Chapter

Deep Learning-Based Detection of Pipes in Industrial Environments

Edmundo Guerra, Jordi Palacin, Zhuping Wang and Antoni Grau

Abstract

Robust perception is generally produced through complex multimodal perception pipelines, but these kinds of methods are unsuitable for autonomous UAV deployment, given the restriction found on the platforms. This chapter describes developments and experimental results produced to develop new deep learning (DL) solutions for industrial perception problems. An earlier solution combining camera, LiDAR, GPS, and IMU sensors to produce high rate, accurate, robust detection, and positioning of pipes in industrial environments is to be replaced by a single camera computationally lightweight convolutional neural network (CNN) perception technique. In order to develop DL solutions, large image datasets with ground truth labels are required, so the previous multimodal technique is modified to be used to capture and label datasets. The labeling method developed automatically computes the labels when possible for the images captured with the UAV platform. To validate the automated dataset generator, a dataset is produced and used to train a lightweight AlexNet-based full convolutional network (FCN). To produce a comparison point, a weakened version of the multimodal approach—without using prior data—is evaluated with the same DL-based metrics.

Keywords: deep learning, autonomous robotics, UAV, multimodal perception, computer vision

1. Introduction

Robotics, as a commercial technology, started to be widespread some decades ago, but instead of decreasing, it has been growing year by year with new contributions in all the related fields that it integrates. The introduction of new materials, sensors, actuators, software, communications and use scenarios converted Robotics in a pushing area that embraces our everyday life. New robotic morphologies are the most shocking aspect that society perceives (i.e., the first models of each type generally produce the largest impact), but the long-term success of robotics is found in its capability to automate productive processes. Manufacturers and developers know that the market is found not only in large-scale companies (car manufacturers and electronics mainly) but also in the SME that provides solutions to problems that are manually performed so far. Also, robotics has opened the doors to new applications that did not exist some years ago and are also attractive to investors. These facts, together with lower prices for equipment, better programming and communication tools, and new fast-growing user-friendly collaborative robotic frameworks, have pushed robotics technology at the edge in many areas.

It is clear that industrial robotics leads the market worldwide, but social/gaming uses of robots have increased sales. Nevertheless, the most promising scenario for the present time and short term is the use of robots in commercial applications out of the plant floor. Emergency systems, inspection, and maintenance of facilities of any kind, rescues, surveillance, agriculture, fishing, border patrolling, and many other applications (without military use) attract users/clients because their use increases the productivity of the different sectors, low prices and high profitability are the keys.

There exist many robot morphologies and types (surface, underwater, aerial, underground, legged, wheels, caterpillar, etc.) but authors want to draw attention in the unmanned aerial vehicles (UAVs), which have several properties that make them attractive for a set of application that cannot be done with any other type of robot. First, those autonomous robots can fly, and therefore, they can reach areas that humans or other robots cannot. They are light, easy to move from one area to another, and can be adapted to any area, terrain, soil, building, or facility. The drawback is the fragility in front of adverse meteorological events, and their autonomy is quite limited compared with unmanned surface vehicles (USVs).

UAVs have seen the birth of a new era of unthinkable cheap, easy applications up to now. The authors would like to focus its use in the maintenance and inspection of industrial facilities, but specifically in the inspection of pipes in big, complex factories (mainly gas and oil companies) where the manual inspection (and even location and mapping) of pipes becomes an impossible task. Manned helicopters (with thermal engines) cannot fly close to pipes or even among a bunch of pipes. Scaffolds cannot be put up in complex, unstable, and fragile pipes to manually inspect them. Therefore, a complex problem can be solved through the use of UAVs for inspecting pipes of different diameters, colors, textures, and conditions in hazardous factories. This problem is not new and some solutions have been brought to an incipient market. Works as those in [1, 2] propose the creation of a map of the pipe set navigating among it with odometry and inertial units [3]. Obstacle avoidance in a crowded 3D world of pipes becomes of great interest when planning a flight; in [4], some contributions are made in this direction although the accuracy of object is deficient to be a reliable technology. Work in [5] overcomes some of the latter problems with the use of a big range of sensors, cameras, laser, barometer, ultrasound, and a computationally inefficient software scheme made the UAV too heavy and unreliable due to the excessive sensor fusion approach.

Many of the technical developments that have helped robotics grow have had a wider impact, especially those related with increasing computational power and parallelization levels. Faster processors, with tens of cores and additional multiple threat capabilities, and modern GPUs (graphics processing unit) have led to the emergence of GPGPU (general-purpose computing on GPU). These type of computing techniques have led to huge advances in the artificial intelligence (AI) field, producing the emergence of the "deep learning" field. The deep learning (DL) field is focused in using artificial neural networks (ANNs) that present tens or hundreds of layers, exploiting the huge parallelization capabilities of modern GPU. This is used in exploiting computational cores (e.g., CUDA cores), which compared on a one-to-one basis with a processor core, they are less powerful and slower, but can be found in amounts of hundreds or thousands. This has allowed the transition from shallow ANN to the deeper architectures and innovations such as several types of convolutional layers. In this work, the authors present a novel approach to detect pipes in industrial environments based in fully convolutional networks (FCNs). These will be used to extract the apparent contour of the pipes, replacing most of the architecture developed in [6] and discussed in Section 2. To properly train these networks, a custom dataset relevant to the domain is required, so the authors

