Memoirs of the Scientific Sections of the Romanian Academy
Tome XLVIII, 2025

COMPUTER SCIENCE

DEEP LEARNING-BASED TRAFFIC PARTICIPANT DETECTION, CLASSIFICATION AND TRACKING FRAMEWORK USING NONLINEAR DIFFUSION-BASED MULTI-SCALE ANALYSIS

BARBU TUDOR[1,2], BEJINARIU SILVIU-IOAN[1] AND LUCA RAMONA[1]

[1] Institute of Computer Science of the Romanian Academy - Iasi Branch, Iasi, Romania
[2] The Academy of the Romanian Scientists, Bucharest, Romania
Corresponding author: tudor.barbu@iit.academiaromana-is.ro

A novel detection, recognition and tracking framework for the road traffic participants is introduced in this research work. A deep learning-based participant detection technique is proposed first. Voluminous vehicle, pedestrian and animal image databases have been developed for this purpose and some convolutional neural networks (CNN) have been trained, validated and tested on those datasets. Next, the detections generated by the obtained CNN-based detectors are subjected to a tracking-by-detection (TBD) process that applies a deep learning-based multi-scale feature extraction using a nonlinear fourth-order diffusion-based scale-space and then an instance matching algorithm. The performed simulations and method comparison are finally discussed here.

Keywords: Road traffic participant detection and tracking; Vehicle, pedestrian and animal databases; CNN training, validation and testing; Tracking-by-detection scheme; Multi-scale feature extraction; Nonlinear diffusion.

1. INTRODUCTION

The continuous increase in the number of vehicles has contributed to a significant rise in road accidents in recent years. Consequently, there has been a growing research interest in smart city technologies aimed at enhancing the safety of traffic participants, including intelligent surveillance systems and smart traffic management solutions [1].

The most common road traffic participants represent vehicles, pedestrians and animals. Detection, recognition and tracking of these participants represent essential tasks for these smart city technologies and their application fields include also the video security systems, law enforcement, biometrics, human–computer interaction, robotic vision, autonomous vehicles, human activity recognition and augmented reality.

The detection approaches use frame differences [2], Gaussian Mixture Models (GMM) [3], optical flow [4], adaptive boosted classifiers [5], Support Vector Machines (SVM) with Histograms of Oriented Gradients (HOG) [6], active contours [7], Aggregate Channel Features (ACF) [8] and Deep Learning (DL) [9]. Tracking, or counting, methods apply mean-shift algorithms [10], instance matching [11], Kalman filters [3], Hidden Markov Models (HMM) [12], optical flow estimation and active contours [13], and Convolutional Neural Networks (CNNs) [14].

This research, which has been conducted within a road traffic monitoring project, introduces a new framework for multiple traffic participant detection, classification and counting that combines successfully deep learning and partial differential equation (PDE) – based models. It's CNN – based traffic participant detection and recognition component is described first, in the following section. Then, the counting component is presented in the third section. It applies a tracking by detection (TBD) procedure using a deep learning-based multi-scale high-level feature extraction based on a nonlinear fourth-order diffusion-based scale-space representation, and then a participant instance matching algorithm.

The simulations performed using the proposed framework are discussed in the fourth section, where method comparison results are also provided. The conclusions and future research directions in this domain are outlined in the last section of this work.

2. DEEP LEARNING-BASED VEHICLE, PEDESTRIAN AND ANIMAL DETECTION

We have developed and annotated some voluminous image databases for vehicles, pedestrians, which are available on the website of our SIMPATIA project[1] [15]. The vehicle database contains images obtained from 20 traffic videos recorded at 30 fps, filtered with 3D anisotropic diffusion-based models [16, 17] and containing more than 400K frames. We have extracted 52331 frames and annotated almost 400K objects in the images with 2D bounding boxes represented as [*x-center*, *y-center*, *width*, *height*]. Annotated objects represent most common land vehicles: cars, buses, trucks, trams and bikes (including motorcycles).

Our pedestrian database has been obtained from 20 videos recorded at 30 fps with a mobile camera and containing around 300 K frames of [960 × 540] size. The video sequences have been pre-processed by applying 3D diffusion-based filters to remove AWGN and quantum noises [16, 17]. One has extracted around 50K frames and annotated 150K pedestrian objects in them with bounding boxes represented in the same normalized format.

