ROMAI J., v.21, no.2(2025), 21–35

MULTIPLE ANIMAL DETECTION, RECOGNITION, TRACKING AND BEHAVIOR ANALYSIS FRAMEWORK COMBINING DEEP NETWORKS TO REACTION-DIFFUSION MODELS

Tudor Barbu[1,2], Silviu-Ioan Bejinariu[1], Costică Moroșanu[3]

[1]*Institute of Computer Science of the Romanian Academy - Iași Branch*

[2]*Academy of the Romanian Scientists, București, Romania*

[3]*Alexandru Ioan Cuza" University of Iasi, Iasi, Romania*

tudor.barbu@iit.academiaromana-is.ro, silviu.bejinariu@iit.academiaromana-is.ro, costica.morosanu@uaic.ro

Abstract A novel computer vision framework for the detection, classification and counting of the animals on roads is introduced in this research work, which is part of a traffic monitoring project and combines successfully some deep learning and mathematical models. A CNN-based animal detection and classification technique using a voluminous annotated animal database for training, validation and testing is proposed first. Then, a tracking by detection algorithm using a nonlinear reaction-diffusion based multi-scale analysis of the animal detections is introduced for the animal counting task. An analysis of the animal movements in the traffic sequence is also provided. Thus, one performs a transfer learning-based video semantic segmentation to identify the main components of the traffic scene, representing roadways, sidewalks and crosswalks, and the animal tracks that intersect them. The performed experiments and the method comparison results are finally discussed in this paper.

Keywords: annotated animal database, CNN-based multi-animal detection and classification, tracking by detection algorithm, nonlinear diffusion-based multi-scale analysis, crosswalk detection, semantic video segmentation, road crossing behavior analysis.

2020 MSC: 35Axx, 35Q68, 65D19, 68Txx, 68T07, 97U99.

1. INTRODUCTION

The growth of the vehicle and pedestrian traffic has generated an increasing number of road accidents in recent years. For this reason, there has been a considerable amount of interest in the last years in developing some smart city technologies that improve the safety of the traffic participants [1].

These traffic participants represent vehicles, pedestrians and animals. The detection, tracking, classification and behavior analysis of these participants represent essential tasks for those smart city technologies [1]. Their application fields include the video surveillance systems, law enforcement, biometrics,

human–computer interaction, robot vision, autonomous vehicles, action recognition and augmented reality.

A high amount of research has been performed in the vehicle and pedestrian detection and tracking domains in the last decades [2]. We also developed many detection, recognition and tracking techniques for both pedestrians [3, 4, 5] and vehicles [6, 7, 8]. Somewhat lesser research has been conducted on the monitoring of the animals on roads, although their presence in the traffic areas is quite concerning and could generate serious traffic events.

The existing animal detection techniques use background subtraction [9], power spectrum [10], template matching [11], optical flow [12], adaptive boosted classifiers with Haar features [13], Support Vector Machines (SVM) with Histograms of Oriented Gradients (HOG) [14], and Deep Learning models [15]. Animal tracking (counting) approaches are based on mean-shift algorithms [16], Kalman filtering [17], optical flow estimation and active contours [18], convolutional neural networks (CNN) [19] and other tools used for multi-object tracking [20].

This research, which is performed within a road traffic monitoring project, introduces a novel automatic framework for multiple animal detection, recognition and tracking that combines successfully deep learning and PDE models. Its CNN-based animal detection and recognition component is described first, in the next section. The proposed detector and classifier uses a voluminous animal image database created and annotated by us.

The animal tracking process is then performed by applying a novel tracking by detection (TBD) technique that treats both the appearance and motion of the animal detections. Their appearance is modeled applying a multi-scale analysis-based high-level feature extraction on them using a nonlinear reaction-diffusion equation based scale-space and a deep neural network. Their motion is treated using an Intersection over Union (IoU)-based metric and the distances between the centroids. Some bounding box size related conditions are also involved in the TBD algorithm that is detailed in the third section.

Numerous traffic accidents result from domestic and wild animals unexpectedly crossing paths with vehicles on the roads. So, an animal behavior analysis [21] is considered in the fourth section. One performs also a transfer learning-based video semantic segmentation, to identify the main regions of the traffic scene: roadways, sidewalks and crosswalks. The animal detections and trajectories that intersect these zones are determined next. Thus, some walking behaviors of these animals could be modeled. The performed experiments and method comparison results are disscused in the fifth section and the conclusions of this work and the future research plans are outlined in the last section.