captured a dataset and developed an automatic label generation procedure base in previous works. Two different state-of-the-art semantic segmentation approaches were trained and evaluated with the standard metrics to prove the validity of the whole approach. Thus, in the following section, some generalities about the pipe detection and positioning problem are discussed, and the authors' previous work [6] on it, as it will be relevant later. The next section discusses the semantic segmentation problem as a way to extract the apparent contour, both surveying classical methods, considered for earlier works, and state of the art deep-learning-based methodologies. The fourth section describes how the automatic label generator using multimodal data was derived and some features to the process. The experimental section starts discussing the metrics employed to validate the results, the particularities of the domain dataset generated and describes how an AlexNet FCN architecture was trained through transfer learning and the results achieved. To conclude, some discussion on the quality of the results and possible enhancements is introduced, discussing which would be the best strategies to follow continuing this research.

2. Related work

As it has been discussed, inspection and surveying are a frequent problem where UAV technologies are applied. The most common scenario found is that of a hard to reach infrastructure that is visually inspected through different sensors onboard a piloted UAV. Some projects have proposed the introduction of higher level perception and automation capacities, depending on the specific problem. In these cases, it is common to join state-of-the-art academic and industrial expertise to reach functional solutions.

In one of these projects, the specific challenge of accurately detecting and positioning a pipe in real time using only the hardware deployable in a small (per industry standards) UAV platform was considered (**Figure 1**), with several solutions studied and tested (including vision- and LIDAR-based techniques).

In the case of LIDAR-based detection, finding a pipe is generally treated as a segmentation problem in the sensor space (using R3 data collected as "*point clouds*"). There are many methods used for LIDAR detection, but the most successful are based on stochastic model fitting and registration, commonly in RANSAC (Random Sample Consensus [7]) or derived approaches [8, 9]. Three different data density levels were tested using the libraries available through ROS: using RANSAC over a map estimated by a SLAM technique, namely LOAM [10]; detecting the pipe

Figure 1.
One of the UAV used for the development of perception tasks in the AEROARMS project. Several sensors were deployed, processing them with a set of SBCs (single-board computers), including a Velodyne LiDAR, two different cameras, ultrasonic range-finder (height), and optical flow.

in a small window of consecutive point clouds joined by an ICP-like approach [11]; and finally to simply work using the most recent point cloud. The first approach probed to be computationally unfeasible, no matter what optimization was tested, as even working with a single datum cloud point could be prohibitive if not done carefully. To enhance the performance, the single cloud point approach was optimized employing spatial and stochastic filtering to reduce the data magnitude, and a curvature filter allowed to reduce fake positives in degenerate configurations, producing robust results at between 1 and 4 Hz. To solve the same problem with visual sensors, a two-step strategy was used. In order to estimate the pose of the pipes to be found, they were assumed to be circular and regular enough to be modeled as a straight homogeneous circular cylinder. This allowed using a closed-form conic equation [12], which related the axis of the pipe (its position and orientation as denoted in Plücker coordinates) with the edges of its projection in the image space. While this solves the positioning problem, the detection probed to be a little more challenging: techniques based on edge detection, segmentation, or other classical computer vision methods used to work under controlled light but failed to perform acceptably in outdoor scenarios. This issue was solved by introducing human supervision, where an initial seed for the pipe in the image sensor space was provided (or validated) by a human and then tracked robustly through vision predicting it with the UAV odometry.

With these results, discussed in [6], it was apparent that a new solution was needed, as the LiDAR approaches were too slow and the vision-based techniques probed themselves unreliable. The final proposed solution was based on integrating data from the laser and the vision sensors: the RANSAC over LiDAR approach would detect robustly the pipe and provide an initial position, which would then be projected into the image space (accounting for displacements if odometry is available) and used as a seed for the vision-based pipeline described.

In that same work [6], a sensibility analysis studying the effects of the relative pose between the sensor and pipes is provided. Once the pipe is detected in the LiDAR's space sensor, the cylinder model is projected into the R^2 image space using a projection matrix derived from the calibrated camera model (assumed to be a thin lens pinhole model, per classic literature [13]). This provides a region or band of interest where to look for the edges of the pipe in the image and is useful to solve the degenerate conic equation up to scale (i.e., being a function of the radius). An updated architecture version of the process is depicted in **Figure 2**.