The animal database has been obtained from 30 videos recorded by using a high-resolution video camera, at a recording speed of 30 fps. We set 7 classes of

[1] SIMPATIA – http://iit.academiaromana-is.ro/grantSIMPATIA.html

animals to be detected: 0 – *Dog*, 1 – *Cat*, 2 – *Bear*, 3 – *Cattle*, 4 – *Horse*, 5 – *Sheep*, 6 – *Poultry*. We have extracted around 60K video frames of [640 x 640] size from the videos, inserting them in the database. Around 200K animal objects have been detected in the frames by using DL detectors and annotated with 2D bounding boxes. Their coordinates have been represented in the normalized [*x-center*, *y-center*, *width*, *height class number*] format and an annotation file containing the coordinates of the bounding boxes has been assigned to each frame.

Each of these three annotated image databases has been split into the next three datasets: the training set, containing 70% of the data, the validation set, which contains 10% of image data and the testing set, with 20% of the data. Data augmentations have been performed to increase the variety of the training and validation datasets, and model accuracy, and to avoid the over-fitting. Therefore, we have applied some random transformations, such as rotations, to those images and their bounding boxes.

We have considered You-Only-Look-Once (YOLO) – v9 deep neural networks for all traffic participant detection tasks [14]. We have set their Input Layer to the [400 × 400 × 3] dimension, so a resizing has been applied to the augmented training and validation images. The YOLO models have been trained, validated and tested on the three datasets of each database by applying various training and validation options.

Thus, for the vehicle detection task we have applied a stochastic gradient descent (SGD) algorithm with momentum, 45 training epochs, a mini-batch size value of 16 and an initial learning rate of 1e-3. A network training accuracy of 95%, a validation accuracy of 93.5% and a testing accuracy of 90% have been achieved.

The obtained vehicle detector is then applied on the analyzed traffic video sequence and the vehicles in the video are located successfully. Next, these detections are recognized as *Car*, *Bus*, *Truck*, *Tram* or *Bike* by applying a GoogLeNet – based classification model developed by us [18].

In the pedestrian case we have used the following options: the SGDM algorithm for optimization, 40 training epochs, mini-batch size value = 8, initial learning rate = 3e-4, validation frequency = 30 and validation patience = 6. One has obtained a training accuracy of 94.5%, a validation accuracy of 93% and a testing accuracy of 91%. The pedestrians are properly determined by the obtained DL-based detector, but no classification process is performed in their case.

For the animal detection task the used training options are: SGDM optimization, 50 epochs, mini-batch size value = 4 and initial learning rate = 0.001. The validation options applied here are: validation frequency = 30 and validation patience = 6. High accuracies are achieved for training – 94%, validation – 92% and testing – 91%. The generated YOLOv9-based detector produces also an animal classification output. Thus, the resulted animal detections are also recognized by their classes, which are labeled from 0 to 6.

Let T_j^i represent the traffic participant detections in the video frame sequence $V = [F_1, \ldots, F_M]$, where j = 1, ... , n_i and n_i is the number of detections in the frame F_1. The proposed CNN-based detection locates successfully both static and moving

participants in both fixed- and moving-camera videos and also detects properly the partially occluded ones.

3. MULTI-SCALE ANALYSIS-BASED TRACKING BY DETECTION TECHNIQUE

A tracking-by-detection (TBD) methodology has been developed to address the task of counting traffic participants. The proposed tracking solution takes into consideration both the motion and the visual aspect of the participants. Their aspect is treated here through a multi-scale analysis based on the next nonlinear fourth-order nonlinear diffusion model:

$$\begin{cases} \frac{\partial u}{\partial t} + \alpha\delta(|\Delta u|)\Delta(\varphi(\|\nabla u_\sigma\|)\nabla^2 u) - \beta(u - u_0) = 0 \\ u(x, y, 0) = u_0(x, y), \quad \forall (x, y) \in \Omega \\ u(x, y, t) = 0, \ \forall (x, y) \in \partial\Omega \end{cases}, \tag{1}$$

where $\alpha \in (1,2), \beta \in 0{,}0.5$, the diffusivity function $\varphi(s) = \varepsilon^{r+1}\sqrt{\frac{\lambda}{|\eta s^r + \xi|}}$, the function $\delta(s) = \lambda\sqrt[r]{\gamma s^{r+1} + \nu}$, where $\gamma \in (0,1)$ and $\lambda, \xi, \varepsilon, r, \nu \in [1,4]$ [19, 20].