2. DEEP LEARNING-BASED MULTI-ANIMAL DETECTION AND CLASSIFICATION METHOD

An automatic deep learning-based multi-animal detection and recognition technique has been developed. For this purpose, a voluminous image database containing several common classes of domestic and wild animals seen on or near roads was created and annotated by us. That database, which is available at [22], was obtained from 30 videos recorded by using a high-resolution video camera, at a recording speed of 30 fps. Since the videos could be contaminated by multiple noises during the acquiring process, we have applied our own spatio-temporal 3D anisotropic diffusion-based video filters for additive white Gaussian noise (AWGN) and mixed noise [23, 24].

Then, we extracted around 60K video frames of $[640 \times 640]$ dimension from the filtered sequences and inserted them in the database. Around 150K objects representing multi-view animals were detected in those frames by using some DL-based detectors and then annotated with 2D bounding boxes. The coordinates of these boxes have been represented in the normalized format, which is [*x-center, y-center, width, height, class number*], and each video frame has been assigned an annotation file containing the bounding box coordinates of its animal objects. We have considered 7 classes of domestic and wild animals that are more often encountered on our country roads and labeled them with the following *class number* values: $0 = Dog$, $1 = Cat$, $2 = Bear$, $3 = Cattle$, $4 = Horse$, $5 = Sheep$ and $6 = Poultry$.

Next, the annotated animal database has been split into the following three image datasets: the **training set** that contains approximately 70% of the data (almost 42K images), the **validation set** comprising 10% of data (around 6K images) and the **testing set** including 20% of this data (around 12K images). The training and validation datasets are then pre-processed. Thus, a data augmentation process has performed to increase the variety of the training and validation datasets and the detector accuracy and to avoid the over-fitting. It applies some random transformations, such as rotations, to the images from the two datasets and their bounding boxes.

Then, a convolutional neural network has been applied to the three animal image datasets for the detection and recognition tasks. We have considered a You-Only-Look-Once (YOLO)-v9 deep neural network for these processes [25].

The Input Layer of this YOLO network has been set to the $[400 \times 400 \times 3]$ format. The pre-processed images have been fed into this CNN. It has been trained and validated and tested on the two datasets by applying the following trainining and validation options: the Stochastic Gradient Descent (SGD) optimization algorithm with momentum, 50 training epochs, the minimum

batch size value = 8, initial learning rate = 0.001, validation frequency = 30 and validation patience = 6. It has achieved a training accuracy of 93.5% and a validation accuracy of 92%. It has been also tested on the testing set of the database and a testing accuracy of 91% has been obtained.

An effective animal detector and recognizer is generated by this YOLO v9-based training and validation process. It has a motion-insensitive character and locates and classifies successfully both the static and dynamic *animal objects*, in both fixed-camera and moving-camera videos. Let A_j^i represent those animal detections, with the class labels $c(A_j^i) \in \{0, \ldots, 6\}$, in the video frame sequence $V = [F_1, \ldots, F_N]$, with $i = 1, ..., n_i$ and n_i is the number of detections in the frame F_i. An example of animal detection and classification result achieved by our DL-based technique is described in Fig. 1, where 2 animals are detected in a traffic video frame and labeled as 1 (*Cat*) and 0 (*Dog*).