The detailed architecture of the multimodal approach reveals how the LiDAR-based pipeline minimizes the data dimensionality by filtering non-curved surfaces (i.e., remove walls, floor, etc.) and also by removing entirely regions of the sensed space if priors or relevant data or the expected relative position of the pipe to the sensor is available. This was aimed at minimizing the size of the point cloud to be processed by the RANSAC step. To be able to project the detected pipe from the LiDAR sensor space into the camera image, some additional information was required: the rigid transformation between sensors (i.e., the calibration between LiDAR and camera) and an estimation of the odometry of the UAV. This is due because, even in the best assumption, with a performance slightly over 4 Hz, the delay between the captured point cloud and the produced estimation of the pipe would be over 200 ms. Therefore, the projection of the detected pipe to predict the area of interest to search the apparent contour has to consider the displacement during this period, not only the rigid LiDAR to camera transformation. This predicted region of interest is used in the vision process pipeline, with predictions of the appearance of the pipes into image space used to refine the contour search. This contour search relies on stacking a Hough transform to join line segment detector (LSD) detected segments (to overcome partial obstructions) on the relevant area

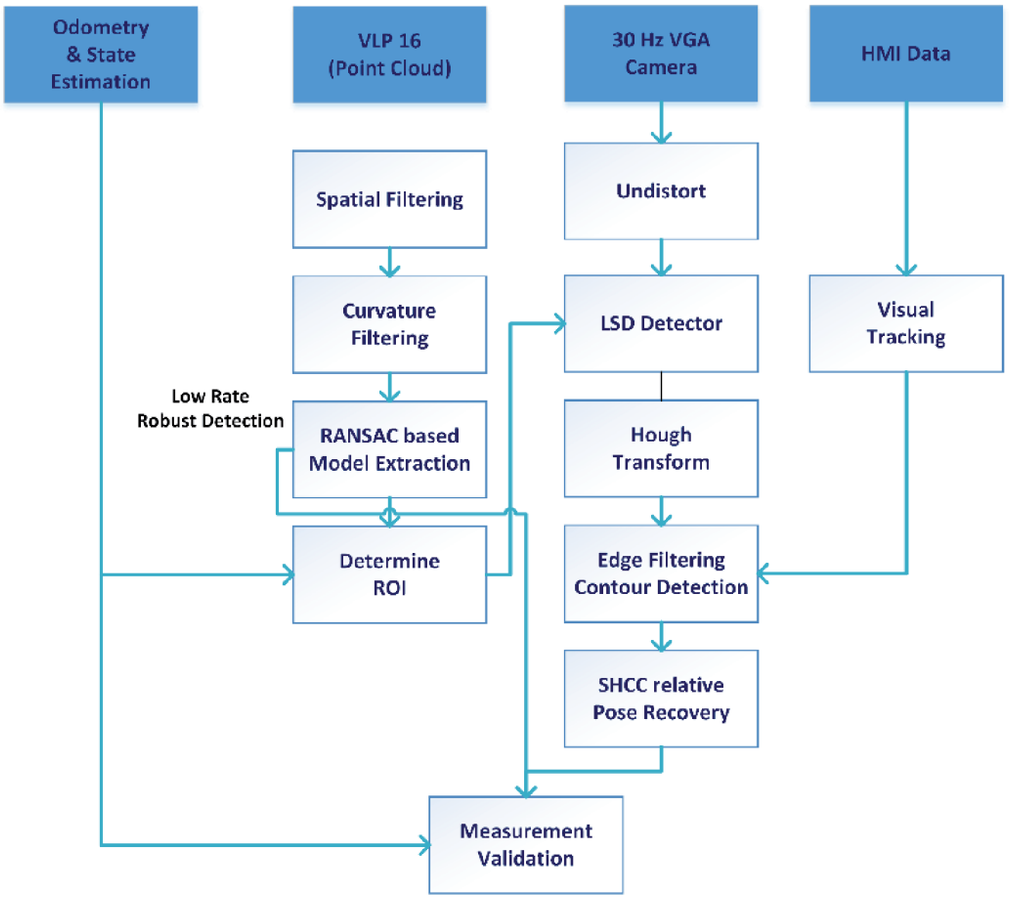

Figure 2.
The architecture of the multimodal perception pipeline combining LiDAR and camera vision. An updated version adds to previous works a validation step using odometric measurements.

and allows to choose the nearest correctly aligned lines. Notice that using a visual servoing library [14], an option to use data provided through human interaction was kept as available, though the integration of LiDAR detections as seeds into the visual pipeline made it unnecessary. To avoid degenerate or spurious solutions, a validation step (based on reprojection and "matching" of the Plückerian coordinates [15] for a tracked piped) was later introduced.

This architecture leads to a fast (limited by the performance of the vision-based part) and robust (based on the RANSAC resilience to spurious detections) pipe detector with great accuracy, which was deployed and test in a UAV. The main issue of the approach is the hardware requirements: access to odometry from the avionics systems, LiDAR, and camera sensors, and enough computing power to process them (beyond any other task required from the UAV). All this hardware is focused on solving what can be described as a semantic segmentation problem. This is relevant given the enormous changes produced in the last decade in the computer vision field, and how classic problems like semantic segmentation are currently solved.

3. Semantic segmentation problem: classic approaches and deep learning (DL)

In the context of computer vision, the semantic segmentation problem is used to determine which regions of an image present an object of a given category, that is, a class or label is assigned to a given area (be it a pixel, window, or segmented region). The different granularity accepted is produced by how the technique and

its solution evolved: for a long time, it was completely unfeasible to produce pixel-wise solutions, so images were split according to different procedures, which added a complexity layer to the problem.

