Section 3.2, the TensorRT engine was introduced. This engine is used to optimize, using a binding component between TensorFlow network graphs and TensorRT itself, the implementation of a neural network on a compatible NVIDIA GPU. There are several tunable parameters for customizing the implementation, and the most relevant ones were described in Section 3.2 as well. As varying these parameters changes the model size and the inference time, an experiment has been conducted in order to test the inference time of each model. The optimization script performs a grid search between a set of values for each parameter (MSS, MCE and precision mode, as described on Section 3.2), and tests the performance on a specific ROSBag sequence, storing the detections and the inference times on a YAML file, besides the optimized graph to be loaded without requiring to perform the optimization again.

The inference times for the fastest SSD-based model and the Tiny YOLOv3 implementations are shown below in Table 4.4 and Table 4.5. The impact of this optimization on the precision is studied on Subsection 4.4.2. The performance tables for the rest of models can be found in Chapter 6.

Precision	MSS	MCE	Avg. inference time (ms)
FP32	3	1	59,223
		3	57,139
		5	58,210
	20	1	58,398
		3	58,240
		5	57,910
	50	1	41,077
		3	41,410
		5	41,080
FP16	3	1	57,423
		3	56,777
		5	57,286
	20	1	56,783
		3	56,591
		5	56,637
	50	1	40,053
		3	**39,738**
		5	40,115
INT8	3	1	62,859
		3	61,105
		5	62,383
	20	1	62,439
		3	61,810
		5	63,477
	50	1	46,123
		3	46,835
		5	47,387
GPU without TensorRT			172,269
CPU			112,111

Table 4.4. Grid search results for the `ssd_mobilenet_v1_0.75_depth_coco` model. The lowest inference time is **boldfaced**.

Precision	MSS	MCE	Avg. inference time (ms)
FP32	3	1	20,898
		3	21,032
		5	21,112
	20	1	21,373
		3	21,208
		5	21,639
	50	1	22,506
		3	22,301
		5	22,239
FP16	3	1	16,180
		3	**15,922**
		5	16,061
	20	1	16,200
		3	16,208
		5	16,183
	50	1	18,294
		3	18,110
		5	18,248
INT8	3	1	35,266
		3	36,329
		5	36,289
	20	1	36,305
		3	35,420
		5	35,734
	50	1	35,195
		3	34,815
		5	35,178
GPU without TensorRT			35,996
CPU			NHWC

Table 4.5. Grid search results for the `yolo_v3_tiny` model. The lowest inference time is **boldfaced**. The CPU inferences could not be performed due to hardware incompatibility issues.

4.4.2. Optimized graphs vs. standard graphs

As it has been studied, tuning the TensorRT optimization parameters greatly varies the inference time required for processing an image. However, as it was explained in Section 3.2, this acceleration additionally entails a reduction on the precision, as the weights of the neural network layers are trimmed in the process. The precision mode choice determines the precision of the weights. In the case of the SSD-MobileNet detector, the best inference time (Table 4.4) was yielded by the FP16 precision mode, which trims the weights to a 16-bit long floating point number. This will cause the inference precision to be reduced as the operations are performed on a coarser mode.

This experiment aims to quantify the loss of precision when the SSD model is optimized by TensorRT using the FP16 precision model, which is the fastest mode to infer, as shown in Table 4.4. To do so, the test sequence (Figure 4.2) is used again, passing each frame forward on the standard neural network and storing the detected persons. Later, the same video sequence is passed through the TensorRT version of the same graph, storing the detections of each person as well. When both passes are performed, the IoU score is computed on each frame between the standard inferences (considered as ground-truth labels) and the TensorRT inferences. This IoU score on each frame, along with the inference times for each network model, can be seen on Figure 4.8.

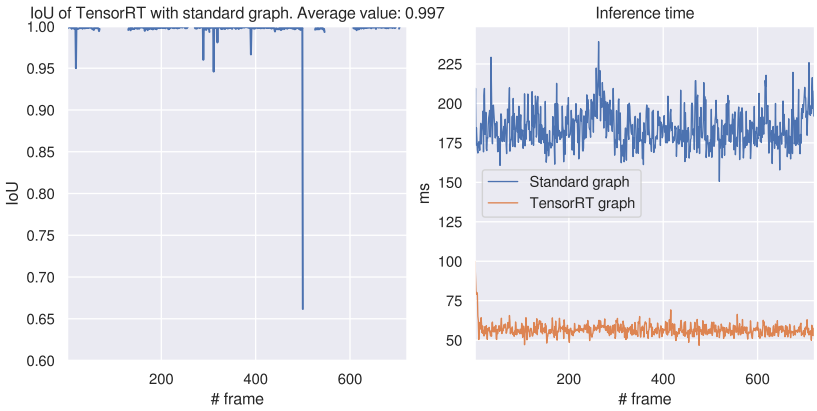


Fig. 4.8. IoU between the standard graph and the TensorRT graph inferences (left) and inference times for both networks (right). The IoU graph has been rescaled between 0.6 and 1 to have a better visualization of the IoU variability.