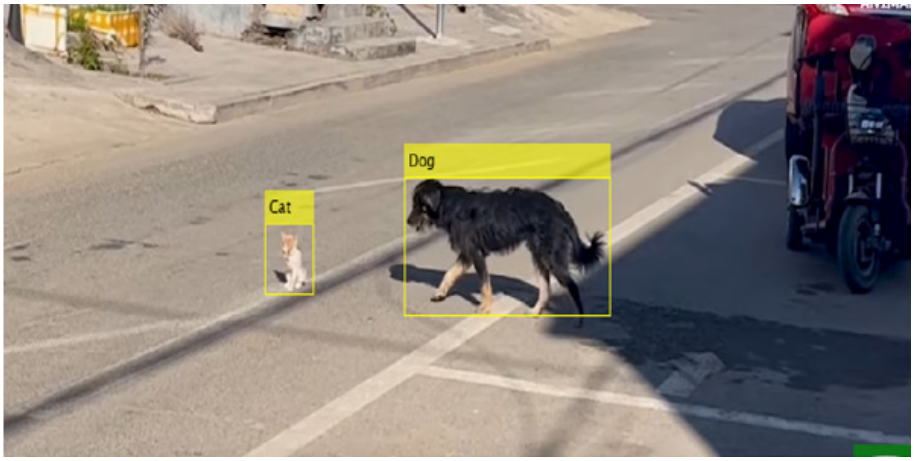

Fig. 1. YOLO-based animal detection and recognition example

3. AUTOMATIC ANIMAL TRACKING BY DETECTION APPROACH

A novel tracking-by-detection (TBD) technique has been developed for the animal counting process. This animal tracking algorithm treats both the motion and appearance of the obtained and classified detections. Their appearance has been modeled by applying a high-level multi-scale feature extraction scheme that combines partial differential equations (PDE) to deep neural networks. The scale-space representation has been created by using the following nonlinear second-order reaction-diffusion model [5, 26]:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \alpha\varphi(|\nabla^2 u|)\nabla \cdot (\psi(\|\nabla u_\sigma\|)\nabla u) + \beta(u - u_0) = 0 \\ u(x, y, 0) = u_0(x, y),\ \forall (x, y) \in \Omega \\ u(x, y, t) = 0,\ \forall (x, y) \in \partial\Omega,\ t \in (0, T] \end{cases} \tag{1}$$

where the evolving function u corresponds to the subimage of the analyzed animal object, $\Omega \subseteq R^2$, $\alpha, \beta \in [0.4, 2]$, $u_\sigma = u^* G_\sigma$, 2D Gaussian filter kernel $G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$, its edge-stopping function $\psi(s) = \xi \sqrt[r+1]{\frac{\lambda}{|\eta s^r + \varepsilon|}}$, with $\lambda, \xi, \varepsilon, r \in (1, 4]$ and $\eta \in (0, 1]$ and $\varphi(s) = \lambda(\gamma s^{r+1} + \nu)^{\frac{1}{r}}$, with $\gamma \in (0, 1]$ and $\nu \in (1, 4]$.

This nonlinear PDE-based model is non-variational, since it cannot be obtained from an energy functional minimization process. It is also valid, or well-posed, its weak (variational) solution being determined numerically applying a finite difference method-based approximation algorithm [27]. Let u^n represent the state of the evolving image after applying n steps of the numerical scheme that solves the reaction-diffusion model (1). This model works for 2D image functions u only and we need a multi-scale space for RGB animal objects. Since R, G, B channels could be highly correlated. Therefore, an animal detection A is converted to the de-correlated color space CIELAB, its converted form $[l(A), a(A), b(A)]$ being obtained. Thus, the next scale-space representation is generated for A:

$$S_c(A) = \{A, RGB([(L(A))^n, a(A), b(A)]), ..., RGB([(L(A))^{nK}, a(A), b(A)])\} \tag{2}$$

where the number of scales $K \geq 2$, $n \in [2, 8]$ and $RGB(\ )$ converts the argument to the RGB form.

A CNN-based feature extraction process has been applied at each scale k from 0 to K on the image object $S_c(A)[n] = RGB([(L(A))^{nk}, a(A), b(A)])$. Thus, one has extracted the high-level characteristics from a deep layer of DenseNet-201 [28], which is Fully Connected Layer of 1000 neurons (FC 1000). The current image has been pre-processed by using the specifications of the input layer of DenseNet-201 model. It has been resized to the $[224 \times 224 \times 3]$ format and then normalized. The minimum batch size parameter has been set to 32 and the pre-processed animal object has been fed into the DenseNet-201 network. Its FC 1000 layer has generated an activation that produced a feature vector of 1000 coefficients, $V(S_c(A)[n])$. These feature vectors determined at multiple scales have been concatenated into the final $[(K + 1) \times 1000]$ feature vector of $A : V(A) := [V(S_c(A)[0]); \ldots; V(S(A)[K])$.

The proposed TBD algorithm combines the distances between these feature vectors characterizing the animal appearance to some measures related to the animals' motion and sizes. Thus, each animal detection in the first frame is labeled with its index. For each F_i, one searches for each A_j^i the unlabeled

A_τ^{i+1} of the same class in the next frame, that is closest to A_j^i in terms of image location, appearance and size. Its index is determined as follows:

$$\begin{aligned} \tau &= \arg\min_{k\in\{1,\ldots,n_{i+1}\}} d_{IoU}(A_j^i, A_k^{i+1}) \mid c(A_j^i) \\ &= c(A_k^{i+1}), d(V(A_j^i), V(A_k^{i+1})) \leq \kappa, \\ &\frac{h(A_j^i)}{h(A_k^{i+1})} \in \left[\frac{1}{T_h}, T_h\right], \ \frac{w(A_j^i)}{w(A_k^{i+1})} \in \left[\frac{1}{T_w}, T_w\right] \end{aligned} \tag{3}$$