This nonlinear PDE model is non-variational, since it cannot be derived through energy functional minimization, and well-posed, since it admits a weak solution that is computed numerically applying a finite difference-based approximation algorithm [19-21]. Let u^n represent the image filtered after n steps of the numerical scheme that is next used to create the scale-space. Since one requires a scale-space for RGB objects and R, G, B layers are correlated, a participant detection T is converted to the decorrelated color space CIELAB, the obtained conversion being $[L(T), a(T), b(T)]$. So, one obtained the next multi-scale space for T:

$$Sp(T) = \left\{T, RGB\left(\left[\left(L(T)\right)^l, a(T), b(T)\right]\right), \ldots, RGB\left(\left[\left(L(T)\right)^{lK}, a(T), b(T)\right]\right)\right\} \tag{2}$$

where $\left(L(P)\right)^l$ is the luminance filtered by $l \in [2{,}10]$ steps of the numerical solving algorithm of (1) and K > 2.

A machine – and deep learning – based feature extraction is applied to T at each scale k from 0 to K on the image $Sp(T)[l] = RGB\left(\left[\left(L(T)\right)^{lk}, a(T), b(T)\right]\right)$. So, we extract the high-level characteristics from prediction layer of an Inception-ResNet-V2 network [22]. Current image is pre-processed based on the specifications of the input layer of this CNN. It is resized to [299 × 299 × 3] format and then normalized. Batch size value is set to 32 and the pre-processed image is fed into the Inception-ResNet-V2 model. Prediction layer generates an activation that produces a feature

vector of 1000 coefficients. We concatenate it to a HOG-based feature vector with 81 coefficients, determined by combining 9 histograms with 9 bins, and a LBP feature vector of 59 coefficients [23], also determined by us, and obtain the vector $V(Sp(T)[l])$. These features vectors computed at multiple scales are concatenated next into the final $[(K+1)\times 1140]$ feature vector of T: $V(T)$: = $[V(Sp(T)[0]); \ldots; V(Sp(T)[K])]$.

First, the instance matching algorithm labels each detection in the first frame with its index value. Then, for each F_i, one identifies for each T_j^i the closest unlabeled T_ρ^{i+1} from the same class (if any), in the next frame, as follows:

$$\rho = arg \min_{k\in\{1,\ldots,n_{i+1}\}} D\left(T_j^i, T_k^{i+1}\right) | d\left(V\left(T_j^i\right), V\left(T_k^{i+1}\right)\right) \le t_1, \frac{h\left(T_j^i\right)}{h\left(T_k^{i+1}\right)} \in \left[\frac{1}{t_2}, t_2\right], \frac{w\left(T_j^i\right)}{w\left(T_k^{i+1}\right)} \in \left[\frac{1}{t_3}, t_3\right] \quad (3)$$

Where t_1 and $t_3 > t_2 > 2$ are properly selected thresholds (for example, some good values are $t_1 = 35$, $t_2 = 2.5$ and $t_3 = 3$), h and w return the height and width of the argument, d represents the Euclidean distance and the proposed *Intersection over Union* (IoU) – based metric [24] has the form:

$$D\left(T_j^i, T_k^{i+1}\right) = 1 - IoU\left(T_j^i, T_k^{i+1}\right) = 1 - \frac{\left|T_j^i \cap T_k^{i+1}\right|}{\left|T_j^i \cup T_k^{i+1}\right|} \quad (4)$$

If $D\left(T_j^i, T_k^{i+1}\right) < 1$, then there exists an overlap between these detections and T_ρ^{i+1} represents the next instance and receives the track label of T_j^i. If D = 1, then one searches for the next state in a neighborhood of T_j^i , as following:

$$\rho = arg \min_{k\in\{1,\ldots,n_{i+1}\}} d\left(V\left(T_j^i\right), V\left(T_k^{i+1}\right)\right) | d_c\left(T_j^i, T_k^{i+1}\right) \le t_4, \frac{h\left(T_j^i\right)}{h\left(T_k^{i+1}\right)} \in \left[\frac{1}{t_2}, t_2\right], \frac{w\left(T_j^i\right)}{w\left(T_k^{i+1}\right)} \in \left[\frac{1}{t_3}, t_3\right] \quad (5)$$

where $d\left(V\left(T_j^i\right), V\left(T_k^{i+1}\right)\right) \le t_1$, d_c computes the distance between the detections' centroids and the threshold value $t_4 = \frac{w(F_i)}{2}$. If no such index ρ can be determined from (5), then the trajectory of T_j^i ends in F_i, which means that participant has exited the traffic scene. When all the tracks of traffic participants in F_i have determined, the unlabeled detections from the next frame are considered new participants and labeled with the next available numbers.