	Original graph	**TensorRT graph**
Inference time (ms)	184.477 ± 11.827	56.769 ± 4.148

Table 4.6. Numeric summary (average ± standard deviation) for the inference time with and without TensorRT.

Figure 4.8 and Table 4.6 illustrate the differences between a standard graph and an optimized one. One of the premises of the optimization process is the reduction of the

precision of the parameters in the neural network, which can be reduced from 64-bit values up to 16-bit or even 8-bit (performing an additional quantization process), as it was described on Section 3.2.

The loss of precision is clear as well on Figure 4.8, as the IoU of the optimized graph drops at several frames. On the one hand, this loss of precision is small, with some observable exceptions with a loss above 5% of the original performance. On the other hand, the inference time gap can be observed as well. The difference is more notorious, as the TensorRT optimized model performs the inferences 3 times faster than the original graph (56.769 ms vs. 184.477 ms), as Table 4.6 shows.

Given these results, the TensorRT optimizations are a convenient tool to greatly increase the performance of the system, allowing the slower component (the neural pipeline) to experiment an important reduction on the inference time. As a result, the overall performance is greatly improved, receiving reliable neural updates more often.

As it was described on Subsection 4.4.1, a set of parameters can be tuned when optimizing a graph with TensorRT, yielding different performances. As Table 4.4 and Table 4.5 show, important reductions on the inference time can be obtained when the rest of parameters (*Maximum Cached Engines* and *Minimum Segment Size*) are tuned as well. However, these parameters only affect the inference time, as the precision loss is only due to the *Precision Mode* parameter, which has been already analyzed.

The resulting models can be loaded in the program, instead of the original TensorFlow graphs, and offer an overall higher performance, as it has been demonstrated.

4.5. Motion tracker experiments

In Section 3.2, the Lucas-Kanade tracker was described. This tracker aims to follow the movements of the person between two consecutive inferences from the neural pipeline. As in embedded systems these inferences might take a long time, an interpolation of the detections using optical flow can be crucial for avoiding a loss of the location of the person, especially if a partial occlusion of the person causes that the network does not detect them for a while.

This experiment aims to identify the conditions under which a tracker can palliate these drawbacks of the neural detection pipeline, depending on the parameter k. This k modulates the number of elapsed frames between two consecutive neural detections, and takes a higher value if the inferences take longer to be computed by the neural pipeline. On the test, a specific test sequence was recorded and labeled, using a hanging blanket with

the purpose of partially occlude the person, making the network to lose the detections. Several frames of the sequence can be visualized on Figure 4.9.

(a) Frame 285. (b) Frame 788. (c) Frame 864.

Fig. 4.9. 3 frames from the test video sequence.

A correctly tuned tracker keeps the detection active and updates the bounding box for a number of frames (determined by the *patience* parameter, as described in Section 3.2). The video sequence was evaluated using $k = 10$ and $k = 20$, checking the influence of the tracker in the IoU with the ground truth labels of the sequence. The result for both values of k can be observed on Figure 4.10, where the lapse corresponding to the person occlusion has been emphasized.

