EVALUATION OF THE NEURAL NETWORK OBJECT DETECTION IN MULTI-MODAL IMAGES

Adam Ligocki
Doctoral Degree Programme (4), FEEC BUT
E-mail: adam.ligocki@vutbr.cz

Supervised by: Luděk Žalud
E-mail: zalud@feec.vutbr.cz

Abstract: This paper studies the information gain of various data domains that are commonly used in the modern Advanced Driving Assistant Systems (ADAS) to develop robust systems that would increase traffic safety. We could see a fast growth of many Deep Convolutional Neural Networks (DCNN) based solutions during the last several years. These methods are state-of-the-art in object detection and semantic scene segmentation. We created a small annotated dataset of synchronized RGB, grayscale, thermal, and depth map images and used the modern DCNN framework tool to evaluate the object detection robustness of different data domains and their information gain process understanding the surrounding environment of the semi-autonomous driving agent.

Keywords: Multi-modal, Object Detection, Convolutional Neural Network, RGB, Grayscale, Thermal, IR, Depth Map

1 INTRODUCTION

These days, we can see the dramatic development of various ADAS systems that improve all traffic members’ safety. However, the question of which sensory equipment setup is the best for a fail-safety and robust autonomous agent’s orientation stays unanswered. It is also quite clear that in the future, we will need to develop a wide-range of data fusion methods that will combine information from different data sources into a single robust model that will help to understand the situation around the agent.

In this paper, we focus on the several data domains commonly used in automotive. We took the RGB camera, thermal camera, and 3D LiDAR scanners, and we processed these data, so all are represented as a synchronized and unified 2D images of the traffic situations, and we tested the performance of these days state-of-the-art neural network in the task of object detection.

2 RELATED WORKS

In the publically available literature, the topic of combining multimodal data by a single deep neural network model is not new. When we talk about the thermal and visible-spectrum images data fusion methods, the most frequent topic is enhancing pedestrian detection in bad lighting or weather conditions [3] [15].

Another way to improve visible spectrum neural network object detections in traffic is to extend the RGB information with depth map (LiDAR data) [1], [9]. Literature also covers the fusion of all three domains studied by this paper [8], [10]. However, these works usually focus only on pedestrian detections and do not cover the ADAS topic or deep learning methods.

However, non of the papers mentioned above discuss the information gain from the specific data domain (visible spectrum, thermal, and depth map), which is my work’s main aim.
3 MATERIALS AND METHODS

3.1 DATA GENERATION

To create the training and validation data for this experiment, we used our existing Brno Urban Dataset [6], and Atlas Fusion Framework [5].

In the first phase, we detected objects using the YOLOv5 on the raw RGB video data. Later, using the principle described in [7], to project detections from RGB images into the thermal ones. This way, we generated annotated therm images. Simultaneously, for every thermal image, the full 3D point cloud model was projected into the camera plane, so the image’s corresponding depth map was generated. Last, we estimated the perpendicular plane in the thermal camera’s field of view (FoV) in the distance of 50m from the camera. We projected the area of this plain into the RGB camera to approximate the common FoV for both the thermal and RGB cameras. Cutting out this area, we created the RGB image that covers approximately the same scene as the thermal image with minimal epipolar distortion, as both the RGB and the thermal cameras are placed near each other on the sensory framework. We also converted the RGB images into the grayscale to compare the object detection on the images with reduced information. In the case of this paper, we annotated only vehicles. All other classes were left unused.

![Figure 1: Tuple of annotated multi-modal data. From left, RGB image, information reduced grayscale image, thermal image, depth map image and annotations visualized on RGB image on the right. Annotations are same for all images.](image)

This way, we generated about 7500 annotated image tuples containing RGB image, grayscale image, thermal image, and the depth map. We split this set of tuples by 4:1 ratio, training to test set.

3.2 NEURAL NETWORK TRAINING

To train a neural network on our dataset, we used the existing YOLOv5 framework [4], which is publically available on Github. The framework provides modified YOLOv3 [13] architecture implemented in the Pytorch framework, mainly updated with training process augmentations and API for training models on custom datasets.