4. EXPERIMENTS AND METHOD COMPARISON

The detection, recognition and tracking system described here has been implemented by using Python 3.11 and Matlab R2024b on a processor Intel Core I7 12700 2.1 GHz, Turbo mode 4.9 GHz, 32 GB RAM, SSD: EMTECX 150 240 GB, and then applied successfully on numerous traffic video datasets. This framework has been experimented on some free available video traffic sequences from Pond5

[25], which describe pedestrians, vehicles and animals on the roads and contain more than 11K frames.

Its CNN-based motion-insensitive detection component represents a much better solution than the motion-sensitive detectors, such as those based on temporal differencing, optical flow or GMM, because it can detect both static and moving participants in both fixed and moving-camera video sequences. It also locates efficiently the partially-occluded traffic participants, thus working well for the crowded traffic scenes. The proposed TBD technique provides effective pedestrian tracking results but depends much on the pedestrian detections. Also, it tracks the partially-occluded subjects successfully, but fails to count the fully-occluded ones. That means, if a participant disappears because of a full occlusion and then reappears in the scene, the TBD algorithm creates a new track for it.

Our detection and tracking approach outperforms many machine learning and deep learning-based techniques, achieving higher performance scores, as illustrated by the method comparison results presented in Table 1. For detection and tracking, *Precision* is the percentage of the current trajectory that is right and *Recall* provides the percent of the ground-truth trajectory overlapping with that trajectory. We have tested also this YOLO-v9 + TBD approach for K = 0 and achieved some lower metric scores. This demonstrates the usefulness of the proposed PDE-based multi-scale analysis.

Table 1
Detection and tracking method comparison results

Technique	**Precision**	**Recall**
The proposed YOLO-v9 + TBD framework	0.8942	0.8795
YOLO-v9 + TBD without multi-scale analysis	0.8672	0.8533
YOLO–X + DeepSORT	0.8694	0.8677
Faster–R CNN + DeepSORT	0.8562	0.8435
ACF + IoU tracker	0.8231	0.8057
AdaBoost + HOG	0.7985	0.7866
HOG-SVM + Kalman filter	0.7547	0.7452
GMM + Kalman filtering	0.6971	0.6873

Figure 8. Traffic participant detection, recognition and tracking example.

Some results are also shown in Fig. 1, where a traffic participant detection and tracking example is described. An animal, a pedestrian and two vehicles are detected, recognized by class if the case and then tracked properly in a video sequence. Their states in the 1st, 10^{th}, 20^{th} and 30^{th} video frames, marked by bounding boxes and labeled with class and count labels, are displayed in the images a) – d).

5. CONCLUSIONS

A multiple traffic participant detection, classification and tracking technique that combines successfully nonlinear diffusion-based models to machine and deep learning algorithms was described here. The diffusion models were used successfully for the image preprocessing tasks and the multi-scale analysis-based feature extraction.

The motion-insensitive CNN-based detectors and the three voluminous databases represent important contributions of our research, but the major novelties are proposed in the tracking stage. Our TBD approach, modeling the visual appearance by using a nonlinear diffusion-based multi-scale analysis, and the participants' motion by applying some properly chosen metrics, outperforms some well-known ML - and DL - based counting methods. The novel well-posed fourth-order nonlinear diffusion-based model that is used for creating the multi-scale space represents an important contribution of this work.

Our future research will focus on solving some drawbacks of this system, such as that related to the full occlusions. Since this research is part of the computer vision project that is acknowledged below, we also intend to improve and include this detection, recognition and counting framework into a larger traffic monitoring system. The detection and tracking results achieved here could lead successfully to the recognition of some actions of the road traffic participants and some interactions between them, thus contributing to the road collision prevention and increasing the safety of all participants. They could also be applied in domains like biometric systems [26], autonomous vehicles [9] and robot vision.

Acknowledgments. This research work was supported by a grant of the Romanian Academy, GAR-2023, Project Code 19.