Current off-the-shelf technologies have changed the paradigm, as GPUs present huge capabilities in terms of parallelization, while solid-state disks make fast reliable storage cheap. These technical advancements have increased dramatically the performance, complexity, and memory available for data representation, especially for techniques inherently strong in highly parallelized environments. One of the fields where the impact has been more noticeable has been the artificial intelligence community, where the artificial neural network (ANN) has seen a resurgence thanks to the support this kind of hardware provides to otherwise computationally unfeasible techniques. The most impactful development in recent years has been the convolutional neural networks (CNNs), which have become the most popular computer vision approach for several of the classic problem and the default solution for semantic segmentation.

To understand the impact of deep learning into our proposed solution, we will discuss briefly how the classical segmentation pipeline worked and how the modern CNN-based classifier became the modern semantic segmentation techniques.

3.1 Classic semantic segmentation pipeline

The classic semantic segmentation pipeline can be split into two generic blocks, namely image processing for feature extraction and feature level classification. The first block generally includes any image preprocessing done and resizing/resampling, splitting the image into the regions/windows, defining the granularity level of the classification, and finally, extracting the features itself. The features can be of any type and frequently the ones feed to the classification modules will be a composition of several individual features from different detectors. The use of different window/region-based approaches helps build up higher level features, and the classification can be refined at later stages with data from adjacent regions.

Notice that this kind of architecture generally relies on classifiers which required very accurate knowledge or a dataset with the classes to learn specified for each input so it can be trained. **Figure 3** shows the detection of pipelines in classic semantic segmentation. Notice that to train the classifier, the image mask or classification result becomes also an input for the training process.

So, it can be seen that solving the semantic segmentation problem though classic pattern recognition methods requires acute insight into the specifics of the problem domain, as the features to be detected and extracted are built/designed specifically. This implies (as mentioned earlier) working from low-level features and explicitly deriving the higher level features from them is a very complex problem itself, as they are affected by the input characteristics, what is to be found/discriminated, and which techniques will be used in the classification part of the pipeline.

3.2 The segmentation problem with deep learning

Modern semantic segmentation techniques have organically evolved with the rise of the deep learning field to its current prominence. This evolution can be seen as a refinement in the scale of the inference produced from very coarse (image level probabilistic detection) to very fine (pixel level classification). The earliest ANN examples made probabilistic predictions about the presence of an object of a given class, that is, detection of objects with a probability assigned. The next step, achieved thanks to increased parallelization and network depth, was starting to tackle the localization problem, providing centroids and/or boxes for the detected

Figure 3.
Block diagram of a classical architecture approach for semantic segmentation using computer vision.

classes (the use of *classes* instead of *objects* here is deliberate, as the instance segmentation problem, separating adjacent objects of the same class, would be dealt with much later).

The first big break into the classification problem was done by AlexNet [13] in 2012, when it won the ILSVRC challenge, with a score of 84.6% in the top-5 accuracy test, while the next best score was only 73.8% (based on classic techniques). AlexNet has since then become a known standard and a default network architecture to test problems, as it is actually not very deep or complex (see **Figure 4**). It presents five convolutional layers, with max-pooling after the first two, three fully connected layers, and a ReLU to deal with non-linearities. This clear victory of the CNN-based approaches was validated next year by Oxford's VGG16 [16], one of the several architectures presented, winning the ILSVRC challenge with a 92.7% score.

While several other networks have been presented with deeper architecture, relevant development focused on introducing new types of structures into the networks. GoogLeNet [17], the 2014 ILSVRC winner, achieved victory thanks to the novel contribution of the inception module, which validated the concept that the CNN layers of a network could operate in other orders different from the classic sequential approach. Another relevant contribution produced by technology giants was ResNet [18], which scored a win for Microsoft in 2016. The introduction of residual blocks allowed them to increase the depth to 152 layers while keeping initial data meaningful for training the deeper layers. These residual blocks architecture essentially forwards a copy of the received inputs of a layer; thus, later layers received the results and same inputs of prior layers and can learn from the residuals.

More recently, ReNet [19] architecture was used to extend recurrent neural networks (RNNs) to multidimensional inputs.

The jump from the classification problem with some spatial data to pixel level labeling (refining inference from image/region to pixel level) was presented by Long [20], with the fully convolutional network (FCN). The method they proposed was based on using the full classifier (like the ones just discussed) as layers in a convolutional network architecture. FCN architecture, and its derivatives like U-Net [21] are the best solutions to semantic segmentation for most domains. These derivatives may include classic methods, such as DeepLab's [22] conditional random fields [23], which reinforces the inference from spatially distant dependencies, usually lost due to CNN spatial invariance. The latest promising contributions to the semantic segmentation problem are based on the encoder-decoder architecture, known as autoenconders, like for example SegNet [24].