where d is the Euclidean distance, κ, T_h, T_w are some properly selected thresholds, $T_w > T_h > 2$, and Intersection over Union (IoU)-based metric $d_{IoU}(A_j^i, A_k^{i+1})=1 - IoU(A_j^i, A_k^{i+1})=1 - \frac{|A_j^i \cap A_k^{i+1}|}{|A_j^i \cup A_k^{i+1}|}$ [29]. If $d_{IoU}(A_j^i, A_k^{i+1}) < 1$, then the 2 detections overlap, so A_τ^{i+1} is the next state of A_j^i, and receives its track label. Otherwise, next instance is found in a neighborhood of radius R as follows:

$$\begin{aligned} \tau &= \arg\min_{k\in\{1,\ldots,n_{i+1}\}} d(V(A_j^i), V(A_k^{i+1})) \mid c(A_j^i) \\ &= c(A_k^{i+1}), d_c(A_j^i, A_k^{i+1}) \leq R, \\ &\frac{h(A_j^i)}{h(A_k^{i+1})} \in \left[\frac{1}{T_h}, T_h\right], \ \frac{w(A_j^i)}{w(A_k^{i+1})} \in \left[\frac{1}{T_w}, T_w\right] \end{aligned} \tag{4}$$

where $d(V(A_j^i), V(A_k^{i+1})) \leq \kappa$ and d_c returns the distance between their centroids. If there is no such τ, then the trajectory of A_j^i ends in F_i (animal exited the scene). When the tracks of all animals in F_i have been found, the unlabeled detections of F_{i+1} are considered new animals and receive the next available labels. An animal counting example obtained by applying this TBD approach and describing 2 animal tracks is presented in Fig. 2.

4. ANIMAL BEHAVIOR ANALYSIS

An animal behavior analysis that is mainly related to road crossing has been performed next. It uses a semantic video segmentation that divide the frames in 3 categories of regions: *sidewalks, crosswalks* and *roadways*. The animal trajectories generated by the tracking-by-detection process were analyzed using the obtained segments.

A deep learning-based crosswalk detection approach has been applied first [30, 31]. We have built a sufficiently-large crosswalk dataset for this purpose by using 35K zebra crossing images obtained from Roboflow Universe collection. Those crosswalk images have been filtered and then annotated with 2D bounding boxes represented in the normalized form [*x-center, y-center, width, height*] and a file containing the bounding box coordinates of has been assigned to each image. This annotated crosswalk dataset has been split into a training

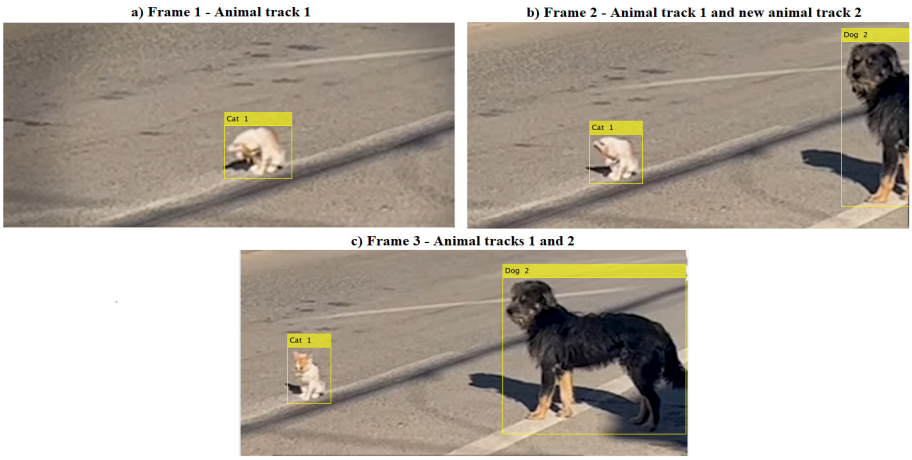

Fig. 2. Animal tracking by detection example

set containing 80% of the image data, and validation and testing sets, each of them comprising 10% of data. Training and validation datasets have been pre-processed applying some data augmentation processes.

One has used a YOLO v9 network in this case, too. First, an Input Layer of $[256 \times 256 \times 3]$ dimension has been created for it. This CNN has been applied to the three zebra image datasets. It has been trained successfully on the training dataset with the following options: a SGDM algorithm for optimization, 40 training epochs, batch size value = 4 and initial learning rate = $3e^{-4}$. It has obtained a high training accuracy of 93.5%. It has been also validated on the validation set with the options: validation frequency = 50, validation patience = 7. A validation accuracy of 92% has been achieved. This YOLO v9 model has got a testing accuracy of 90%. The resulted motion-insensitive CNN-based zebra image detector locates properly the cross-walks in any video sequence, labeling them accordingly. Let the n^i zebra images detected in the frame F_i be Z_j^i, $j = 1, ..., n^i$. A crosswalk detection example is provided in Fig. 3, which describes a detected animal, recognized as *Dog*, crossing the street on a detected zebra.