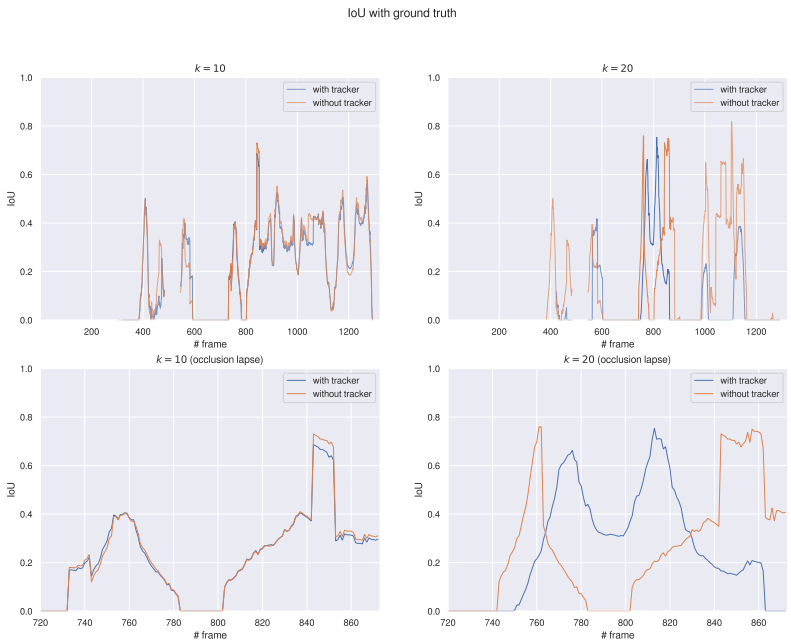


Fig. 4.10. Results of the motion tracker test, for $k = 10$ (left) and $k = 20$ (right). The lapse corresponding to the person occlusion has been emphasized and zoomed in in the bottom graphs.

The results on Figure 4.10 show the IoU score between the persons and the ground truth labels on the test sequence. Regardless of the value of k (the number of frames elapsed between neural detections), a similar performance can be expected under standard

conditions (the faded portion of the graph). However, the emphasized region corresponds to an occlusion behind a hanging blanket (Figure 4.9), and it is zoomed-in on the bottom plots. On this lapse, a better performance is perceptible, especially when the inference time of the neural pipeline is higher ($k = 20$). The hanging blanket occludes the person, causing the neural network to stop detecting them. However, as the tracker retains the detection for several updates because of the patience parameter, the person is not lost until several frames later. Additionally, the Lucas-Kanade algorithm allows to determine the displacement of the person even when it is not being detected by the neural pipeline. This explains the higher IoU when the tracker is active due to the bounding box shifting computed using Lucas-Kanade, confirming the improvement in the performance when using the tracker. Outside this region (top plots), some regions can be detected, such as the ending lapse of the sequence for $k = 20$. This is probably due to non-optimal values for the parameters of the tracker, which do not correctly shift the bounding box towards the true direction of movement of the person. A proper in-depth tuning of the parameters can potentially fix this lower performance for situations similar to that lapse.

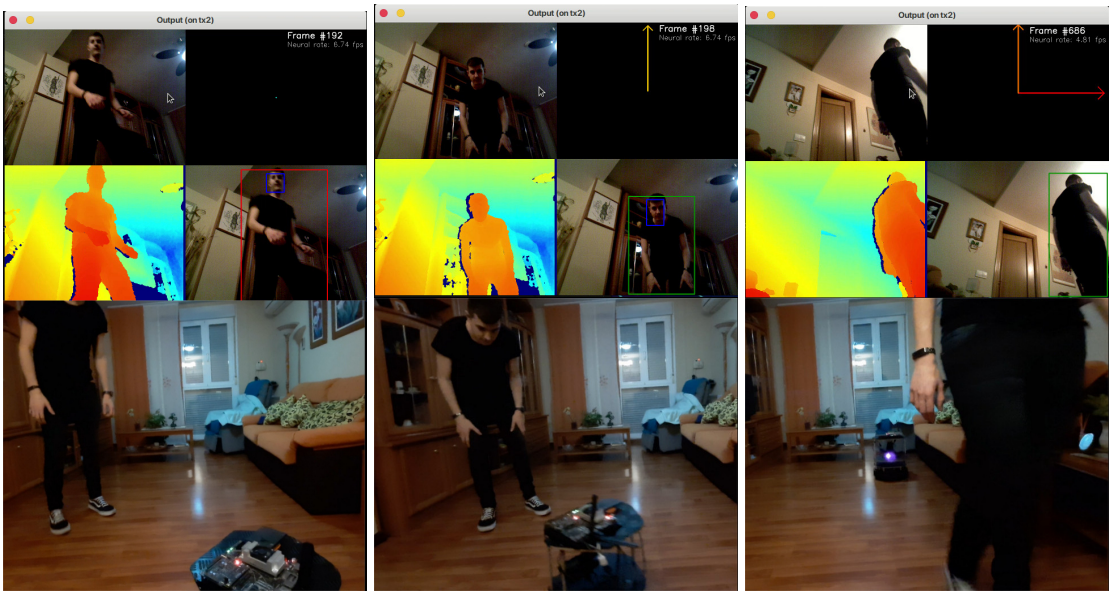
Additionally, while the neural pipeline runs on the GPU of the board, the Lucas-Kanade tracker, whose calculations are much lighter, runs on the CPU. This separation allows to combine both systems asynchronously without affecting the overall load of the system.

