Computational Intelligence Based Point of Interest Detection by Video Surveillance Implementations

Emre Tercan
General Directorate of Highways, 13th Region, Department of Traffic Safety, Türkiye
etercan87@gmail.com

Serkan Tapkın
Department of Civil Engineering, Bayburt University, Türkiye
serkantapkin@bayburt.edu.tr

Furkan Küçük
Department of Electrical Engineering, TOBB University of Economics and Technology, Türkiye
furkankucuk1994@gmail.com

Ali Demirtaş
Department of Electrical Engineering, TOBB University of Economics and Technology, Türkiye
ademirtas@etu.edu.tr

Ahmet Özbayoğlu
Department of Computer Engineering, TOBB University of Economics and Technology, Türkiye
mozbayoglu@etu.edu.tr

Abdussamet Türker
Department of Computer Engineering, TOBB University of Economics and Technology, Türkiye
a.turker@etu.edu.tr

Abstract: *Latest advancement of the computer vision literature and Convolutional Neural Networks (CNN) reveal many opportunities that are being actively used in various research areas. One of the most important examples for these areas is autonomous vehicles and mapping systems. Point of interest detection is a rising field within autonomous video tracking and autonomous mapping systems. Within the last few years, the number of implementations and research papers started rising due to the advancements in the new deep learning systems. In this paper, our aim is to survey the existing studies implemented on point of interest detection systems that focus on objects on the road (like lanes, road marks), or objects on the roadside (like road signs, restaurants or temporary establishments) so that they can be used for autonomous vehicles and automatic mapping systems. Meanwhile, the roadside point of interest detection problem has been addressed from a transportation industry perspective. At the same time, a deep learning based point of interest detection model based on roadside gas station identification will be introduced as proof of the anticipated concept. Instead of using an internet connection for point of interest retrieval, the proposed model has the capability to work offline for more robustness. A variety of models have been analysed and their detection speed and accuracy performances are compared. Our preliminary results show that it is possible to develop a model achieving a satisfactory real-time performance that can be embedded into autonomous cars such that streaming video analysis and point of interest detection might be achievable in actual utilisation for future implementations.*

Keywords: *Point of interest detection, YOLO algorithm, R-CNN, TOOD, deep learning.*

Received November 24, 2022; accepted January 23, 2023
https://doi.org/10.34028/iajit/20/6/7

1. Introduction

Autonomous image analysis has always attracted researchers and industry professionals due to the wide implementation areas ranging from text-to-speech systems to autonomous vehicles. With the advancements in the computational power, data collection capabilities through Internet of Things (IoT) and edge computing devices, smart sensors, faster communication through high-speed wireless networks and the rise of the artificial intelligence and machine learning models, these autonomous image analysis systems became widely available and more feasible than ever. Within the field, video surveillance and autonomous object detection draws more attention, mostly from a security viewpoint. However, these systems are also valuable for autonomous vehicle advancements [34, 36].

Driverless cars and intelligent transportation systems have become an important research field [24, 31, 38]. Lots of industries and manufacturers are trying to introduce their fully autonomous driverless systems into the market. Object recognition is an essential part of such systems, and their flawless performance is crucial for the successful future of driverless car technology. Point Of Interest (POI) detection is gaining importance within the autonomous object recognition area. POI is considered as a specific location/object that is visible or nearby and its presence can affect the decision process of the observer. Even though there have been some prior studies for accurate detection of POIs like restaurants, gas stations, historical sites, etc., most of these studies are based on location estimation systems like Global Positioning System (GPS) [30, 42]. Due to the dynamic nature of driving and road conditions, the GPS based geocaching systems might not reflect the actual environmental surroundings accurately [12]. In addition, within the urban areas, the GPS signals are not always accessible in some occasions. Vision based POI detection models can be a viable solution to address such challenges [1, 2, 41].

In this paper, it is aimed to provide detailed information about the previous studies implemented on POI detection, in particular in autonomous systems. Since main focus is on the application of POI detection on autonomous vehicles, particular importance is given to image and/or video-based POI detection. The existing studies are only reviewed, but also some of the learning models that have been used in such systems are also introduced. In addition, the problems and opportunities for the future are pinpointed. In the second part of the study, a deep learning-based POI detection model is introduced that has been developed as a case study. Running performances of different state-of-the-art techniques are compared and a preliminary analysis is implemented to a roadway section in order to present how such models might be embedded into the autonomous cars.

The main contributions and important aspects of this paper are as follows:

1. A thorough literature survey of POI detection for road assistance and vehicle systems is provided.
2. The basic Deep Learning (DL) models that are used for POI detection problem are introduced.
3. A proof-of-concept use case for deep learning-based POI detection model is proposed.
4. Different DL models are compared based on their runtime and classification accuracy performances.
5. Open issues and future research opportunities for the POI detection problem are discussed.

2. POI Detection

A POI refers to the location of a defined point in a particular coordinate system that users may find relatively advantageous or interesting. Hotels, camping areas, tourism facilities, restaurants, fuel stations are particular examples of POIs. In the literature, various studies have been conducted about the POI detection/discovery in different application areas. Ruta *et al.* [30] proposed a new discovery tool for mobile devices at Augmented Reality (AR) for semantic addition of nodes in crowd-sourced Open Street Map (OSM) mapping. As an AR discovery facilitator, this tool provides an automatic match between the utiliser characteristics and the source definitions. Yu *et al.* [44] utilised machine learning algorithms and similarity metrics to estimate and classify the multi-feature resemblance measurement outcomes for multi-vendors' POI data. They stated that the Support Vector Machine (SVM) method could find duplicate POIs more efficiently compared to the naïve Bayesian classifier and decision trees. Hao *et al.* [10] presented a novel system that leverages camera locations and viewing directions. This system has features that can automatically detect interesting regions and objects POIs. Rohella and Singh [29] developed an algorithm that looks for nearby spatial objects for potential POIs such as travel and commercial centers. Shu *et al.* [33] presented a three-step algorithm that uses deep neural networks in order to learn and estimate POIs on 3D images by utilizing multiple feature identifiers.

Meanwhile, different studies focus on objects or POIs that have versatile features/characteristics along with their corresponding objectives. Even though various methods have been proposed for POI detection, image and video-based models have been of particular interest.