A roadway detection has been performed next, by applying a transfer learning-based semantic segmentation technique [32, 33, 34]. An annotated video image dataset based on the Cambridge-driving Labeled Video Database (CamVid) [35], has been used here. The dataset is composed of 701 images and their associated ground-truth (GT) labels and it has been divided into 3 subsets: training dataset, composed of 575 images, validation dataset with 61 images and testing dataset containing 65 images.

Fig. 3. Example of animal detected on crosswalk

While CamVid provides GT labels for 32 semantic classes, but we are interested in two classes only: *Road* and *Not-road*. The number of classes has been reduced to 2, by merging multiple classes and their corresponding labels.

Then, we have assembled a MobileNet-v2 network with 154 layers, whose weights have been trained on the ImageNet collection and applied it in the transfer learning process [36]. Thus, a DeepLab-v3+ neural network has been trained with the weights initialized from thid pre-trained MobileNet-v2 model. The Input Layer of this semantic network has been set to the $[720 \times 960 \times 3]$ format. This DeepLab-v3+ neural network has been trained and validated successfully on the corresponding subsets using the next options: SGDM optimization, 30 epochs, minimum batch size = 2, L2 Regularization hyperparameter = 0.005, initial learning rate = 1e-2, validation patience = 5 and validation frequency = 30.

The generated CNN-based segmenter has been applied succesfully on the frames of V. The labeled segmentation result has been represented as a binary image, whose connected components correspond to the drivable regions. Some morphological operations have been applied to binary image next to improve the road detection [37]. Also, the connected components with low areas or inappropriate shapes have been discarded. Let the obtained road regions in F_i be noted R_j^i, $j = 1, ..., n(i)$.

We have determined three main traffic areas by using these detection and segmentation results: *safe road, unsafe road* and *no road*. The *safe road* area

of the frame F_i is determined as the union of all crosswalks: $SR := \bigcup_{j=1}^{n^i} Z_j^i$, $\forall i = 1, \ldots, K$. The *unsafe road* area represents the set of the roadway pixels outside the safe road area: $UR_i := \left(\bigcup_{j=1}^{n(i)} \right) \setminus SR_i$, $\forall i = 1, \ldots, K$. The *no road* is determined as area outside both safe and unsafe regions:

$$NR_i := F_i \setminus NR_i := F_i \setminus \left[SR_i \cup \left(\bigcup_{j=1}^{n(i)} \right) \right], \forall i = 1, \ldots, K.$$

The moving behavior of an animal in the traffic scene has been modeled by investigating the overlappings between the states of its trajectory and those 3 areas. Since we have been interested in the lower part of the animal body, which contains the feet, only the bottom third part of each detection A_j^i has been considered. We denoted it $A_j^i/3$ and determined its overlaps with the three areas. That corresponding to the largest overlapping segment provides the instance label:

$$l(A_j^i) := l \left(\arg \max_{Area \in \{SR_i, UR_i, NR_i\}} \left\{ \frac{A_j^i}{3} \cap Area \right\}, \ \forall i \in 1, ..., K \right), \quad (5)$$

with $l(SR_i)$ = *safe road*, $l(UR_i)$ = *unsafe*, $l(NR_i)$ = *no road*. The animal behavior in the traffic video is represented by the sequence of all labels in its track: $[l(A_{j_1}^1), \ldots, l(A_{j_K}^K)]$, where the indices $j_i \in \{1, ..., n_i\}$ are computed by the TBD algorithm in the previous section. This label sequence shows if an animal has a right moving in traffic or if the safe movements alternate with the unsafe ones.

An animal labeling example based on this behavior modelling is provided in Fig. 4. Three animals are detected in the frame displayed in (a), recognized as dogs, and labeled as *no road* and *unsafe*, based on the semantic frame segmentation result described in (b).

5. EXPERIMENTS AND METHOD COMPARISON

The presented detection, recognition, tracking and behavior analysis technique was implemented soft using Python 3.13.11 and MATLAB R2024b on a processor Intel Core I7 12700 2.1 GHz, Turbo mode 4.9 GHz, 32 GB (2*16GB DDR4) RAM, SSD: EMTECX 150 240 GB, and then successfully experimented on multiple traffic video datasets. One achieved very good results representing animals that were properly located, classified and tracked, crosswalks that were detected correctly and properly labeled animal moving actions.