4.6. Global system experiments

Finally, a visual assessment can be derived from a sample of the fully functional system[20]. Figure 4.11 shows the behavior of the robot, expected to follow the person properly. The top region of the images shows several screen captures of the program output, with the RGB and depth images (left), and the movement commands and tracked persons (right). The bottom region of the plots show the scene recorded from a mobile phone, allowing to observe the performance externally.

This video presents a sequence of the robot following the typical use case: the reference person enters into the field of view of the robot, showing their face to the camera. After some consecutive frames detecting the person, it is identified using the detected face. When its projection is close enough to the reference one, the robot starts following the person. For each frame, the linear and angular errors are computed, even if the face of the person is not seen anymore, as the system has checked previously that the person has to be followed. If the errors are outside the safe zones, a velocity command is computed and sent to the robot. This routine is executed iteratively until the person gets lost, causing the robot to stop waiting for the person to be seen again.

[20] https://www.youtube.com/watch?v=WZ0riKMwJWA



(a) Person detection.

(b) Person recognition and following.

(c) Following without face feedback.

Fig. 4.11. 3 frames from the full test (available on YouTube, URL on the previous footnote).

5. DISCUSSIONS

On this last chapter, the objectives summarized on Section 1.2 are individually revisited, analyzing the degree of accomplishment of the solution on each one of them. Additionally, suggestions are made for further improvements in future works, that can address the drawbacks of the solution proposed in this dissertation.

5.1. Conclusions

This section revisits the objectives stated in Section 1.2, and reviews the degree of accomplishment of each one of them.

Regarding the first objective, this work has focused on the development and testing of an embedded system that follows a reference person in a robust way, relying on the robustness of deep learning for being capable of working in real environments. This project has been developed using an affordable educational robot and a consumer RGBD sensor.

As the second objective requires, the detection and recognition pipeline has been exclusively designed using deep neural networks, ensuring a robust performance in non-controlled environments. As it has been seen along the project description, this robustness is crucial, especially because the camera is located at a very low position: the lens has an vertical inclination in order to see the full body of the persons in front of the robot. However, this causes as well an excessive amount of light from ceiling lamps to enter into the camera, dimming the persons on the image (Figure 1.4). As it has been tested in Chapter 4, classical systems tend to fail given this issue.

This neural pipeline has been complemented by a tracking component, improving the performance under certain issues, such as partial occlusions, or a higher inference time. This could happen if the networks are more complex or the inference device does not provide a low detection time. This fulfills the third objective of the project.

5.2. Future lines

However, further improvements can be addressed on future works, for example:

- Implement a multimodal tracking using sensor fusion, like in works such as [55]. The depth data of the person also provides information about their position, and bringing this information into the tracker can potentially lead to a better performance.
- Implement a probabilistic tracker, such as an EKF (*Extended Kalman Filter*), relying on the person trajectory. This approach may avoid confusions between two persons if they cross each other, or help the system to follow the trajectory of a person even if it is temporarily lost. In addition, this can solve problems coming from using optical flow, such as a person moving a part of their body. The displacement of the keypoints on that part of the body cause the full bounding box to suffer a displacement even if the person has not changed its position. This can be addressed using probabilistic subsystems to predict the movement of the person.
- Add a navigation component to the robot. The used robot is additionally equipped with a laser scanner, it can be used to detect possible obstacles between the robot and the person. Thus, a simple planning algorithm such as VFF (*Virtual Force Field*) can be combined with this system in order to avoid collisions while the robot is moving.

BIBLIOGRAPHY

[1] *Computer Vision Market to Reach $ 48.6 Billion by 2022*, https://bitrefine.group/11-blog/120-establishing-your-brand-on-college-campuses, Accessed: 2020-06-07.

[2] A. Radford *et al.*, "Language models are unsupervised multitask learners," 2019.

[3] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8599–8603.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, Jan. 2012. DOI: 10.1145/3065386.

[5] P. Martínez-Olmos, "Deep Learning course: Convolutional Neural Networks," University Lecture, 2020.

[6] J. Potel, "Trial by Fire: Teleoperated Robot Targets Chernobyl," *IEEE Computer Graphics and Applications*, 1998. [Online]. Available: https://www.computer.org/csdl/mags/cg/1998/04/mcg1998040010.pdf.

[7] P. Berkelman and J. Ma, "The university of hawaii teleoperated robotic surgery system," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2007, pp. 2565–2566. DOI: 10.1109/IROS.2007.4399550.