3. Computer Vision for POI Detection

Even though POI detection literature is mostly focused on static location-based models like geocaching through GPS, online approaches like real-time POI detection through offline image understanding methodologies can be viable alternatives for various autonomous systems. Autonomous vehicles and satellite systems are two examples for such applications of real-time POI detection, using computational intelligence, in particular deep learning. One of such POI based object detection systems using deep learning is being employed in this paper and will be explained in section 4. The proposed approach is a preliminary model based on a deep learning network presenting a proof of concept use case based on a real video footage on a highway. Various model performances are compared to analyse the effectiveness of different techniques. The aim is to demonstrate the general methodology along with the environmental challenges and difficulties. The main focus on this study is roadside POI detection, in particular gas station detection. This paper addresses that problem, and attempts to provide a working solution on how current state-of-art convolutional networks perform on that task.

3.1. Existing Computational Intelligence/Deep Learning-Based POI Detection Studies in the Literature

There are a number of machine learning based POI detection studies in the literature. Ahmad *et al.* [1] defined the road markings and signs as a segmentation problem and studied the basic principles and algorithms of such implementations through a digital image analysis point of view. As a result, road markings detection is considered as one of the most commonly encountered POI detection problems. They also studied road markings as a POI detection problem, however, their focus was not only on detection of these markings, but also recognizing and classifying them into 10 classes using CNN with a recognition rate of 99%. Wu and Ranganathan [41] developed a template matching based road markings detection model using video input from a car camera. They used Maximally Stable External Regions (MSER) features for their template generations and training their model. Then they tested their model based on the generated templates. They

trained their system based on 10 different road signs and the classification results indicate their POI detector was able to successfully detect designated POIs in real time. Bailo *et al*. [2] also used MSER features for their road markings detection model, however they embedded a density-based clustering technique to provide a more robust POI detection model that would not be effected from varying road and meteorological conditions. Li *et al.* [17] presented an algorithm for road markings detection to be integrated with the autonomous navigation system of intelligent transportation systems. They used an Inverse Perspective Mapping (IPM) transformation through low-level image processing techniques within their algorithm in order to enhance the detection performance. Greenhalg and Mirmehdi [9] and Kheyrollahi and Breckon [14] also developed autonomous road markings detection models in their studies. Lee *et al.* [16] used a deep learning model called Generative Adversarial Networks (GAN) for unconstrained roadside marking recognition in a similar fashion, such that they generated a Tiny-YOLO detector for their baseline, then created samples for their training set for the generator and discriminator network for their adversarial learning model. Their results indicate a classification performance of over 95% using the same dataset as Wu and Ranganathan [41].

Careful analysis of the POI detection studies in the literature indicates that most of these studies were concentrated on the road markings, roadside signs and static or dynamic objects alongside the road. However, there are more possibilities for POI detection, such as roadside buildings, restaurants, phone boots, and gas stations. These kind of studies have not been undertaken in an intense manner throughout the vast amount of literature published until date. Also, it is important to consider the possibilities for the lacking of a communication network. A robust system that can work offline can overcome this difficulty. At this point, we will introduce a POI detection model focused on gas station detection through offline model based on computer vision and deep learning.

4. Case Study: POI Detection from Car Camera Footage with Convolutional Neural Networks

To assess the efficacy of deep learning on the POI detection problem, a proof of concept model is developed and demonstrated using real video through a car surveillance camera. The developed model is clarified in depth in the subsequent sub-sections. Each sub-section will include the particular methods adopted for each process step within the model. On or after an application viewpoint, the methods in each step are employed to demonstrate how current state-of-the-art convolutional networks performs on the task of POI detection.

4.1. Data Set Acquisition and Preparation

Source of the data used to train the model was a 47 minute video with 25 fps captured from a moderately busy divided Antalya-Manavgat highway in the Mediterranean region of Türkiye. The acquired data was sufficient for providing discriminatory features, however was not large enough for training a deep learning model. Since that is the case, some augmentation methods were implemented to increase the number of data points within the dataset, as such the new dataset became adequate for training and testing.

Each image was passed from an augmentation pipeline for a larger dataset. The augmentation methods used are as follows:

1. Random Distortion: an image is distorted randomly with the probability of 0.5.
2. Random Erasing: it was observed that the model over fits to undesired locations of POIs. To address that issue, some parts of the picture that are irrelevant to the region of interest are filled with Gaussian noise to prevent overfitting. As a result, during the training process, better model convergence is achieved avoiding overfitting issues.
3. Random Brightness: depending on the time of the day, road conditions and weather, lighting conditions may vary from frame to frame. In order to provide more robustness for appropriate region extraction, some form of normalisation between individual frame variances might be helpful. Hence, augmenting data with random brightness is expected to make model more robust to these variances.
4. Random Colour Shifting: in some tests, it is observed that the model tends to pick a colour and gives it a higher chance to classify the designated region as a positive, even though it is not the desired region of interest. To address this issue, a random colour shifting augmentation policy was employed.

4.2. State-of-the-Art Architectures and Transfer Learning

Obtaining good results from a machine learning model can be a challenging task and it might require countless iterations to achieve a satisfactory outcome [11]. Many factors, ranging from data characteristics, feature properties, network topology to model choices and parameters, effect the overall performance during network training and it is always very difficult, if not impossible, to find the optimum machine learning model. “No free lunch theorem” states that there is no optimal architecture to fit in all data available [39]. However, we can achieve a reasonably good performance with a sub-optimal architecture. Employing state-of-the art architectures not only ensures the performance would be high enough to fulfil some expectations, but it also enables to use the transfer learning technique [23].

In general, the use of Artificial Neural Networks (ANNs) is not preferred in computer vision tasks, since ANNs require 1-dimensional inputs, and converting 2d images to 1d vectors increases the number of the input nodes dramatically [32]. The dramatic increase in the number of input nodes also exponentially increases the number of trainable parameters of models. Increasing trainable parameters not only increases memory and computational requirements but also requires large datasets for training [5]. To overcome these problems CNNs are proposed, and the first modern application of CNNs is implemented and trained by LeCun *et al.* [15]. CNN-based models decrease the number of trainable parameters, extract local spatial features from the image and combine extracted features in hierarchical architecture. Nowadays, CNN-based architectures are the dominant choice in Computer Vision tasks [3]. CNN basically applies small filters through images to extract spatial features (Figure 1).