For the works discussed in this chapter, a FCN16 model with AlexNet as a semantic segmentation model was used. The main innovation introduced by the general FCN was exploiting the classification power via convolution of the common semantic segmentation DL network, but at the same time, reversing the downsampling effect of the convolution operation itself. Taking AlexNet as an example, as seen in **Figure 4**, convolutional layers apply a filter like operation while reducing the size of the data forwarded to the next layer. This process allows producing more accurate "deep features" but at the same time also removes high-level information describing the spatial relation between the features found. Thus, in order to exploit the features from the deep layers while the keeping information from spatial relation, data from multiple layers has to be fused (with element-wise summation).

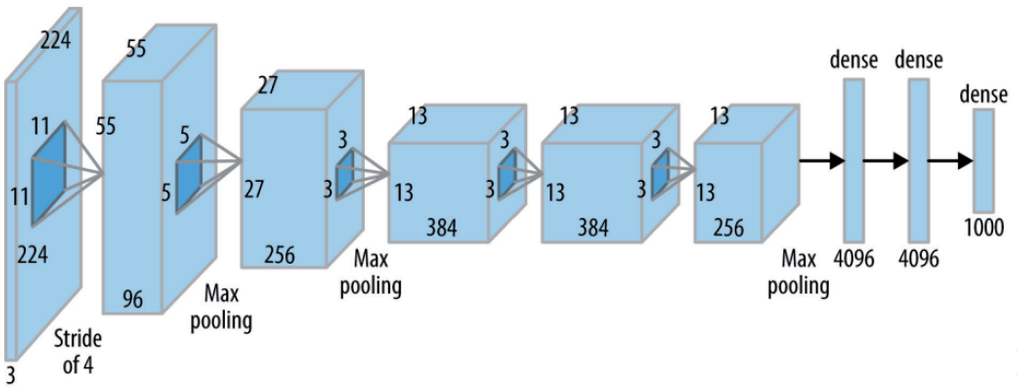

Figure 4.
Diagram of the AlexNet architecture, showcasing its pioneering use of convolutional layers.

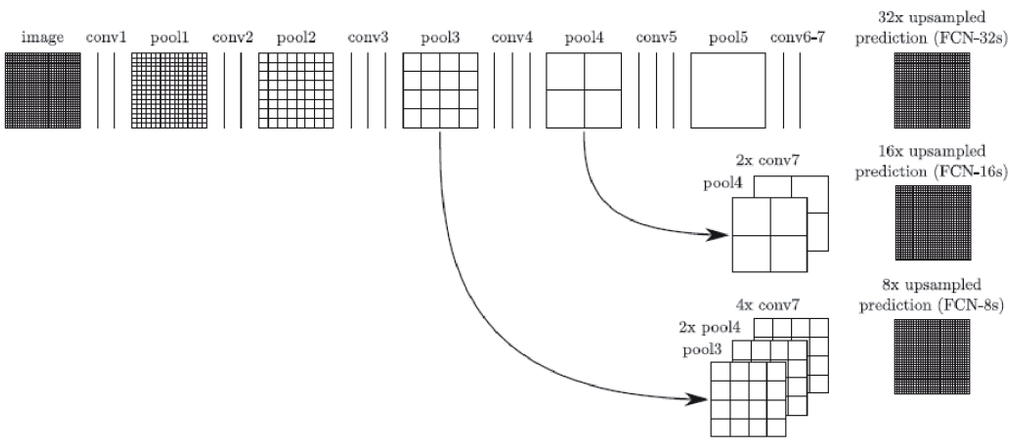

Figure 5.
Detail of the skip architectures (FCN32, FCN16, and FCN8) used to produce results with data from several layers to recover both deep features and spatial information from shallow layers (courtesy of [25]).

In order to be able to produce this fusion, data from the deeper layers are upsampled using deconvolution. Notice that data from shallow layers will be coarser but contain more spatial information. Thus, up to three different levels can be processed through FCN, depending on the quantity of layers deconvoluted and fused, as seen in **Figure 5**.

More information on the detailed working of the different FCN models can be found in [25]. It is still worth noting that the more shallow layers are fused, the more accurate the model becomes, but according to the literature, the gain from FCN16 to FCN8 is minimal (below 2%).

4. Automated ground truth labeling for multimodal UAV perception dataset

Classic methods using trained classifiers would pick designed features (based on several metrics and detectors, as discussed earlier) to parametrize a given sample and assign a label. This would allow creating small specific datasets, which could be used to infer the knowledge to create bigger datasets in a posterior step. The high specificity of the features chosen (generally with expert domain knowledge applied implicitly) with respect to the task generally made them unsuitable to export learning to other domains.