[8] A. M. Okamura, "Methods for haptic feedback in teleoperated robot-assisted surgery," *Ind Rob*, vol. 31, no. 6, pp. 499–508, Dec. 2004, 16429611[pmid]. DOI: 10.1108/01439910410566362. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1317565/.

[9] D. Girimonte and D. Izzo, "Artificial intelligence for space applications," in *Intelligent Computing Everywhere*, A. J. Schuster, Ed. London: Springer London, 2007, pp. 235–253. DOI: 10.1007/978-1-84628-943-9_12. [Online]. Available: https://doi.org/10.1007/978-1-84628-943-9_12.

[10] R. Gockley, J. Forlizzi, and R. Simmons, "Natural person-following behavior for social robots," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '07, Arlington, Virginia, USA: Association for Computing Machinery, 2007, pp. 17–24. DOI: 10.1145/1228716.1228720. [Online]. Available: https://doi.org/10.1145/1228716.1228720.

[11] I. Condés and J. Cañas, "Person Following Robot Behavior Using Deep Learning: Proceedings of the 19th International Workshop of Physical Agents (WAF 2018), November 22-23, 2018, Madrid, Spain," in. Jan. 2019, pp. 147–161. DOI: 10.1007/978-3-319-99885-5_11.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," vol. 1, Feb. 2001, pp. I–511. DOI: 10.1109/CVPR.2001.990517.

[13] I. González-Díaz, "Computer Vision: Image classification," University Lecture, 2020.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, 886–893 vol. 1.

[15] I. González-Díaz, "Computer Vision: Local Invariant Features," University Lecture, 2020.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation*, 2013. arXiv: 1311.2524 [cs.CV].

[17] R. Girshick, *Fast R-CNN*, 2015. arXiv: 1504.08083 [cs.CV].

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Lecture Notes in Computer Science*, pp. 346–361, 2014. DOI: 10.1007/978-3-319-10578-9_23. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10578-9_23.

[19] W. Liu *et al.*, "Ssd: Single shot multibox detector," *Lecture Notes in Computer Science*, pp. 21–37, 2016. DOI: 10.1007/978-3-319-46448-0_2. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_2.

[20] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014. arXiv: 1409.1556 [cs.CV].

[21] A. G. Howard *et al.*, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017. arXiv: 1704.04861 [cs.CV].

[22] J. Hosang, R. Benenson, and B. Schiele, *Learning non-maximum suppression*, 2017. arXiv: 1705.02950 [cs.CV].

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, 2015. arXiv: 1506.02640 [cs.CV].

[24] J. Redmon and A. Farhadi, *Yolo9000: Better, faster, stronger*, 2016. arXiv: 1612.08242 [cs.CV].

[25] ——, *YOLOv3: An Incremental Improvement*, 2018. arXiv: 1804.02767 [cs.CV].

[26] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

[27] P. Li, H. Wu, and Q. Chen, "Color distinctiveness feature for person identification without face information," *Procedia Computer Science*, vol. 60, pp. 1809–1816, Dec. 2015. DOI: 10.1016/j.procs.2015.08.291.

[28] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull*, pp. 99–109, 1947.

[29] B. Johnston and P. Chazal, "A review of image-based automatic facial landmark identification techniques," *EURASIP Journal on Image and Video Processing*, vol. 2018, p. 86, Sep. 2018. DOI: 10.1186/s13640-018-0324-4.

[30] R. Gottumukkal and V. Asari, "An improved face recognition technique based on modular pca approach," *Pattern Recognition Letters*, vol. 25, pp. 429–436, Mar. 2004. DOI: 10.1016/j.patrec.2003.11.005.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015. DOI: 10.1109/cvpr.2015.7298682. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298682.

[32] C. Szegedy *et al.*, *Going deeper with convolutions*, 2014. arXiv: 1409.4842 [cs.CV].

[33] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *In NIPS*, MIT Press, 2006.

[34] I. Itzcovich, *faced: CPU Real Time face detection using Deep Learning*, towardsdatascience.com, Ed., [Online; consulted 9-June-2020], Sep. 2018. [Online]. Available: https://towardsdatascience.com/faced-cpu-real-time-face-detection-using-deep-learning-1488681c1602.

[35] J. Vega and J. Cañas, *PiBot: An Open Low-Cost Robotic Platform With Camera for STEM Education*, Oct. 2018. DOI: 10.20944/preprints201810.0372.v1.