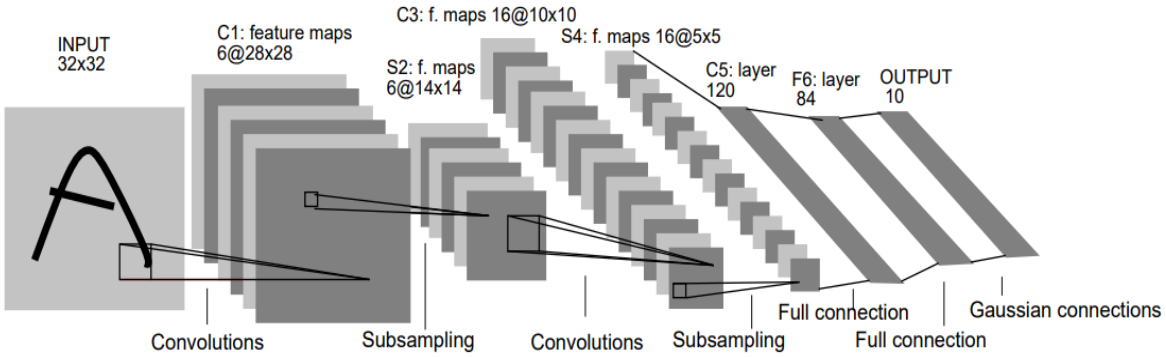

Figure 1. An example of basic CNN-based model architecture taken from LeCun *et al.* [15].

Every year a competition called "ImageNet" is held [13]. It is a challenge to classify a given large dataset and trying to obtain best results via the developed architecture. The contributors of this competition are highly motivated researchers and largest companies in the world. Since they have enough computational power to train on a whole ImageNet dataset, it is only logical to use those architectures to ensure performance and use the pre-trained models they constructed. Transfer learning is mostly based on this paradigm [23]. In recent years, ResNet [28] was more commonly used as the backbone network for CNN architectures. ResNet architecture employs deep residual connections between layers to address the vanishing gradient issue. Vanishing gradient issue is one of the main problems that the researchers encounter during training deeper neural networks [22]. The deeper network gets, the harder to get good gradients to train the first few layers. The gradients might get too small to have any meaningful effect during training and it becomes impossible to learn better filters. An example of the residual connection is shown in Figure 2.

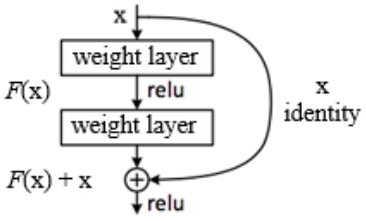

Figure 2. Residual Connection [6].

Transfer learning is a crucial part of the training process. The philosophy behind the idea is using the information obtained from other datasets and training experiences [23]. Considering a human interpretation of classifying images, humans tend to capture the features like edges, shapes and textures and obtain a result from this interpretation. Machine learning methods, in some level, rely on the same principals. However, since training a deep neural network takes lots of data to make the model capture good quality features from images, it is hard to train a CNN with a small dataset. Therefore, employing a state-of-the-art CNN architecture and using pre-trained model weights as a starting point for the training procedure generally achieves better performance than random weight initialization. Meanwhile, since the datasets indicate some similarities between them, the pre-trained network is expected to construct a good set of feature extractor filters with less effort [23].

4.3. Employed Model Architectures

CNNs are the main components powering the most state-of-the-art methods for classifying images [37]. With the employment of the filtering techniques adapted from signal processing applications, which are also employed in classical computer vision techniques, they are able to detect hidden or visible patterns in a given image or any kind of data which can be represented by 2-D matrices. However, classifying an image is only one of many scopes within the computer vision literature. In some areas, the accurate localization of the objects in a frame may be required for some tasks. Since CNNs are capable of extracting the patterns from the images [37], as an implicit side benefit, object detection, or region of interest determination from the images also

become feasible achievements that can be accomplished through the application of these versatile deep learning models, which is precisely what was aimed in this implementation.

Increasing demand for real time object detection led to development of a variety of algorithms. Some of the most commonly used algorithms were Region-based Convolutional Neural Networks (R-CNN) based. However, for the sake of keeping the delay of the object detection pipeline low, one-shot detection algorithms [19, 20] were considered. You Only Look Once (YOLO) algorithm is one of the most recognized models for object detection [25, 27, 48]. YOLO is semi-flexible in the terms of choosing backbone networks. It generally consists of 2 parts, convolutional network for feature extraction and fully connected dense network for prediction. Freedom of backbone network selection comes with some perks like employing state-of-the-art models like ResNet. This enables training by transfer learning and ensures working on a good network architecture for a computer vision task [25]. YOLOv5 [43] algorithm is basically an improvement on the previous YOLO [25, 26] algorithms. Overview of the YOLOv5 architecture is presented in Figure 3. During training of the object detectors Intersection over Union (IoU) is used to define positives and negatives. After training keeping same IoU threshold detectors tends to produce lots of noisy detections. This mainly happens because of overfitting during training, due to exponentially vanishing positive sample and inference-time mismatch between the IoUs for which the detector is optimal and those of the input hypotheses. Cascade R-CNN is proposed to overcome these issues.