REFERENCES

1. CAMERO A., ALBA E., *Smart City and information technology: A review*, Cities, 2019, 93, pp. 84–94.
2. SENGAR S., MUKHOPADHYAY S., *A novel method for moving object detection based on block based frame differencing*, 3rd International Conference on Recent Advances in Information Technology (RAIT), March 2016, pp. 467–472. IEEE.
3. BAKTI R. Y., ARENI I. S., *Vehicle detection and tracking using gaussian mixture model and kalman filter*, 2016 International Conference on Computational Intelligence and Cybernetics, 2016, pp. 115–119. IEEE
4. RAMZAN H., FATIMA B., SHAHID A. R., ZIAUDDIN S., SAFI A., *Intelligent pedestrian detection using optical flow and hog*, International Journal of Advanced Computer Science and Applications, 2016, 7(9).
5. BURGHARDT T., CALIC J., *Real-time Face Detection and Tracking of Animals*, 8th Seminar on Neural Network Applications in Electrical Engineering, Belgrade, Serbia, 2006, pp. 27–32.
6. DALAL N., TRIGGS B., *Histograms of oriented gradients for human detection*, IEEE computer society conference on computer vision and pattern recognition (CVPR'05), June 2005, Vol. 1, pp. 886–893.
7. FEIFEI J., AIBO Z., SHENGKE W., *Pedestrian detection based on active contour models*, 3rd Intl. Symp. on Inf. Proc., 2010, pp. 121–125.
8. HUA J., SHI Y., XIE C., ZHANG H., ZHANG J., *Pedestrian-and vehicle-detection algorithm based on improved aggregated channel features*, IEEE Access, 2021, 9, pp. 25885–25897.
9. GALVAO L. G., ABBOD M., KALGANOVA T., PALADE V., HUDA, M. N., *Pedestrian and vehicle detection in autonomous vehicle perception systems—A review*. Sensors, 2021, *21*(21), 7267.

10. SAHI A., TALELE K., *Person tracking using mean shift with Gray Level Grouping*, International Conference on Communication, Information & Computing Technology – ICCICT, 2015, pp. 1–5. IEEE
11. CHANTARA W., MUN J. H., SHIN D. W., HO Y. S., *Object tracking using adaptive template matching*, IEIE Transactions on Smart Processing and Computing, 2015, 4(1), pp. 1–9.
12. NIU X., *et al*. *WTrack: HMM-based walk pattern recognition and indoor pedestrian tracking using phone inertial sensors*, Personal and ubiquitous computing, 2014, 18, pp. 1901–1915.
13. KALAFATIC Z., RIBARIC S., STANISAVLJEVIC V., *A system for tracking laboratory animals based on optical flow and active contours,* Proceedings 11th International Conference on Image Analysis and Processing, 2001, September, pp. 334–339. IEEE.
14. DJARAH D. *et al*. *Online Multi-object Tracking with YOLOv9 and DeepSORT Optimized by Optical Flow*, Engineering, Technology & Applied Science Research, 2024, 14(6), pp. 17922–17930.
15. SIMPATIA project: http://iit.academiaromana-is.ro/grantSIMPATIA.html – last accessed on 19.11.2025.
16. BARBU T., MOROȘANU C., *Spatio-temporal Video Restoration Technique using a 3D Anisotropic Diffusion-based Scheme*, Bulletin of the Transilvania University of Brasov. Series III, 2025, Vol. 5(67), No. 2, pp. 211–220.
17. BARBU T., MOROȘANU C., *Nonlinear Reaction-Diffusion Based Video Restoration Technique for Noise Mixtures*, *IEEE 19th Conf. on Industrial Electronics and Appl. (ICIEA)*, Kristiansand, Norway, 2024, pp. 1–4.
18. BARBU T., BEJINARIU, S.-I., LUCA, R*., Transfer Learning-based Framework for Automatic Vehicle Detection, Recognition and Tracking*, 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2024, pp. 1–6. IEEE.
19. BARBU T., *Digital Image Processing, Analysis and Computer Vision Using Nonlinear Partial Differential Equations*, Springer Nature, 2025.
20. BARBU T., *Deep Learning-based Multiple Moving Vehicle Detection and Tracking using a Nonlinear Fourth-order Reaction-Diffusion based Multi-scale Video Object Analysis*, Discrete & Continuous Dynamical Systems – Series S, AIMS Journals, 2023, 16 (1), pp. 6–32.
21. BOOLE G., *Calculus of finite differences*, BoD–Books on Demand, 2022.
22. THOMAS A., *et al*. *Moving vehicle candidate recognition and classification using inception-resnet-v2*, 44th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, 2020, pp. 467–472.
23. ZHOU W., *Pedestrian detection with improved LBP and hog algorithm*, Open Access Library Journal, 2018, 5(04), 1.
24. H. REZATOFIGHI *et al*., *Generalized intersection over union: A metric and a loss for bounding box regression*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.
25. POND5 – https://www.pond5.com.
26. BARBU T., CIOBANU, A., LUCA, M*., Multimodal biometric authentication based on voice, face and iris*, *2015 E-Health and Bioengineering Conference (EHB)*, Iasi, Romania, pp. 1–4, 19–21 Nov. 2015.