[36] M. J. Islam, J. Hong, and J. Sattar, "Person-following by autonomous robots: A categorical overview," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1581–1618, 2019. DOI: 10.1177/0278364919881683. eprint: https://doi.org/10.1177/0278364919881683. [Online]. Available: https://doi.org/10.1177/0278364919881683.

[37] M. J. Islam, M. Fulton, and J. Sattar, *Towards a generic diver-following algorithm: Balancing robustness and efficiency in deep visual detection*, 2018. arXiv: 1809.06849 [cs.RO].

[38] M. Stommel and M. Beetz, "Sampling and clustering of the space of human poses from tracked, skeletonised colour+depth images," Jan. 2013.

[39] I. González-Díaz, "Computer Vision: Image registration," University Lecture, 2020.

[40] Y. Robotics, *Kobuki User Guide*, English, version Version 1.1.0, 24 pp. [Online]. Available: https://docs.google.com/document/d/15k7UBnYY_GPmKzQCjzRGCW-4dIP7zl_R_7tWPLM0zKI/edit, consulted on 2020/06/14.

[41] NVIDIA, *NVIDIA Jetson TX2: datasheet*, English, version Version 1.6, 68 pp. [Online]. Available: https://developer.download.nvidia.com/assets/embedded/secure/jetson/TX2/docs/Jetson_TX2_Series_Module_DataSheet_v1.6.pdf?Q_eTPkb4IeUZi3rN5gB7N0v6ZNPZJwCNZxPvj9Ct8Sc_LlQgmY12RNuTrJ-qovqrtMX6yUoYcSHbAE1mjhZ3FL59_UxlubPypJB7l7doHcbtGLBaGMzSdiT_6TyVOC2H9klPyl0KcEo48G-XtdkdSfBugtRDYMYn1ouZjffwy5NdPfEyyiSe0T5T204iiO2SUQ, consulted on 2020/06/14.

[42] ros.org, *What is ROS?* wiki.ros.org/ROS/Introduction.

[43] TensorFlow, *TensorFlow Object Detection: Model Zoo*, [Online; consulted 28-June-2020]. [Online]. Available: https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md.

[44] I. González-Díaz, "Computer Vision: Dense Motion Estimation," University Lecture, 2020.

[45] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI)," vol. 81, Apr. 1981.

[46] OpenCV, *OpenCV: Optical Flow*, [Online; consulted 22-June-2020]. [Online]. Available: https://docs.opencv.org/master/db/d7f/tutorial_js_lucas_kanade.html.

[47] Jianbo Shi and Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[48] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, 1150–1157 vol.2.

[49] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, 2005, 1508–1515 Vol. 2.

[50] S. El-mashad and A. Shoukry, "Evaluating the robustness of feature correspondence using different feature extractors," Sep. 2014. DOI: 10.1109/MMAR.2014.6957371.

[51] NanoNets, *Introduction to Motion Estimation with Optical Flow*, [Online; consulted 22-June-2020]. [Online]. Available: https://nanonets.com/blog/optical-flow/.

[52] K. J. Åström and R. M. Murray, "Feedback systems: An introduction for scientists and engineers," Tech. Rep., 2004.

[53] K. Wada, *labelme: Image Polygonal Annotation with Python*, https://github.com/wkentaro/labelme, 2016.

[54] G. A. Blog, *Accelerating Training and Inference with the Tensorflow Object Detection API*, [Online; consulted 28-June-2020]. [Online]. Available: https://ai.googleblog.com/2018/07/accelerated-training-and-inference-with.html.

[55] T. Ophoff, K. Van Beeck, and T. Goedemé, "Exploring rgb+depth fusion for real-time object detection," *Sensors*, vol. 19, p. 866, Feb. 2019. DOI: 10.3390/s19040866.

6. ANNEXES

6.1. Optimization results for all the models

This annex shows the optimization results using TensorRT for all the neural network models used and tested on this work. The inference time results have been stored in several spreadsheets, classified according to the optimization parameters (described in Section 3.3), and a color scale has been added according to the FPS (*Frames per Second*) the resulting graph is capable to process. These numbers are compared as well to the original implementation (GPU without using TensorRT) and to the CPU implementation. The results can be observed in the following subsections.

6.1.1. Object detection models

The optimized models correspond to the different implementations of SSD-based and YOLO-based object detectors, as described in Chapter 2.

6.1.2. Face detection models

These models are the specifically trained for the `faced` library [34]. They have been optimized as well, swapping the originally included models in the package for the TensorRT optimized ones.