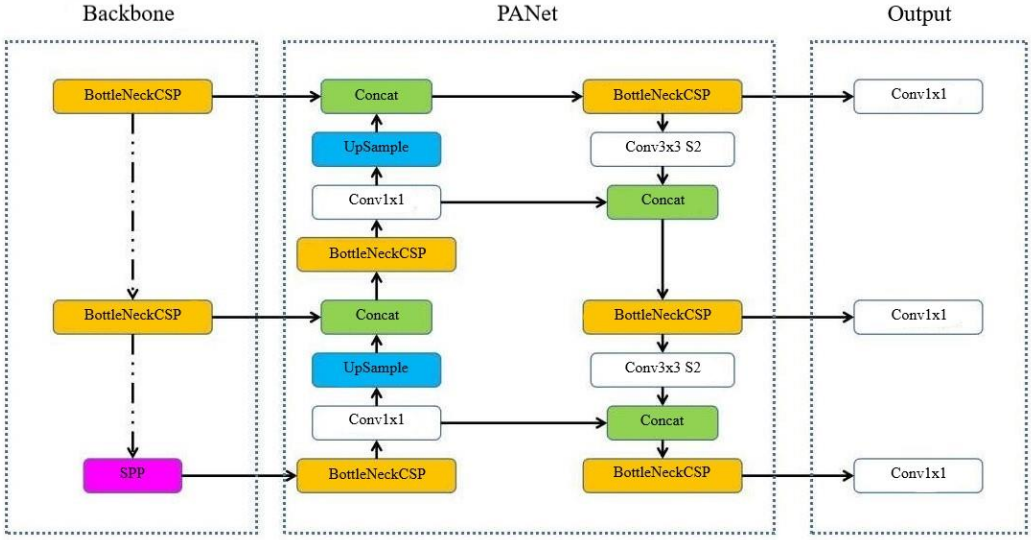

Figure 3. Overview of the YOLOv5 architecture taken from [43].

Object detection algorithms use classification scores to eliminate redundant bounding box candidates, however, using only classification scores is not sufficient to represent the candidate box qualities. Therefore, recent algorithms additionally predict either IoU scores [40] or centerness scores [35] to choose better box candidates. Zhang *et al.* [46] find a surprising result that there are accurately localized bounding boxes in box candidate pools. If we select them correctly, detector algorithms' performance can be increased by up to 94%. IoU-aware Classification Score proposed to rank detections which aim to merge localization accuracy with classification score.

One-stage object detectors commonly consist of classification and regression branches. However, using such architectures with two-branches might cause spatial alignments between classification branch and regression branch. To address this problem in Task Aligned One Stage Object Detection (TOOD), Feng *et al.* [8] propose a novel architecture with T-Head and alignment metric TAL (Figure 4). T-Head predicts classification and regression scores then during backpropagation computed TAL signals are used to align between two predictions. The details about this model can be obtained in Localization Distillation (LD) proposed a novel knowledge distillation method to efficiently transfer localisation knowledge from the teacher to the student (Figure 5). Proposed algorithm can be applied to different object detectors without sacrificing inference speed. The details about these models can be obtained in [4, 8, 46, 47]. Locating uncertain bounding box edges for the target object is ambiguous for the detectors. To address this problem using Knowledge Distillation (KD) may alleviate the problem. Zheng *et al.* [47] proposed a novel KD method called Localization Distillation to overcome this problem. The proposed method can be applied to any object detector without sacrificing inference speed.

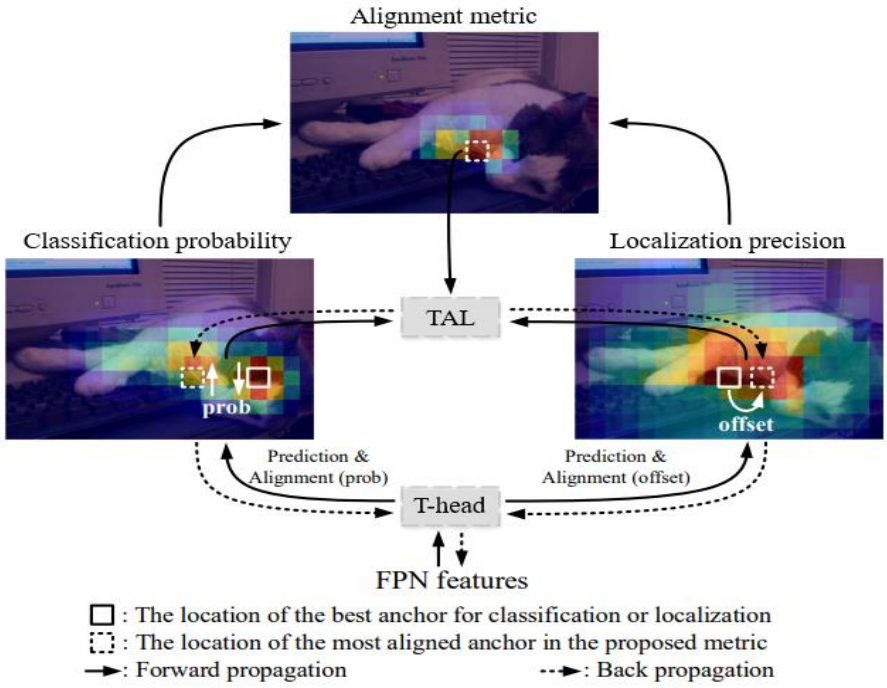

Figure 4. Overview learning mechanism of the TOOD taken from [8].

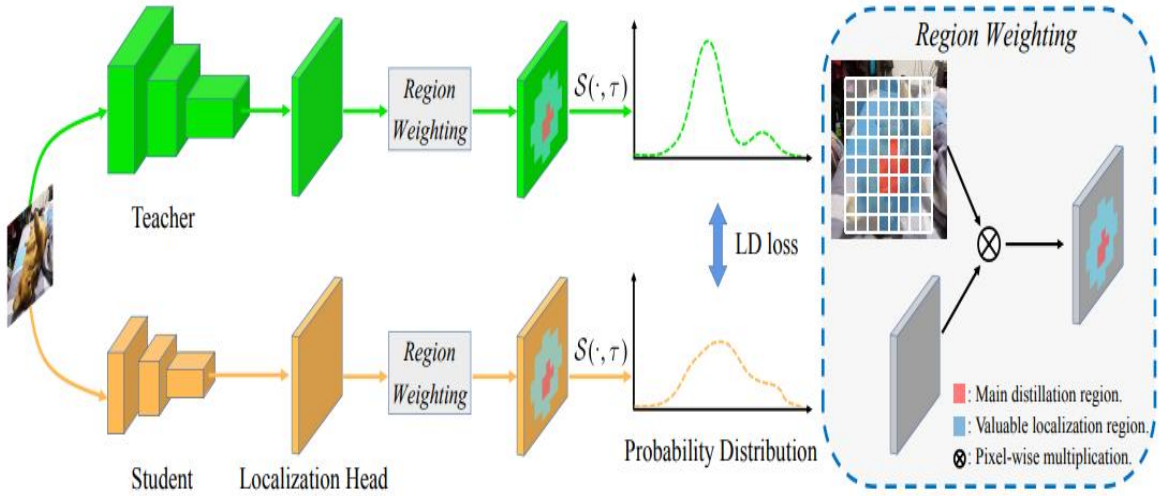

Figure 5. Overview of the location distillation pipeline taken from [47].

