ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 3, May 2023, pp.229-238

Road conditions monitoring using semantic segmentation of smartphone motion sensor data

Emad Mahmood[1, *,] Nizar Zaghden[2], and Mahmoud Mejdoub[3]

[1] National School of Electronics and Telecommunication of Sfax, Universit of Sfax
[2] Higher School of business, University of Sfax, [3] Faculty of sciences Sfax, University of Sfax

[1.2.3] Research Lab: Smart systems for Engineering & E-health based on Technologies of Image & Telecommunications. (SETIT), ISBS. University of Sfax, Tunisia

ABSTRACT

Many studies and publications have been written about the use of moving object analysis to locate a specific item or replace a lost object in video sequences. Using semantic analysis, it could be challenging to pinpoint each meaning and follow the movement of moving objects. Some machine learning algorithms have turned to the right interpretation of photos or video recordings to communicate coherently. The technique converts visual patterns and features into visual language using dense and sparse optical flow algorithms. To semantically partition smartphone motion sensor data for any video categorization, using integrated bidirectional Long Short-Term Memory layers, this paper proposes a redesigned U-Net architecture. Experiments show that the proposed technique outperforms several existing semantic segmentation algorithms using z-axis accelerometer and z-axis gyroscope properties. The video sequence's numerous moving elements are synchronised with one another to follow the scenario. Also, the objective of this work is to assess the proposed model on roadways and other moving objects using five datasets (self-made dataset and the pothole600 dataset). After looking at the map or tracking an object, the results should be given together with the diagnosis of the moving object and its synchronization with video clips. The suggested model's goals were developed using a machine learning method that combines the validity of the results with the precision of finding the necessary moving parts. Python 3.7 platforms were used to complete the project since they are user-friendly and highly efficient platforms.

Keywords: Strips, Format Semantic Analysis, Dynamic Objects, U-Net Architecture, Machine Learning

Corresponding Author:

Emad Mahmood

National School of Electronics and Telecommunication of Sfax
University Sfax

E-mail: emadmah236@gmail.com

1. Introduction

There are a huge number of videos that are needed to describe and classify so that these videos can be retrieved easily. The work has contributed to the growth of techniques and tools for automated video analysis and interpretation. In fact, when a semantic understanding of contents is required to detect (and eventually retrieve) objects, actions, or events within a video stream, many video-based applications, such as video surveillance, road traffic control, and sports events detection, still heavily rely on human intervention. Manual analysis of video sequences is a time-consuming process that frequently yields misleading results due to the "video blindness" phenomenon, which affects human operators while watching video for an extended period. A video surveillance operator is thought to be able to overlook up to 95% of scene movements within just 22 minutes [1]. Semantic analysis is a technique for studying the idea or meaning of any image or video that is kept there, but it also exemplifies how a computer will understand and grasp each part of a particular movie in order to transform it into things that will assist the plot come to a satisfactory finish. [2]. Semantic analysis of video sequences presents several difficulties for automation and building a correct architecture for a particular film that must be analyzed. These issues will be the focus of the search for the best solution since automation requires ontology-based algorithms and methodologies that result in excellent application performance for search engines, video recommenders, and video summarizers. As a result, the semantic analysis focuses on the characteristics of video and their related characteristics, which include colors, edges, and arcs… etc. [3]. There are three levels of semantic analysis, (1) low-level semantic analysis techniques focus on the capacity to identify the visual areas that correspond to items of interest (detection), follow those objects over many frames, and keep their identities (tracking), (2) Simple or "atomic" actions or behaviors including loitering, falls, direction changes, group forms, and separations might be difficult for mid-level semantic analysis tools to identify. (3) High-level semantic analysis techniques focus on the identification of "complex" events or behaviors including hostility, fights, pickpocketing, thefts, general "strange occurrences," daily activities (particularly in the healthcare domain), vehicle theft, and so on. [4] [5]. The research of moving object speed and the evaluation of the proposed model using three levels of semantic analysis are the goals of this work. The article focuses on the following issues: the importance of moving elements in video sequences, the importance of each element's coordinates and qualities, and the detestation of static and repetitive elements like the background of a photograph. The essay will go more into deep learning and machine learning techniques. [6]. In this study, the motion sensor data from smartphones is divided up into several categories for analyzing the road surface using the U-Net architecture and BiLSTM networks. Flat road, pothole, speed bump, uneven surface, human movement, and machine vibration are a few of the categories in question. evaluating the effectiveness of the suggested method in comparison to other published methods in the literature. The creation of a novel scheme for evaluating movement objects at video sequences and diagnosing them as human movements, the study aims to improve the operation of analysis for movement objects. Three layers make up the work: the first deals with the characteristics of moving objects; the second, ontology-based classification; and the third, labeling or tag development. [7]. The paper includes five sections, first is related works and the second is the proposed method and the third section is talking about empirical evaluations and the results of paper and last section is a conclusion.