6.1.3. Face encoding model

This is the FaceNet implementation [31], which has been optimized as well using TensorRT:

Model	Precision	MSS	MCE	Mean inf. time	Avg. FPS
ssd_mobilenet_v1_coco	FP32	3	1	59,473	16,814
			3	59,405	16,834
			5	59,894	16,696
		20	1	58,791	17,010
			3	60,542	16,518
			5	58,944	16,965
		50	1	45,643	21,909
			3	45,730	21,867
			5	45,458	21,998
	FP16	3	1	56,802	17,605
			3	56,895	17,576
			5	57,810	17,298
		20	1	57,616	17,356
			3	56,580	17,674
			5	58,027	17,233
		50	1	43,509	22,984
			3	43,993	22,731
			5	44,194	22,628
	INT8	3	1	66,161	15,115
			3	65,682	15,225
			5	66,154	15,116
		20	1	63,711	15,696
			3	64,674	15,462
			5	64,890	15,411
		50	1	53,062	18,846
			3	52,652	18,993
			5	52,105	19,192
	GPU without TensorRT			163,633	6,111
	CPU			152,126	6,573

(a) Optimization results for the object detection model `ssd_mobilenet_v1_coco`.

Model	Precision	MSS	MCE	Mean inf. time	Avg. FPS
ssd_mobilenet_v2_coco	FP32	3	1	67,9234	14,722
			3	67,9204	14,723
			5	67,208	14,879
		20	1	68,2606	14,650
			3	67,8827	14,731
			5	67,6093	14,791
		50	1	52,095	19,196
			3	52,0465	19,214
			5	51,9552	19,247
	FP16	3	1	64,6886	15,459
			3	64,2897	15,555
			5	64,7186	15,452
		20	1	64,5814	15,484
			3	64,6789	15,461
			5	63,6143	15,720
		50	1	48,5507	20,597
			3	48,4325	20,647
			5	48,3583	20,679
	INT8	3	1	116,2488	8,602
			3	114,8495	8,707
			5	116,6604	8,572
		20	1	115,1101	8,687
			3	115,4840	8,659
			5	115,0603	8,691
		50	1	101,5242	9,850
			3	101,3347	9,868
			5	100,6825	9,932
	GPU without TensorRT			207,413	4,821
	CPU			194,621	5,138

(b) Optimization results for the object detection model `ssd_mobilenet_v2_coco`.

Model	Precision	MSS	MCE	Mean inf. time	Avg. FPS
ssd_mobilenet_v1_0.75_depth_coco	FP32	3	1	59,223	16,885
			3	57,139	17,501
			5	58,210	17,179
		20	1	58,398	17,124
			3	58,240	17,170
			5	57,910	17,268
		50	1	41,077	24,344
			3	41,410	24,149
			5	41,080	24,343
	FP16	3	1	57,423	17,415
			3	56,777	17,613
			5	57,286	17,456
		20	1	56,783	17,611
			3	56,591	17,671
			5	56,637	17,656
		50	1	40,053	24,967
			3	39,738	25,165
			5	40,115	24,929
	INT8	3	1	62,859	15,909
			3	61,105	16,365
			5	62,383	16,030
		20	1	62,439	16,016
			3	61,810	16,179
			5	63,477	15,754
		50	1	46,123	21,681
			3	46,835	21,351
			5	47,387	21,103
	GPU without TensorRT			172,269	5,805
	CPU			112,111	8,920

(c) Optimization results for the object detection model `ssd_mobilenet_v1_0.75_depth_coco`.

Model	Precision	MSS	MCE	Mean inf. time	Avg. FPS
ssdlite_mobilenet_v2_coco	FP32	3	1	64,864	15,417
			3	63,414	15,770
			5	63,317	15,794
		20	1	64,936	15,400
			3	64,117	15,597
			5	63,980	15,630
		50	1	49,617	20,154
			3	49,107	20,364
			5	49,804	20,079
	FP16	3	1	63,408	15,771
			3	65,282	15,318
			5	64,823	15,427
		20	1	65,206	15,336
			3	63,855	15,661
			5	64,432	15,520
		50	1	49,570	20,174
			3	49,444	20,225
			5	49,243	20,308
	INT8	3	1	77,383	12,923
			3	76,993	12,988
			5	75,952	13,166
		20	1	77,520	12,900
			3	73,522	13,601
			5	76,221	13,120
		50	1	59,531	16,798
			3	59,681	16,756
			5	59,999	16,667
	GPU without TensorRT			164,628	6,074
	CPU			194,621	5,138

(d) Optimization results for the object detection model `ssdlite_mobilenet_v2_coco`.