4.4. Experiments

In this section, we analyse the results from the obtained dataset. In all experiments we used RTX 3060 12 GB with AMD Ryzen 5 5600X on Ubuntu 20.04 to train and evaluate the models. In our Anaconda environment Python 3.8, PyTorch 1.10, and CUDA Toolkit 11.1 libraries are used. We have presented the architectures of object detectors in section 4.3. The prepared dataset is extracted from multiple road-trip videos which the environment and target object differed substantially. We created a dataset consisting of over 5000 images, all of which are 1920x1080 resolution. Due to the imbalance in the number of images that contain the target object versus those that do not, we employed oversampling techniques on the images that do contain the target object in order to balance the dataset. The obtained dataset was divided into training set (80%) and test set (20%). Images in the dataset that were spared for the training set was allowed to have augmented data. Validation set only consisted of the original acquired images, and augmented versions of these images were not used in the training set. The experiments consisted of comparing YOLOV5, Cascade R-CNN, VFNet, LD, TOOD, YOLOV3 and SSD300 algorithms. In this study, the chosen evaluation metrics for performance comparison are the widely used mAP0.5:0.95 mean Average Precision (AP), mAP0.75 and mAP0.95. Mean AP is a commonly used evaluation metric for object detection tasks. It is defined as the mean of the AP of all classes in the dataset. AP is defined as the integral of the precision-recall curve, which is a plot of precision versus recall at various confidence thresholds of the detector's output. The precision is defined as the ratio of

true positive detections to the number of true positive and false positive detections, while recall is defined as the ratio of true positive detections to the number of true positive and false negative detections. As it can be seen from the Table 1, relatively old algorithms YOLOv3 and SSD300 were not able to adequately extract the discriminative features to achieve higher accuracy. Most probably, a larger dataset might help getting better results. However, gas stations may include different features and the region determination can be complicated since some specific views of gas stations may include cars, a part of sky, a big sign, etc. In most cases, it can be observed that these variations increase the number of loose bounding boxes as shown in Figure 6. This may be the main reason why YOLOv3 and SSD300 approach were not able to perform as well as expected.

Figure 6. A correct localization with loose bounding box prediction.

Table 1. Object Tracking algorithms evaluations scores on our dataset. We chose the most common evaluation metrics. We collected mAP 0.5:0.95, mAP 0.5, mAP 0.75 scores and inference speed (fps). The first and second scores are marked as red and blue, respectively. In the YoloV5 repository we used, mAP 0.75 was not defined, so we filled that cell with X.

Detector	Backbone	mAP 0.5:0.95	mAP 0.5	mAP 0.75	Speed (fps)
Cascade R-CNN	Resnet50	0.612	0.937	0.731	12.4
YOLOv5L	CSPDarknet	0.608	0.967	X	58.7
YOLOv5S	CSPDarknet	0.575	0.949	X	111
VFNet	Resnet101	0.556	0.930	0.541	10.7
TOOD	Resnet50	0.541	0.948	0.536	16.4
VFNet	Resnet50	0.540	0.952	0.511	14.2
LD	Resnet34	0.495	0.951	0.478	16.3
LD	Resnet50	0.477	0.943	0.418	25.0
LD	Resnet18	0.400	0.948	0.360	29.8
YOLOv3	Darknet-53	0.329	0.670	0.218	17.3
SSD300	VGG16	0.208	0.362	0.225	13.8

Meanwhile, the Cascade-RCNN with ResNet50 backbone performs the best, achieving a 0.612 mAP0.5:0.95 score at 12.4 fps. YOLO family also performed well, YOLOv5L achieved 0.608 mAP0.5:0.95 scores and YOLOv5S achieved a 0.575 mAP0.5:0.95 score. The precision-confidence plot of YOLOv5L is presented in Figure 7, highlighting the precision achieved by the model at different levels of confidence. Considering inference speed YOLO models outperform other models by a big margin. Especially in edge devices, YOLO models can be useful. In our experiments, we observed that a bigger backbone does not always increase the performance of the algorithm. For instance, LD with Resnet34 achieved a slightly better mAP score than LD with Resnet50. Increasing the number of training samples may change this observation. One of the fundamental reasons behind the success of the Cascade-RCNN is probably due to the fact that its cascade architecture (backbone, neck, and other components) can generalize with less data than other algorithms. However, the fast execution performance of YOLOv5 makes it a desirable choice among the alternatives for object detection applications that require real-time video processing. We also trained bigger models like YOLOX-l but, the models did not converge.

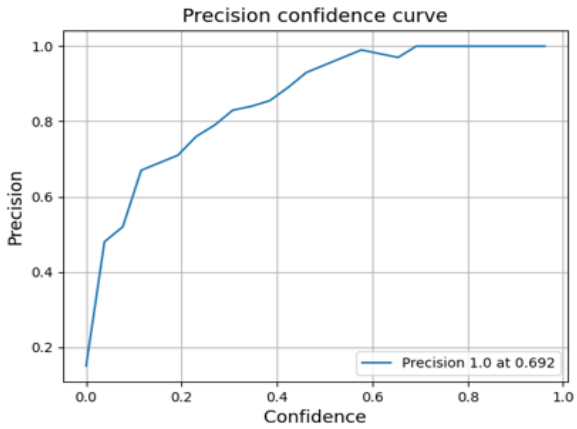

Figure 7. Precision confidence curve of YOLOv5L.

5. Discussion

Using the proposed approach to detect POI has several advantages. Google-Map, Apple Map and most mapping applications need to update their maps periodically. This process for periodic updating can take a substantial amount of time, it might be days, or sometimes weeks. As a result, solely depending on online maps and their corresponding POI databases might result in not fully up-to-date correspondences in real-time. Many users might have experienced recommendations for non-existent buildings, changed

roads, or the online application might not be able to properly recommend the optimum output due to the lack of identifying a new route, new building, new store, etc. It would have been very beneficial to have a system that combines the best of both worlds: An online POI database, which is periodically updated by the application server and an offline application that can detect POIs dynamically in real time.