2. Related works

Semantic algorithms have been used in several video processing studies to evaluate moving objects to find them and use them in retrieval activities. The following names are given to these studies:

Jiadai Sune et. al. [8]: The authors supported the return of motion data and put forth the motion uncertainty-aware framework (MUNet) motion uncertainty-aware framework for semi-supervised video object segmentation (VOS). Secondly, based on a correlation cost volume, they provided an implicit technique for discovering the spatial correspondences between adjacent frames. They included the uncertainty in dense matching and established motion uncertainty-aware feature representation to tackle the difficult circumstances of occlusion and texture less areas while creating dense correspondences. To successfully combine the motion information with the semantic features, the authors secondly developed a motion-aware spatial attention module. Extensive tests on difficult benchmarks demonstrate that a modest quantity of data combined with strong motion information may significantly improve performance.

Sirine Ammar, Nizar Zaghden, et.al., 2020 [9]: A model of identifying and segmenting moving objects in video sequences was created using Generative Adversarial Networks and DeepSphere, an unsupervised anomaly detection framework (GANs). According to the data, the proposed method outperforms cutting-edge algorithms for segmenting and categorizing moving objects in surveillance video. Using the Deep Sphere architecture, the proposed Deep DC technique locates, isolates, and classifies moving objects in video sequences. To begin, moving objects are segmented using the Deep Sphere architecture. When the network outputs are thresholder, binary segmentation labels and morphological filters are generated. Deep Sphere-based object segmentation is outperformed by all BGS Library methods. The proposed approach initially collects deep features before classifying extracted images utilizing the GAN discriminator's ability to reliably classify data.

The approach employed in this study is the same, but the Gaussian method is utilized to identify visually moving objects and to comprehend their behavior utilizing a track system. The model to offer the diagnostic of moving item status is supported by the machine learning system.

Luca Greco et. al. 2017 [10] : The authors of this study provided a survey of several pertinent works on the use of Semantic Web technologies for video analysis. The purpose of the study was to describe the possibilities given by semantic web technologies for enhancing the functionality of sophisticated video analytic algorithms and solutions while also enhancing the efficiency of current algorithms and solutions. The assessed publications have been examined considering a proposed taxonomy of the SW technology adoption for video analysis. Additionally, an examination of the publications under consideration was done in terms of their timeline and use of SW languages. Due to improvements in reasoning techniques that enable efficient inference of complicated activities from atomic events, the study found a rising trend in the usage of Semantic Web technologies for event detection and activity recognition. Since semantic annotation is especially inexpensive for combining data obtained from many sensors in the big data and Internet of things sectors, the usage of ontologies is also growing rapidly in ubiquitous computing.

Budi Darma Setiawan et. al 2022 [11]: In the study presented, a novel U-Net architecture and BiLSTM networks were coupled to separate data from motion sensors in smartphones and categorize road surfaces according to their conditions. The suggested technique can effectively detect motion sensor data segments that correlate to road surface conditions, according to tests done with data gathered from four different cellphones. According to the results, adding BiLSTM to current network topologies like U-Net might improve segmentation performance. To improve segmentation and classification performance, future research will examine more reliable feature extraction (e.g., signal decomposition) and noise reduction (e.g., various filters) techniques.

Karishma Pawar et. al. 2022 [12] the authors proposed a deep learning based approach for detection and localization of road accidents. The benefit of the proposed approach is that they need to only train the model on normal traffic events and anomalies are detected as out of distribution samples due to one-class learning paradigm. They assessed the performance of the model over three bench marked datasets.

Olaf Ronneberger et. al 2015 [13]: The authors offered a network and training technique that makes heavy use of data augmentation to make better use of the readily accessible annotated examples. The design comprises of a symmetric expanding path that permits exact localization and a contracting path to capture context. On the ISBI challenge for segmenting neural structures in electron microscopic stacks, we demonstrate that a network of this type can be trained end-to-end from a very small number of photos and exceeds the previous best solution (a sliding-window convolutional network). The authors significantly outperformed the competition in the ISBI cell tracking challenge 2015 in several categories using the same network trained on pictures from transmitted light microscopy (phase contrast and DIC). The network is also fast. With a modern GPU, segmentation of a 512×512 picture takes less than a second.

YONG ZHANG et. al. 2019 [14] the authors developed a unique HAR approach based on U-Net, doing activity categorization and prediction at each sample point. In order to perform the pixel-level activity detection function, the motion sensor data gathered from the wearable sensors is translated into a picture with a single-pixel column and multi-channel. To achieve the dense prediction of motion sensor data, the authors created a comprehensive HAR framework based on U-Net, which includes data preprocessing, dense prediction, and post analysis. They suggested the post-correction algorithm for the dense prediction results based on the activity misalignment analysis in order to further enhance the dense prediction performance. The comprehensive experimental findings show that our U-Net technique outperforms both deep learning and conventional machine learning methods based on sliding window prediction. Also, based on the dense prediction on the four datasets, it performs better than the full convolutional network (FCN), SegNet, and Mask R-CNN. Moreover, it demonstrates improved resilience and great recognition performance for short-term activities and minority classes. To assess the effectiveness of the HAR algorithm, the authors provided a new dataset called Sanitation, which contains seven different categories of daily work activity data of sanitation personnel. The study suggested a strategy for utilizing a machine learning system to diagnose moving objects to determine their state and build a map of them to understand how they behave.