By contrast, deep learning offers several transfer learning options. That is, as it was proven by Yosinski [26], trained with a distant domain dataset are generally

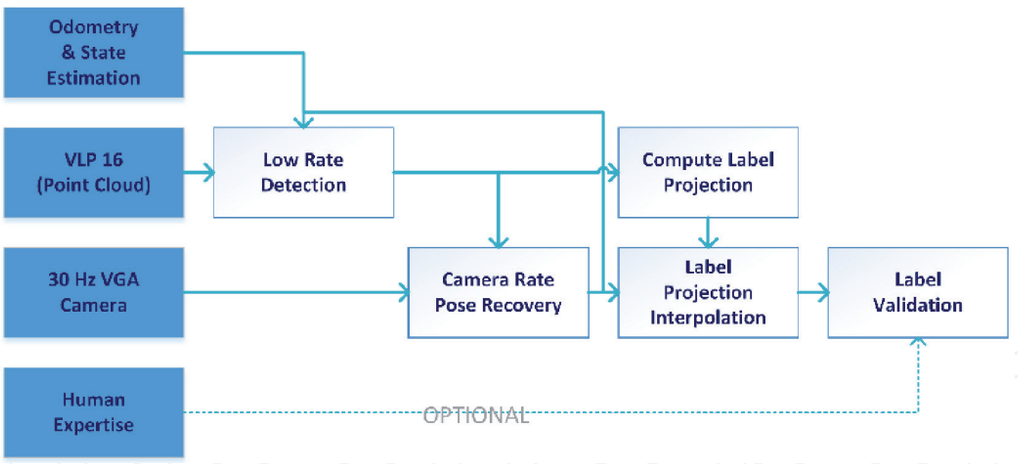

Figure 6.
The framework proposed to automatically produce labeled datasets with the multimodal perception UAV.

useful for different domains and usually better than training from an initial random state. Notice that the transferability of features decreases with the difference between the previously trained task and the target one and implies that the network architecture is the same up to the transferred layers at least.

With this concept in mind, we decided to build a dataset to train an outdoor industrial pipe detector with pixel level annotation to be able to determine the position of the pipe. While the ability of transfer learning allows us to skip building a dataset with several tens of thousands of images, and therefore, the authors will work with a few thousand, which were used to fine-tune the network. These orders of magnitude are required as a "shallow" deep network. For instance, the AlexNet already presents 60 million parameters.

Capturing and labeling a dataset is a cumbersome task, so we also set to automatize this task with minimal human supervision/interaction, exploiting the capabilities of the sensing architecture proposed in earlier works described in Section 2.

This framework, see **Figure 6**, uses the images captured by the UAV camera sensor, the data processed by the localization approach chosen (see Section 2) to obtain the UAV odometry, and pipe detection seeds from the RANSAC technique treating the LiDAR point cloud data. When a pipe (or generally a cylinder) is detected and segmented in the data sensor provided by the LiDAR, this is used to produce a label for the temporally near images, to identify the region of the image (the set of pixels) containing the pipe or cylinder detected and its pose w.r.t. the camera. Notice that even running the perception part, the camera works at a higher rate than the LiDAR, so the full odometric estimation is used to interpolate between pipe detections, to estimate where the label should be projected into the in-between images (just as it was described for the pipe prediction in Section 2).

This methodology was used to create an initial labeled dataset with actual data captured in real industrial scenarios during test and development flights, as it will be discussed in the next section.

5. Experimental evaluation

To evaluate the viability of the proposed automated dataset generation methodology, we apply it to capture a dataset and train several semantic segmentation networks with it. To provide some quantitative quality measurement for the solutions produced, we use modified standard metrics for state-of-the-art deep learning, accounting that in our problem we are dealing with only one semantic class:

- **PA** (pixel accuracy): base metric, defined by the ratio between properly classified pixels *TP* and the total number of pixels in an image, pix_{total}:

$$PA = \frac{TP}{pix_{total}} \tag{1}$$

Notice that usually, besides the PA the mean pixel accuracy (MPA) is provided, but in our case, it reduces to the same value of PA, thus it will not be provided.

- **IoU** (intersection over union): standard metric in segmentation. The ration is computed between the intersection and union of two sets, namely the found segmentation and the labeled ground truth. Conceptually, it equals to the ratio between the number of correct positives (i.e., the intersection of the sets) TP, over all the correct positives, spurious positives FP and false negatives FN (i.e., the union of both the ground truth and segmentation provided). Usually, it is used as mean IoU (MIU), averaging the same ratio for all classes.

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

An additional metric usually computed along with the MIU is the frequency weighted MIU, which just weighs the average IoU computed at MIU according to the relative frequency of each class. The MIU, in our case, IoU is the most relevant metric and the most widely used when reporting segmentation results (semantic or otherwise).

5.1 Dataset generation results

The system proposed was implemented over the ROS meta-operating system, just as in previous works [6], where the UAV system used to capture the data is described. A set of real flights in simulated industry environments was performed, where flights around a pipe were done. During these flights, averaging ~240 s, an OTS USB camera was used to capture images (at 640 × 480 resolution), achieving an average frame rate of around 17 fps. This translated in around 20,000 raw images captured, including the parts of flight where no industry-like elements are present, thus of limited use.

Notice that as per the method described, the pipe to be found can be only labeled automatically when the LiDAR sensor can detect it; thus, the number of images was further reduced due to the range limitations of the LiDAR scanner. Other factors, such as vibrations and disruptions in the input or results of required perceptual data, further reduced the number of images with accurate labels.