The proposed POI detection approach finds new places and adds them to the conventional map applications instantly. Moreover, distributed mapping applications, which run on next-generation driverless cars automatically, update their maps and adjust their moves according to a new gas station or restaurant information. For example, cars can decide whether to buy gasoline by looking at the available gasoline and nearby gas station. Cars can also work as a meal recommender system. They propose potential places according to the locations of the restaurants and time of the day.

As it can be seen from the experiments, currently available and widely adopted state-of-the-art solutions are sufficiently powerful for tackling the problem. However, obtained data is far from sufficient for a good generalization for wide-range adaptation. POI features, brands or their appearances may tend to vary for other locations in different countries. So, the dataset can be enlarged by obtaining even more visuals from different locations.

Obtaining visuals and POIs is a time-consuming task if it is performed manually. The data for such mapping applications require human power, mostly through crowdsourcing to acquire the necessary POI data through appropriate tagging of the content and the coordinates. These manual efforts are subject to errors; wrong information content can be tagged for the POI, or the coordinate information might not be accurate. At the same time, there is always the possibility of GPS signal loss or lack of Wireless/4.5G signal preventing the car to be able to locate itself and/or access the online map and POI database. A more robust POI data acquiring might be helpful. Hence, instead of manually associating the POI information, the bounding box of the POI can be detected automatically using corners or edges. Currently the bounding boxes around gas station and restaurant logos are determined manually, as explained above. Instead, an automatic bounding can be generated heuristically. For instance, roadside image can be divided into sections like gas station pole, pavement, or predefined logos. These image sections are exploited to determine the bounding box. Motion estimation can also be applied to estimate the location of the bounding box or a spatio-temporal method can be developed to generate the bounding box. Furthermore, the proposed approach only detects a single POI. The algorithm is improved to detect multiple POIs.

6. Conclusions and Further Recommendations

In this study, the POI detection problem within the car surveillance videos is addressed. This particular problem is a major concern for providing the best Quality of Service (QoS) for autonomous vehicles. Most applications rely on online maps for POI recommendation. However, these maps sometimes may be out-of-synch or unreachable and the recommendation system might not able to recommend the best POI under these circumstances. Various techniques used throughout the literature are analysed including traditional image processing techniques along with new state-of-the-art deep learning models. Careful analysis of these algorithms indicates the deep learning models, especially the CNN based region detection techniques conveniently outperform traditional methods. However, the computational complexity is also high in these deep learning models. For demonstrating the effectiveness of CNN based models in POI detection problem, in this study, a gas station detection model was trained and tested with a real highway video footage. Among the CNN based models, Cascade R-CNN achieved the best accuracy with remarkably accurate prediction performance. Meanwhile, YOLOv5 also performed exceptionally well with not only very good accuracy performance but also with nearly real-time execution speed. 0.612 mAP0.5:0.95 score is achieved with this POI model. The results indicate that it might be possible to use such a POI detection system in real time without the explicit need for GPS access and an online POI database.

The only POI class considered in this paper was gas stations. This work can be easily extended by adding more classes. Furthermore, lots of POI classes tend to have similar properties with each other like having a sign. These properties may enable a hierarchical classification through detecting a parent class first, before detecting the child class. Hence, depending on the POI resolution, dynamic clusters of POIs can be analysed (i.e., restaurants, or specific restaurants such as Chinese Restaurants).

The learning mechanism of the proposed POI detection can be enhanced by integrating the information coming from other cars, GPS, base station internal sensors like Light Detection and Ranging (LIDAR). This additional information can be exploited to improve the detection accuracy of the learning model.

Places like gas stations or restaurants are static point of interests. The proposed approach can be extended to detect dynamic point of interests like police cars, ambulance, fire truck, suspicious vehicles (drunk drivers). To achieve this aim, the proposed approach must work faster. Hence, novel spatio-temporal POI detection mechanism should be developed. Moreover, sequential learning approaches like Recurrent Neural

Network (RNN)/LSTM can be used to estimate the location of the vehicle.

Lastly different computation models can be used to mitigate the computation complexity. In addition to the computational capabilities of cars, edge computing and cloud computing can be used to leverage the computational capabilities of centralized systems. Distributed computing can also be used. For instance, nearby cars share computation resources to distribute the complexity of the learning algorithms.

The exemplary performance of transformers in Natural Language Processing (NLP) tasks has brought interest from the Computer Vision community. Standard transformer architectures are designed to work on the text-based dataset, but Dosovitskiy *et al.* [7] proposed a novel transformer architecture for vision tasks. Nowadays, vision transformers are used in object detection [45, 49], segmentation [18] and classification [21] tasks widely. In POI detection transformer self-attention mechanism may make significant improvements. Furthermore, spatio-temporal transformer-based architectures can be implemented.

References

[1] Ahmad T., Ilstrup D., Emami E., and Bebis G., "Symbolic Road Marking Recognition Using Convolutional Neural Networks," *in Proceedings of the IEEE Intelligent Vehicles Symposium*, Los Angeles, pp. 1428-1433, 2017. doi: 10.1109/IVS.2017.7995910.

[2] Bailo O., Lee S., Rameau F., Yoon J., and Kweon I., "Robust Road Marking Detection and Recognition Using Density-Based Grouping and Machine Learning Techniques," *in Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Santa Rosa, pp. 760-768, 2017. doi: 10.1109/WACV.2017.90.

[3] Bhatt D., Patel C., Talsania H., Patel J., Vaghela R., Pandya S., Modi K., and Ghayvat H., "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope," *Electronics*, vol. 10, no. 20, pp. 2470, 2021. DOI:10.3390/electronics10202470

[4] Cai Z. and Vasconcelos N., "Cascade R-CNN: Delving into High Quality Object Detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, pp. 6154-6162, 2018. doi: 10.1109/CVPR.2018.00644.

[5] Debie E. and Shafi K., "Implications of the Curse of Dimensionality for Supervised Learning Classifier Systems: Theoretical and Empirical Analyses," *Pattern Analysis Application*, vol. 22, no. 21, pp. 519-536, 2019. https://doi.org/10.1007/s10044-017-0649-0

[6] Deep Residual Network, http://primo.ai/index.php?title=(Deep)_Residual_Network_(DRN)_-_ResNet, Last Visited, 2021.