3.2.1 YOLOv3 ARCHITECTURE

The YOLOv3 [13] architecture is an evolution of this end-to-end object detection model’s previous versions. Originally introduced as a “You only look once” (YOLO) [11], authors proposed the neural network model that accepts N-channel image on the input and by single inference with the image, it provides an array of detections on the output, without any region proposition phases, compared to the RCNN methods [2], [14]. The RCNN-like methods, on the other hand, separate object detection into two phases, region proposition, and region classification.

The YOLO architecture firstly passes the input image through the backend. The backed is the convolutional neural network (CNN) that produces feature maps on the output. These feature maps are later processed by the output layers that generate the tensor of object detection propositions. These tensors are NxNxM structures, where the NxN is the cell grid that divides the input image into the smaller
areas when the M is a vector of numbers that represents bounding box positions, dimensions, object detection confidence, and the classification score for all classes that model detects. See the Figure 2.

![Figure 2: Visualization of the output tensor. In this use-case, the input image is divided into the 3x3 cell grid, and for each cell, there are two bounding boxes proposed. Each bounding box is defined by x and y position w.r.t the cell, width, height, detection confidence, and the classification score for three possible classes.](image)

During the time the YOLO architecture developed, authors introduced many improvements [12], [13], like multi-class classification on the output, batch normalization, multi-scale feature map pyramid, or introducing new backend neural networks. For YOLOv3, the model has three output tensors, when each tensor proposes detections on the different cell grid sizes (20x20, 40x40, 80x80), which corresponds to the different scale and size of the detected objects.

3.2.2 METRICS

To measure the performance of our trained models, we used the commonly used Mean Average Precision score (mAP(x)), where x represents the threshold value intersection over union (IoU) between the ground truth and proposed detection to accept the proposed detection as a true positive sample. More specifically, we used the mAP(0.5) as a basic metric, the mAP(0.5:0.95) to test the model’s robustness, and commonly used F1 score to express the precision-recall relation.

3.2.3 TRAINING HYPERPARAMETERS

All trained neural networks are the YOLOv5 version S models with a 640x640 image size and 3-channels per image. The grayscale, thermal, and depth images, which are 1-channel by definition, were extended by two first channel copies.

We trained each model for 100 epochs, and for the performance test, we used the model from the epoch with the best validation score. For training, we used the ADAM optimizer. For testing, we set the IoU threshold to 0.6 and the confidence threshold to 0.1.

4 RESULTS AND DISCUSSION

We tested the object detection performance on each data domain independently. The Table 1 shows the overall results. The best results showed the RGB-based and grayscale-based models, with both mAP(0.5) and mAP(0.5:0.95) on high numbers. That is evidence of the high robustness of the models. Also, the depth-map-based model shows quite good results. Compared to the IR model, it deals quite well with occlusions (Fig 3, column 3) and even with distant objects (Fig 3, column 5).

The grayscale, IR, and depth images were extended from the 1-channel format into the 3-channel by copying the first channel twice. This way, we kept the same number of convolutions applied on each data domain, the same computational requirements, and the similar model capacity.
Table 1: Neural networks performance in given RGB, grayscale, depth and thermal domains. Each data domain were tested independently.

We performed experiments using the RGB and grayscale images with perfect lighting and weather conditions in this work. That makes it hard for the depth and IR domain specialized neural networks to reach similar even better results than visible spectrum-focused networks and show their advantages in the more challenging conditions. In future work, it will be interesting to focus on data where lighting conditions are degraded and where the visible spectrum sensors are partially blinded by the night, fog, or by water on the lens during the rain.

Figure 3: Validation datasets detection visualizations. By the rows (top to bottom) there are the RGB images, grayscale images, depth map images and thermal images.

Overall, there are two interesting discoveries that we did not expect before this study. First, reducing RGB information into the grayscale format did not affect object detection in any significant way. Second, the object detection on a depth map created by projecting the LiDAR data into the camera frame did way better than we originally expected. The results even outperformed the thermal camera-based object detection, see the Table 1 and detection visualization in the Figure 3.

ACKNOWLEDGEMENT

The completion of this paper was made possible by the grant No. FEKT-S-20-6205 - “Research in Automation, Cybernetics and Artificial Intelligence within Industry 4.0” financially supported by the Internal science fund of Brno University of Technology.
REFERENCES