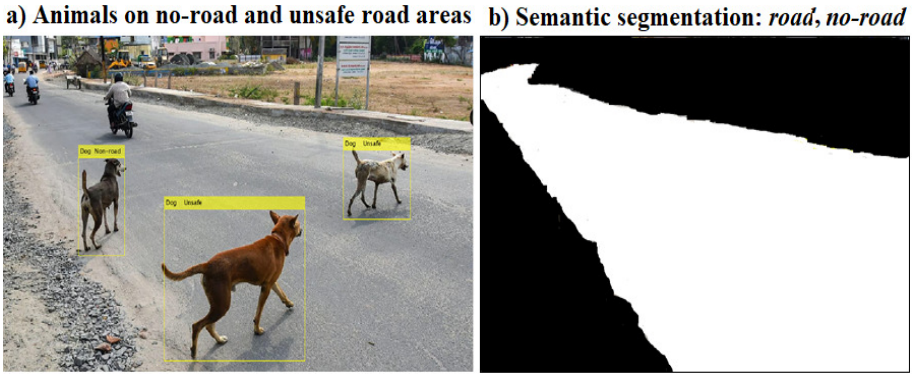

Fig. 4. Animal moving action labeling based on semantic segmentation

Each component of the framework gets successful results, outperforming other methods in its field. Thus, the YOLO-v9 based motion-insensitive detection component is a much better solution than motion-sensitive people detectors, like those based on background subtraction or optical flow, since it detects both the static and moving animals in both stationary- and moving-camera sequences and classifies them properly, too. This CNN-based detector also locates successfully the partially-occluded animals. It also outperform other machine and deep learning solutions when compared on the testing dataset of the animal database, as shown by the values of the performance metric scores in Table 1.

Table 1. Animal detection method comparison results

Detection Approach	**Precision**	**Recall**
This YOLO v9-based detector	0.9057	0.8964
YOLO-X detection	0.8824	0.8711
YOLO–v5	0.8692	0.8534
Faster R-CNN [38]	0.8277	0.8102
SVM+HOG	0.7665	0.7592
Cascade Classifier	0.7822	0.7731
Background subtraction	0.6529	0.6486

The proposed tracking-by detection approach gets successful animal counting results on right detections. Unlike the detector, it has a quite high com-

putational complexity and running time, given its DL- and PDE-based multi-scale feature extraction. Also, while this TBD tracks the highly-occluded animals quite well, it does not track properly the fully-occluded ones. If an animal disappears from the scene because of a full occlusion, our algorithm cannot count it anymore and considers it a new animal if it reappears. The TBD technique outperforms other machine and deep learning-based trackers when applied to the animal detections obtained by our CNN-based detector [8, 20], as illustrated by the animal tracking method comparison results presented in Table 2. However, it may execute slower than IoU and feature matching-based trackers, due to its higher complexity.

Table 2. Animal tracking method comparison

Method	**Precision**	**Recall**
The proposed TBD technique	0.8965	0.8863
Deep SORT algorithm	0.8953	0.8817
SORT algorithm	0.8723	0.8691
Kalman filtering	0.8477	0.8387
IoU tracker	0.7984	0.7871
Feature matching	0.7503	0.7422

The crosswalk detection component also achieves successful results, outperforming other approaches when simulated on the 3500 zebra images of the testing subset of our crosswalk dataset detailed in the third section. The *Precision* and *Recall* scores obtained by our CNN-based detector are higher than those of other ML- and DL-based techniques, as shown by the method comparison results in Table 3.

Table 3. Crosswalk detection method comparison results

Technique	**Precision**	**Recall**
The proposed CNN-based method	0.9141	0.9075
YOLO–X model	0.8865	0.8732
Faster R-CNN	0.8655	0.8431
Cascade classification	0.7433	0.7284

The proposed transfer learning-based semantic segmentation approach for roadway detection, which uses DeepLab-v3 and MobileNet-v2, also gets high performance scores. It outperforms U-Net segmentation network [32] and the transfer learning schemes using DeepLab with residual networks such as ResNet-101 and ResNet-50 [39], when compared on the same testing set of the labeled image dataset mentioned in previous section.

Table 4. Semantic segmentation method comparison

Approach	**Precision**	**Recall**
Our transfer learning-based method	0.8946	0.8795
U-Net model	0.8492	0.8336
DeepLab + ResNet-101	0.8834	0.8702
DeepLab + ResNet-50	0.8821	0.8673

These accurate detection, classification, tracking and semantic segmentation results have determined also some high animal moving action recognition rates, which are illustrated by the performance metric scores *Precision* = 0.8581 and *Recall* = 0.8497. The animal behavior analysis approach has been tested successfully on free available video traffic sequences from Pond5 [40], describing animal on roads and comprising more than 12K frames.