[7] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., and Unterthiner T., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, 2020. arXiv:2010.11929.

[8] Feng C., Zhong Y., Gao Y., Scott M., and Huang W., "TOOD: Task-Aligned One-Stage Object Detection," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, pp. 3490-3499, 2021. DOI:10.1109/ICCV48922.2021.00349

[9] Greenhalgh J. and Mirmehdi M., "Automatic Detection and Recognition of Symbols and Text on the Road Surface," *in Proceedings of the International Conference on Pattern Recognition Applications and Methods*, Lisbon, pp. 124-140, 2015.

[10] Hao J., Wang G., Seo B., and Zimmermann R., "Point of Interest Detection and Visual Distance Estimation for Sensor-Rich Video," *IEEE Transactions on Multimedia*, vol. 16, no. 7, pp. 1929-1941, 2014.

[11] Haykin S., *Neural Networks and Learning Machines*, Pearson, 2009.

[12] Huang H., Gartner G., Krisp J., Raubal M., and Weghe N., "Location Based Services: Ongoing Evolution and Research Agenda," *Journal of Location Based Services*, vol. 12, no. 2, pp. 63-93, 2018. https://doi.org/10.1080/17489725.2018.1508763

[13] ImageNet Challenge, http://www.image-net.org/challenges/LSVRC/, Last Visited, 2021.

[14] Kheyrollahi A. and Breckon T., "Automatic Real-Time Road Marking Recognition Using a Feature Driven approach," *Machine Vision and Applications*, vol. 23, no. 1, pp. 123-133, 2012. DOI 10.1007/s00138-010-0289-5

[15] LeCun Y., Bottou L., Bengio Y., and Haffner P., "Gradient-based Learning Applied To Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, 2278-2324, 1998. DOI: 10.1109/5.726791

[16] Lee Y., Lee J., Hong Y., Ko Y., and Jeon M., "Unconstrained Road Marking Recognition with Generative Adversarial Networks," *in Procceedings of the IEEE Intelligent Vehicles Symposium*, Paris, pp. 1414-1419, 2019. https://doi.org/10.48550/arXiv.1910.04326

[17] Li H., Feng M., and Wang X., "Inverse Perspective Mapping Based Urban Road Markings Detection," *in Proceedings of the IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, Hangzhou, pp. 1178-1182, 2012. DOI:10.1109/CCIS.2012.6664569

[18] Liang J., Homayounfar N., Ma W., Xiong Y., Hu R., and Urtasun R., "Polytransform: Deep Polygon Transformer for Instance Segmentation," *in Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, p Seattle, p. 9131-9140, 2020. https://doi.org/10.48550/arXiv.1912.02801

[19] Lin T., Goyal P., Girshick R., He K., and Dollár P., "Focal Loss for Dense Object Detection," *in Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017. https://doi.org/10.48550/arXiv.1708.02002

[20] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C., and Berg A., "SSD: Single Shot Multibox Detector," *in Proceedings of the European Conference on Computer Vision*, Amsterdam, pp. 21-37, 2016.

[21] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., and Guo B., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012-10022, 2021.

[22] Pascanu R., Mikolov T., and Bengio Y., "On the Difficulty of Training Recurrent Neural Networks," *in Proceedings of the 30th International Conference on Machine Learning*, Atlanta, pp. 1310-1318, 2013. https://doi.org/10.48550/arXiv.2103.14030

[23] Pratt L., "Discriminability-Based Transfer Between Neural Networks," *in Proceedings of the 5th International Conference on Neural Information Processing Systems*, San Francisco, pp. 204-211, 1992.

[24] Ramakrishnan D. and Radhakrishnan K., "Applying Deep Convolutional Neural Network (DCNN) Algorithm in the Cloud Autonomous Vehicles Traffic Model," *The International Arab Journal of Information Technology*, vol. 19, no. 2, pp. 186-194, 2022. https://doi.org/10.34028/iajit/19/2/5

[25] Redmon J., Divvala S., Girshick R., and Farhadi A., "You Only Look Once: Unified, Real-Time Object Detection," *in Procceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 779-788, 2016. 10.1109/CVPR.2016.91

[26] Redmon J. and Farhadi A., "Yolov3: An Incremental Improvement, *arXiv preprint*, arXiv:1804.02767. [Online], 2018. https://doi.org/10.48550/arXiv.1804.02767

[27] Renwei T., Zhongjie Z., Yongqiang B., Ming G., and Zhifeng G., "Key Parts of Transmission Line Detection Using Improved YOLO V3," *The International Arab Journal of Information Technology*, vol. 18, no. 6, pp. 747-754, 2021. https://doi.org/10.34028/iajit/18/6/1

[28] ResNet50, https://keras.io/api/applications/resnet/#resnet50-function, Last Visited, 2022.

[29] Rohella A. and Singh S., "Path Independent Real Time Points of Interest Detection In Road Networks," *in Proccedings of the 2nd International Conference on Contemporary Computing and Informatics*, Greater Noida, pp. 633-638, 2016. doi: 10.1109/IC3I.2016.7918040.

[30] Ruta M., Scioscia F., Filippis D., Ieva S., Binetti M. and Di Sciascio E., "A Semantic-Enhanced Augmented Reality Tool for Openstreetmap POI Discovery," *Transportation Research Procedia*, vol. 3, pp. 479-488, 2014. https://doi.org/10.1016/j.trpro.2014.10.029

[31] Sankaran S., "Pattern Matching Based Vehicle Density Estimation Technique For Traffic Monitoring Systems," *The International Arab Journal of Information Technology*, vol. 19, no. 4, pp. 575-581, 2022. 2021 https://doi.org/10.34028/iajit/19/4/1

[32] Sarker I., "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, no. 6, pp. 420, 2021. https://doi.org/10.1007/s42979-021-00815-1

[33] Shu Z., Xin S., Xu X., Liu L., and Kavan L., "Detecting 3D Points of Interest Using Multiple Features and Stacked Auto-Encoder," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 8, pp. 2583-2596, 2019. https://doi.org/10.1109/TMM.2021.3070977

[34] Sreenu G. and Saleem M., "Intelligent Video Surveillance: A Review Through Deep Learning Techniques for Crowd Analysis," *Journal of Big Data*, vol. 6, no. 1, pp. 48, 2019. https://doi.org/10.1186/s40537-019-0212-5

[35] Tian Z., Shen C., Chen H., and He T., "FCOS: Fully Convolutional One-Stage Object Detection," *in Procceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9626-9635, 2019. https://doi.org/10.48550/arXiv.1904.01355

[36] Tsakanikas V. and Dagiuklas T., "Video Surveillance Systems-Current Status and Future Trends," *Computers and Electrical Engineering*, vol. 70, pp. 736-753, 2018. https://doi.org/10.1016/j.compeleceng.2017.11.011

[37] Venkatesan R. and Li B., *Convolutional Neural Networks in Visual Computing: A Concise Guide*, CRC Press, 2017.