Around ~2100 images were automatically labeled with a mask assigning a ground truth for the pipe in the image. After an initial human inspection of the assigned label, a further ~320 were rejected, obtaining a final set of 1750. The image rejected produced spurious ground truths/masks. Some of them had inconsistent data and the reprojection of the cylinder detected in through RANSAC in LiDAR scans was not properly aligned (error could be produced by spurious interpolation of poses, faulty synchronization data from the sensors, or due to deformation of the UAV frame, as it is impossible for it to be perfectly rigid). Another group presented partial detections (only one of the edges of the pipe is visible in the image), thus making it useless for the apparent contour optimization. A third type of error found was produced by the vision-based pipeline, where a spurious mask was generated,

commonly some shadows/textures displace/retort the edge, or areas not pertaining to the pipe are assigned due similarity of the texture and complexity of delimiting the areas.

A sample of the labeling process can be seen in **Figure 7**, with the original image, the segmented pipe image, and approximations to centroid and bounding bow.

Out of the several options available to test the validity of the dataset produced, the shallow architecture AlexNet was selected, as it could be easily trained and it would provide some insight in the performance that could be realistically expected from a CNN-based approach deployed in the limited hardware of a UAV.

According to previous literature, the dataset was divided into training, validation, and test at the standard ratio of 70, 15, and 15% respectively.

To match the input of AlexNet the images were resized to 256 × 256 resolution. This was mainly done to reduce the computational load, as the input size could be easily fit adjusting some parameters, like the stride. To train and test the network, the Pytorch library was used, which provides full support for its own implementation of AlexNet.

To produce some metrics relevant to the network architecture just trained, a modified version of the technique used to label the dataset was used. Note that this approach, as described in previous sections, uses LiDAR, cameras, and odometry to: acquire an initial robust detection (from LiDAR), track its projection and predict it in the camera image space (using odometric data), and finally determine its edges/contour in the image. The robustness of the LiDAR detection is mainly due to exploiting prior knowledge (in the form of the known radius of the pipe to detect) that cannot be introduced into the AlexNet architecture to produce a meaningful comparison. So, a modified method, referred to as NPMD (no-priors multimodal detector) was employed to estimate the accuracy of earlier work detector without priors. The main difference was modifying the LiDAR pipeline to be able to detect several pipes with different radius (as it should be considered unknown). This led to the appearance of false positives and spurious measurements, which in turn weakened the results produced by the segmentation part of the visual pipeline.

Thus, FCN with AlexNet classification was trained using a pre-trained model for AlexNet, with the standard stochastic gradient descend (SGD) with a momentum of 0.9. A learning rate of 10^{-3} was used, according to known literature, with image batches of 20. The weight decay and bias learning rate were set to standard values of 5.10^{-4} and 2, respectively. Without any prior data, and no benefit to obtain by doing otherwise reported in any previous works, the classifier layer was set to 0, and the dropout layer in the AlexNet left unmodified. This trained model produced the results found in **Table 1**.

Figure 7.
Left: dataset image. Middle: bounding box and centroid of the region detected. Right: segmentation mask image.

	AlexNetFCN	UPMD
PA	73.4	56.7
IoU	58.6	42.1

Table 1.
Experimental results obtained by AlexNet-based FCN.

It can be seen that eliminating the seed/prior data from the multimodal detector made it rather weak, with very low values for IoU, signaling the presence of spurious detections and probably fake positives. The FCN-based solution was around 1.5 times better segmenting the pipe, being a clear winner. This was to be expected as we deliberately removed one of the key factors contributing to the LiDAR-based RANSAC detection robustness, the radius priors, leading to the appearance of spurious detections.

It is worth noting that although the results are not that strong in terms of metrics achieved for a single-class case, there are no other vision-only pipe detectors with better results in the literature, neither other approaches actually tested in real UAV's platforms, like authors' previous works [6].

6. Conclusions

The field of computer vision has been greatly impacted by the advances in deep learning that have emerged in the last decade. This has allowed solving, with purely vision-based approaches, some problems that were considered unsolvable under this restriction. In the case presented, a detection and positioning problem, solved with limited hardware resources (onboard a UAV) in an industry-like uncontrolled scenario through a multimodal approach, has been solved with a vision-only approach. The previous multimodal approach relied in LiDAR, cameras, and odometric measurements (mainly from GPS and IMU) to extract data with complex algorithms like RANSAC and combine them to predict the position of a pipe and produce a measurement. This system was notable thanks to its robustness and performance but presented the huge requirements detailed in [6]. In order to solve the problem in a simpler and more affordable manner, a pure visual solution was chosen as the way to go, exploring the deep learning opportunities.