6. CONCLUSIONS

We described a new multiple animal detection, recognition, tracking and movement analysis technique that combines successfully the nonlinear diffusion-based models to deep neural networks in this article. The diffusion models were used first for pre-processing the animal and crosswalk databases introduced for the detection tasks, and then in the multi-scale feature extraction process used by the proposed tracking-by-detection algorithm. The novel reaction-diffusion based model used for the multi-scale analysis represents an important contribution of this research.

We have been applied here some state-of-the-art deep learning solutions for the detection and semantic segmentation tasks: YOLO-v9 network and DeepLab-v3+ with MobileNet-v2, respectively. The detection, counting and semantic segmentation approaches disseminated in this work outperforms other methods in their fields and provides effective results which are next combined by a behavior analysis model that was also presented here.

The analysis of the animal trajectories, which shows the safe and unsafe animal movements in traffic, could significantly improve the road safety by

preventing the vehicle collisions and the potential injuries to both pedestrians and animals. Also, the achieved results could be applied in important research fields, such as biometric authentication [41], robotic vision and autonomous vehicles.

Since this research is part of the below-acknowledged project in the traffic monitoring domain, we also intend to integrate this animal detection, classification, tracking and behavior analysis framework into a larger traffic monitoring system that includes also pedestrian and vehicle monitoring processes [5, 6, 8]. As part of our future research in this project, we intend also to improve this computer vision framework. Thus, we consider adding more animal classes to the detection and recognition component, investigating other PDE-based models for multi-scale analysis and reducing the computational complexity of the TBD technique.

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

References

[1] A. Khanna, R. Goyal, M. Verma, D. Joshi, *Intelligent traffic management system for smart cities*, International Conference on Futuristic Trends in Network and Communication Technologies, Singapore: Springer Singapore, 2018, 152-164.

[2] C. Premebida, G. Monteiro, U. Nunes, P. Peixoto, *A lidar and vision-based approach for pedestrian and vehicle detection and tracking*, In 2007 IEEE intelligent transportation systems conference, September 2007, 1044-1049. IEEE.

[3] T. Barbu, *Novel Approach for Moving Human Detection and Tracking in Static Camera Video Sequences*, Proceedings of the Romanian Academy, Series A, **13** (3) (2012), 269-277.

[4] T. Barbu, *Pedestrian detection and tracking using temporal differencing and HOG features*, Computers & Electrical Engineering, **40** (4) (2014), 1072–1079.

[5] T. Barbu, S.-I. Bejinariu, R. Luca, *Deep Learning-Based Pedestrian Detection and Tracking Technique Using Reaction-Diffusion Based Multi-Scale Analysis*, 2025 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 2025, 1-4.

[6] T. Barbu, S.-I. Bejinariu, R. Luca, *Transfer Learning-based Framework for Automatic Vehicle Detection, Recognition and Tracking*, 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Iasi, Romania, 2024, 1-6. IEEE.

[7] T. Barbu, *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, **16** (1) (2023), 6-32.

[8] T. Barbu, S.-I. Bejinariu, *CNN-based Moving Vehicle Recognition using GMM-based Foreground Modeling, Level-set based Segmentation and Kalman Filter-based Tracking*, 2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Sept. 2004, 1-6.

[9] L. Picek, L. Neumann, J. Matas, *Animal identification with independent foreground and background modeling*. In DAGM German Conference on Pattern Recognition, 2024, September, Cham: Springer Nature Switzerland, 2024, 241-257.

[10] A. B. Torralba, A. Oliva, *Statistics of natural image categories*, Network: Computation in Neural Systems, **14** (2003), 391–412.

[11] M. Parikh, M. Patel, D. Bhatt, *Animal detection using template matching algorithm*, Int. J. Res. Mod. Eng. Emerg. Technol, **1** (3) (2013), 26-32.

[12] Y. Fang, S. Du, R. Abdoola, K. Djouani, C. Richards, *Motion based animal detection in aerial videos*, Procedia Computer Science, **92** (2016), 13-17.

[13] T. Burghardt, J. Calic, *Real-time Face Detection and Tracking of Animals*, 2006 8th Seminar on Neural Network Applications in Electrical Engineering, Belgrade, Serbia, 2006, 27-32.

[14] M. B. Rangdal, D. B. Hanchate, *Animal detection using histogram oriented gradient*, International Journal on Recent and Innovation Trends in Computing and Communication, **(** 2), (2014), 178-183.