[38] Vosooghi R., Puchinger J., Jankovic M., and Vouillon A., "Shared Autonomous Vehicle Simulation And Service Design," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 15-33, 2019. https://doi.org/10.1016/j.trc.2019.08.006

[39] Wolpert D. and Macready G., "Coevolutionary Free Lunches," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 6, pp. 721-735, 2005. doi: 10.1109/TEVC.2005.856205.

[40] Wu S., Li X., and Wang X., "IoU-Aware Single-Stage Object Detector for Accurate Localization," *Image and Vision Computing*, vol. 97, pp. 103911, 2020. https://doi.org/10.48550/arXiv.1912.05992

[41] Wu T. and Ranganathan A., "A Practical System for Road Marking Detection and Recognition," *in Procceedings of the IEEE Intelligent Vehicles Symposium*, Madrid, pp. 25-30, 2012. doi: 10.1109/IVS.2012.6232144.

[42] Yang W. and Ai T., "POI Information Enhancement Using Crowdsourcing Vehicle Trace Data and Social Media Data: A Case Study Of Gas Station," *ISPRS International Journal of Geo-Information*, vol. 7, no. 5, pp. 178, 2018. https://doi.org/10.3390/ijgi7050178

[43] YOLOv5. https://github.com/ultralytics/yolov5 Last Visited, 2022.

[44] Yu Q., Jiang H., Liu C., and Wu M., "The Application of Data Mining in Multi-Supplier Points of Interest Processing," *in Procceedings of the 9th International Conference on Natural Computation*, Shenyang, pp. 984-989, 2013. DOI:10.1109/ICNC.2013.6818119

[45] Zhang H., Li F., Liu S., Zhang L., Su H., Zhu J., and Shum H., "Dino: Detr With Improved Denoising Anchor Boxes for End-to-End Object Detection," *arXiv preprint*, arXiv:2203.03605. [Online], 2022.

[46] Zhang H., Wang Y., Dayoub F., and Sunderhauf N., "Varifocalnet: An Iou-Aware Dense Object Detector," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, pp. 8514-8523, 2021. https://doi.org/10.48550/arXiv.2008.13367

[47] Zheng Z., Ye R., Wang P., Ren D., Zuo W., Hou Q., and Cheng M. M., "Localization Distillation for Dense Object Detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9407-9416, 2022. https://doi.org/10.48550/arXiv.2102.12252

[48] Zhou S., Bi Y., Wei X., Liu J., Ye Z., Li F., and Du Y., "Automated Detection and Classification of Spilled Loads on Freeways Based on Improved YOLO Network," *Machine Vision and Applications*, vol. 32, no. 2, pp. 44, 2021. https://doi.org/10.1007/s00138-021-01171-z

[49] Zhu X., Su W., Lu L., Li B., Wang X., and Dai J., "Deformable Detr: Deformable Transformers for End-to-End Object Detection," *arXiv preprint*, arXiv:2010.04159. [Online], 2020. https://doi.org/10.48550/arXiv.2010.04159

Emre Tercan received his BS, MSc, and PhD degrees in Geomatics Engineering from Erciyes University, Kayseri, Türkiye. He has been in academia for more than 10 years and an Associate Professor of Geomatics Engineering for 2 years. His research interests include geographic information systems, regional planning, multi-criteria decision-making, renewable energy, photogrammetry, remote sensing, artificial intelligence, and transportation engineering. He is currently working in General Directorate of Highways, Türkiye.

Serkan Tapkın received his BS, MSc, and PhD degrees in Civil Engineering from Middle East Technical University (METU), Ankara, Türkiye. He has been in academia for more than 29 years and a full Professor of Civil Engineering for 8 years. His main research interest is all aspects of Transportation Engineering (mainly asphalt mixtures) and all aspects of artificial intelligence but in the last 6 years, his interests also spanned on geographic information systems, regional planning, multi-criteria decision-making, renewable energy, photogrammetry and remote sensing. He is currently working in the Civil Engineering Department of Bayburt University, Türkiye.

Furkan Küçük is currently a computer vision researcher at DataBoss Analytics. He received his BS from Electrical and Electronics Engineering at TOBB University of Economics and Technology and currently, pursues his MSc in Computer Engineering at Middle East Technical University, in Ankara, Türkiye. His research focus is mainly on deep learning, computer vision, forecasting and anomaly detection. He is currently conducting research on one-shot/few-shot learning, generative AI, domain adaptation, image enhancement, object detection and segmentation.

Ali Demirtaş received his PhD degree from Electrical Engineering and Computer Science Department at University California Irvine, USA in 2015. He is currently an Assistant Professor in the Department of Electrical and Electronics Engineering, TOBB University of Economics and Technology, Ankara, Türkiye. His current research areas are machine learning for communications, wireless networks, wireless communications, optimization, video coding, video streaming and rate-distortion theory.

Ahmet Özbayoğlu received his PhD degree from Engineering Management at Missouri University of Science and Technology, USA in 1996. He is currently an Associate Professor of Computer Engineering at TOBB University of Economics and Technology, in Ankara, Türkiye. His research interests include machine learning, pattern recognition, deep learning, financial forecasting, computational intelligence, machine vision.

Abdussamet Türker is currently a Computer Engineering student at TOBB University of Economics and Technology, Ankara, Türkiye. His current research areas are object detection, semantic segmentation, and machine learning.