Fig. 6.1. Optimization results for the SSD-based object detection networks.

Model	Precision	MSS	MCE	Mean inf. time	Avg. FPS
yolov3_tiny	FP32	3	1	20,898	47,853
			3	21,032	47,546
			5	21,112	47,367
		20	1	21,373	46,788
			3	21,208	47,151
			5	21,639	46,213
		50	1	22,506	44,432
			3	22,301	44,841
			5	22,239	44,966
	FP16	3	1	16,180	61,805
			3	15,922	62,808
			5	16,061	62,264
		20	1	16,200	61,730
			3	16,208	61,698
			5	16,183	61,794
		50	1	18,294	54,662
			3	18,110	55,220
			5	18,248	54,799
	INT8	3	1	35,266	28,356
			3	36,329	27,526
			5	36,289	27,556
		20	1	36,305	27,544
			3	35,420	28,232
			5	35,734	27,985
		50	1	35,195	28,413
			3	34,815	28,723
			5	35,178	28,427
	GPU without TensorRT			35,996	27,781
	CPU			NHWC	N/A

Fig. 6.2. Optimization results for the object detection model `yolov3_tiny` (due to hardware compatibility issues, the CPU testing was impossible to perform).

Model	Precision	MSS	MCE	Mean inf. time	Avg, FPS
face_yolo	FP32	3	1	22,889	43,689
			3	22,743	43,970
			5	23,004	43,470
		20	1	49,985	20,006
			3	50,316	19,874
			5	50,671	19,735
		50	1	50,013	19,995
			3	49,767	20,094
			5	50,305	19,879
	FP16	3	1	17,351	57,633
			3	17,257	57,948
			5	17,278	57,878
		20	1	50,108	19,957
			3	50,348	19,862
			5	50,385	19,847
		50	1	50,269	19,893
			3	50,757	19,702
			5	50,324	19,871
	INT8	3	1	50,350	19,861
			3	50,370	19,853
			5	50,116	19,954
		20	1	49,668	20,134
			3	49,823	20,071
			5	50,479	19,810
		50	1	49,682	20,128
			3	49,964	20,014
			5	50,002	19,999
	GPU without TensorRT			50,609	19,759
	CPU			132,850	7,527

(a) Optimization results for the face detector model `face_yolo`.

Model	Precision	MSS	MCE	Mean inf. time	Avg, FPS
face_corrector	FP32	3	1	5,766	173,436
			3	5,590	178,888
			5	5,605	178,403
		20	1	6,000	166,678
			3	5,693	175,670
			5	5,830	171,541
		50	1	5,916	169,042
			3	5,687	175,846
			5	5,855	170,783
	FP16	3	1	5,604	178,460
			3	5,752	173,868
			5	5,603	178,466
		20	1	5,689	175,781
			3	5,620	177,945
			5	5,898	169,563
		50	1	5,795	172,577
			3	6,048	165,355
			5	5,778	173,061
	INT8	3	1	5,974	167,384
			3	5,805	172,268
			5	5,800	172,411
		20	1	5,633	177,513
			3	5,854	170,835
			5	5,974	167,403
		50	1	5,931	168,611
			3	5,846	171,045
			5	5,621	177,904
	GPU without TensorRT			7,863	127,171
	CPU			8,675	115,268

(b) Optimization results for the face corrector mdoel `face_corrector`.

Fig. 6.3. Optimization results for the face detection (`faced`) networks.

Model	Precision	MSS	MCE	Mean inf. time	Avg. FPS
facenet	FP32	3	1	40,602	24,629
			3	39,490	25,323
			5	40,163	24,899
		20	1	65,552	15,255
			3	66,399	15,061
			5	66,348	15,072
		50	1	63,498	15,749
			3	65,299	15,314
			5	66,123	15,123
	FP16	3	1	44,575	22,434
			3	44,464	22,490
			5	44,889	22,277
		20	1	66,842	14,961
			3	65,593	15,246
			5	65,840	15,188
		50	1	62,697	15,950
			3	66,012	15,149
			5	65,536	15,259
	INT8	3	1	64,552	15,491
			3	NaN	
			5	NaN	
		20	1	66,029	15,145
			3	64,852	15,420
			5	65,507	15,266
		50	1	63,398	15,773
			3	65,662	15,229
			5	64,770	15,439
	GPU without TensorRT			59,562	16,789
	CPU			141,963	7,044

Fig. 6.4. Optimization results for the face encoding model `facenet`.