[15] Z. Xu, T. Wang, A. K. Skidmore, R. Lamprey, *A review of deep learning techniques for detecting animals in aerial and satellite images*, International Journal of Applied Earth Observation and Geoinformation, **128** (2024), 103732.

[16] N. Manohar, Y. S. Kumar, G. H. Kumar, *An approach for the development of animal tracking system*, International Journal of Computer Vision and Image Processing (IJCVIP), **8** (1) (2018), 15-31.

[17] N. I. Dopico, B. Bejar, S. V. Macua, P. Belanovic, P., S. Zazo, *Improved animal tracking algorithms using distributed Kalman-based filters*. In 17th European Wireless 2011-Sustainable Wireless Technologies, 2011, April , 1-8. VDE.

[18] Z. Kalafatic, S. Ribaric, V. Stanisavljevic, *A system for tracking laboratory animals based on optical flow and active contours*. In Proceedings 11th International Conference on Image Analysis and Processing, 2001, September, 334-339. IEEE.

[19] Y. Liu, W. Li, X. Liu, Z. Li, J. Yue, *Deep learning in multiple animal tracking: A survey*, Computers and Electronics in Agriculture, **224** (2024), 109161.

[20] R. Pereira, G. Carvalho, L. Garrote, U.J. Nunes, *Sort and deep-sort based multi-object tracking for mobile robotics: Evaluation with new data association metrics*, Applied Sciences, **12** (3) (2022), 1319.

[21] E. Fazzari, D. Romano, F. Falchi, C. Stefanini, *Animal behavior analysis methods using deep learning: A survey*, Expert Systems With Applications, 2025, 128330.

[22] SIMPATIA Animal Database:
http://iit.academiaromana-is.ro/simpatia/index_animals.html

[23] T. Barbu, C. Moroșanu, *Spatio-temporal Video Restoration Technique using a 3D Anisotropic Diffusion-based Scheme*, Bulletin of the Transilvania University of Brasov. Series III, **5** (67), No. 2, 2025.

[24] T. Barbu, C. Moroșanu, *Nonlinear Reaction-Diffusion Based Video Restoration Technique for Noise Mixtures*, IEEE 19th Conf. on Industrial Electronics and Appl. (ICIEA), Kristiansand, Norway, (2024), 1-4.

[25] M. Yaseen, *What is yolov9: An in-depth exploration of the internal features of the next-generation object detector*, arXiv:2409.07813, 2024.

[26] T. Barbu, *Digital Image Processing, Analysis and Computer Vision Using Nonlinear Partial Differential Equations*, **1211**, 2025, Springer Nature.

[27] G. Boole, *Calculus of finite differences*, BoD–Books on Demand, 2022.

[28] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, *Densely connected convolutional networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708 (2017).

[29] 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.

[30] X. Lu, et al. *X-CDNet: A real-time crosswalk detector based on YOLOX*, Journal of Visual Communication and Image Representation **102** (2024), 104206.

[31] Ö. Kaya, M. Çodur, E. Mustafaraj, *Automatic detection of pedestrian crosswalk with faster r-cnn and yolov7*, Buildings, **13** (4) (2023), 1070.

[32] N. Siddique, et al., *U-net and its variants for medical image segmentation: A review of theory and applications*, IEEE access 9, 2021, 82031-82057.

[33] L.C. Chen, et al. *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*, IEEE transactions on pattern analysis and machine intelligence **40.4** (2017), 834-848.

[34] V. Badrinarayanan, A. Kendall, R. Cipolla, *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*, IEEE transactions on pattern analysis and machine intelligence, **39** (12) (2017), 2481-2495.

[35] G.J. Brostow, J. Fauqueur, R. Cipolla, *Semantic object classes in video: A high-definition ground truth database*, Pattern recognition letters, **30** (2) (2009), 88-97.

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

[37] P. Soille, *Morphological operators*. In Computer Vision and Applications, Academic Press, 2000, 483-515.

[38] B. Liu, W. Zhao, Q. Sun, *Study of object detection based on Faster R-CNN*, In 2017 Chinese automation congress (CAC), oct. 2017, 6233-6236. IEEE

[39] I. Bello, et al., *Revisiting resnets: Improved training and scaling strategies*, Advances in Neural Information Processing Systems **34** (2021), 22614-22627.

[40] Pond5 – https://www.pond5.com.

[41] T. Barbu, A. Ciobanu, M. Luca, *Multimodal biometric authentication based on voice, face and iris*, 2015 E- Health and Bioengineering Conference (EHB), Iasi, Romania, 19-21 Nov. (2015), 1-4. IEEE