Although the switch to a pure visual solution meant that during its use, the procedure would only use the camera sensor, the multimodal approach was still used to capture data, and through a series of modifications, turn it into an automatic labeling tool. This allowed building a small but complete dataset with fully labeled images relevant to the problem that we were trying to solve. Finally, to test this dataset, we train a DL architecture able to solve the semantic segmentation problem. Thus, three different contributions have been discussed in this chapter: firstly, a dataset generator exploiting multimodal data captured by the perception system to be replaced has been designed and implemented; secondly, with this dataset generation tool, the data captured has been properly labeled so it can be used for DL applications; and finally, a sample lightweight network model for semantic segmentation, FCN with AlexNet classification, has been trained and evaluated to test the problem.

By the same reasons that there was no dataset available for our challenge and we had to capture and develop one dedicated to our domain, there were no related works to obtain metrics. In order to have some relevant metrics to compare the results of the developed approach, a modified version of [13] was produced and

benchmarked without the use of prior knowledge. Under these assumptions, the new CNN-based method was able to clearly surpass the multimodal approach, though it still lacks robustness to be considered ready for industrial standards. Still, these initial tests have proven the viability of the built dataset generator and the utilization of CNN-based semantic segmentation to replace the multimodal approach.

Acknowledgements

This research was funded by the Spanish Ministry of Economy, Industry and Competitiveness through Project 2016-78957-R.

Author details

Edmundo Guerra[1], Jordi Palacin[2], Zhuping Wang[3] and Antoni Grau[1*]

1 Automatic Control Department, Technical University of Catalonia UPC, Barcelona, Spain

2 Department of Informatics and Industrial Engineering, University of Lleida - UdL, Lleida, Spain

3 Department of Control Science and Engineering, Tongji University, Shanghai, China

*Address all correspondence to: antoni.grau@upc.edu

References

[1] Wang Z, Zhao H, Tao W, Tang Y. A new structured-laser-based system for measuring the 3D inner-contour of pipe figure components. Russian Journal of Nondestructive Testing. 2007;**43**(6):414-422

[2] Song H, Ge K, Qu D, Wu H, Yang J. Design of in-pipe robot based on inertial positioning and visual detection. Advances in Mechanical Engineering. 2016;**8**(9):168781401666767

[3] Hansen P, Alismail H, Rander P, Browning B. Visual mapping for natural gas pipe inspection. The International Journal of Robotics Research. 2015;**34**(4-5):532-558

[4] Zsedrovits T, Zarandy A, Vanek B, Peni T, Bokor J, Roska T. Visual detection and implementation aspects of a UAV see and avoid system. In: IEEE. 2011. pp. 472-475. [cited 05 March 2018]. Available from: http://ieeexplore.ieee.org/document/6043389/

[5] Holz D, Nieuwenhuisen M, Droeschel D, Schreiber M, Behnke S. Towards multimodal omnidirectional obstacle detection for autonomous unmanned aerial vehicles. ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2013;**1**:201-206

[6] Guerra E, Munguía R, Grau A. UAV visual and laser sensors fusion for detection and positioning in industrial applications. Sensors. 2018;**18**(7):2071

[7] Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM. 1981;**24**(6):381-395

[8] Choi S, Kim T, Yu W. Performance evaluation of RANSAC family. Journal of Computer Vision. 1997;**24**(3):271-300

[9] Raguram R, Frahm J-M, Pollefeys M. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In: ECCV 2008 (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer; 2008. pp. 500-513

[10] Zhang J, Singh S. LOAM: Lidar odometry and mapping in real-time. In: Proceedings of the "Robotics: Science and Systems. 2014" conference, July 12-16, 2014, Berkeley, USA. 2014. pp. 1-9

[11] Besl PJ, McKay ND. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1992;**14**(2):239-256

[12] Doignon C, de Mathelin M. A degenerate conic-based method for a direct fitting and 3-d pose of cylinders with a single perspective view. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. Roma, Italy: IEEE; 10-14 July 2007. pp. 4220-4225

[13] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012. pp. 1097-1105

[14] Marchand E, Spindler F, Chaumette F. ViSP for visual servoing: A generic software platform with a wide class of robot control skills. IEEE Robotics and Automation Magazine. 2005;**12**(4):40-52

[15] Bartoli A, Sturm P. The 3D line motion matrix and alignment of line reconstructions. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Kauai, HI,

USA. Vol. 1. 8-14 December 2001. pp. I-287-I-292

[16] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. Available from: http://arxiv.org/abs/1409.1556

[17] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA; 7-12 June 2015. pp. 1-9

[18] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA; 27-30 June 2016. pp. 770-778

[19] Visin F, Kastner K, Cho K, Matteucci M, Courville A, Bengio Y. ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks. 2015. Available from: http://arxiv.org/abs/1505.00393

[20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA; 7-12 June 2015. pp. 3431-3440

[21] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. (Lecture Notes in Computer Science). Cham: Springer International Publishing; 2015. pp. 234-241

[22] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;**40**(4):834-848

[23] Lafferty J, Mccallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: 2001 Conference on Machine Learning, ICML, Williamstown, MA, USA; 28 June-1 July 2001. pp. 282-289

[24] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;**39**(12):2481-2495

[25] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;**39**(4):640-651

[26] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. Montrea (CANADA): Curran Associates, Inc.; 2014. pp. 3320-3328