3. Method

Data from the collected datasets must be recorded by a smartphone application. The driver's pocket or the car are both home to smartphones that have the app loaded. The suggested technique requires a manually labeled training dataset because it uses supervised learning. The detected object is photographed for labeling purposes using cameras mounted on the cars and directed towards the object in front of the vehicle for example. Based on the videos gathered, the start and finish of a section in a signal sequence may be determined. To use the same time as a reference, the camera and smartphone must be synced. [15]. The proposed approach may quickly discover and extract potential dynamic elements from a static map. Even though the human operators move about often, they are entirely removed from the static map. All potential dynamic objects are bounding boxed and added to a final semantic map to show their state in real time, with the moving human colored red and the picture colored green. Our approach can distinguish and discover multiple targets in a complex dynamic environment. . [16] [17] The rotation angle is calculated using equations (1) and (2), and Equations are used to produce new x, y, and z standard locations. (3) to (5).

$$\alpha = \begin{cases} \tan^{-1}\dfrac{accY}{accZ} + 180°, \text{ if } accZ < 0 \\ \tan^{-1}\dfrac{accY}{accZ}, \text{ otherwise} \end{cases} \qquad \ldots\ldots\ldots\ldots\ldots\ (1)$$

$$\beta = \tan^{-1}\left(\frac{accX}{\sqrt{accY^2 + accZ^2}}\right) \qquad \ldots\ldots\ldots\ldots\ldots\ (2)$$

$$x' = x\cos(\beta) + y\sin(\beta)\sin(\alpha) + z\cos(\alpha)\sin(\beta); \qquad \ldots\ldots\ldots\ldots\ldots\ (3)$$

$$y' = \mathrm{y}\cos(\alpha) - \mathrm{z}\sin(\alpha); \qquad \ldots\ldots\ldots\ldots\ldots\ (4)$$

$$z' = -x\sin(\beta) + y\cos(\beta)\sin(\alpha) + z\cos(\beta)\cos(\alpha); \qquad \ldots\ldots\ldots\ldots\ldots\ (5)$$

Sensor data values for these variables x, y, and z are those from before transformation, while those for variables x′, Y′, and Z′ are those from after transformation. Along the x- and y-axes, the rotation angles are and, respectively. Just the z-axis data (accelerometer-z and gyroscope-z) are preserved during the segmentation process because when the accelerometer and gyroscope signals are normalized, all values along the x- and y-axes will be extremely close to zero. A sliding window is used to turn the accelerometer and gyroscope signals into t-length sequences, hence diversifying the training set and the volume of data. To prevent segment overlap, s is set to equal t for the validation and testing datasets. One window needs to be able to show the length of the longest part while eliminating background segments given the size of the t. (in this case, a flat road). Each window contains a portion of the 1D-BiLSTM Skip-U-Net. The architecture is shown in Fig. 1. Blue arrows represent the several 1-dimensional convolutional layers that make up the encoder section of the network. The solid blue arrows demonstrate that the output is the same size as the input since the layers have a stride of 1. A dash during a blue arrow denotes a two-step stride and twice as many kernels. The convolution method employed in the encoder element is reversed in the decoder component using one-dimensional deconvolution, also known as convolution transposition. Layers with solid green arrows employ strides of 1 and 2, but layers with dashed green arrows always use half as many kernels as the one before them. The kernel is composed of 15 layers. Orange arrows indicate a concatenation of the BiLSTM output and associated decoder layers after red arrows display the skip architecture employing BiLSTM layers. The Rectified Linear Unit (ReLU) is used by all convolutional and deconvolutional layers, whereas BiLSTM levels employ the hyperbolic tangent (tanh) for activation functions. Batch normalization is used prior to each activation, disregarding the input and output layers, to avoid overfitting. Also, every epoch, training batches are rotated. [18].

A word or action is recorded using machine learning so that the computer can recover it the next time the user wishes to type it. Supervised learning and unsupervised learning are the two types of machine learning. A training set with accurate targets is used in supervised or semi-supervised learning, and the technique is modified to perform more accurately depending on the training set. Unsupervised learning doesn't have any proper objectives; instead, it looks for similarities between inputs to group items with comparable characteristics. [20]. By learning typical actions, an unsupervised machine learning model may identify abnormal human behavior. Additionally, by simulating the dynamics and interactions of its characteristics, it is possible to learn the state of moving objects. This model has the benefit of being able to explain how it arrived at its decision to visualize related aspects. The research uses representative tracking techniques to offer baseline findings for a 2D video stream. for tracking single and multiple views. The baselines for the generative and discriminative methods in this study are a credible Bayesian tracker and twin Gaussian processes, respectively. Considering the findings, the proposed model would employ the Gaussian Mixture Model (GMM) estimate technique and provide a range of baseline strategies [21]. The model uses a variety of algorithms, including Luxand, BioID, and others, to determine the right moving item state. Not the whole of a person's body is employed by these algorithms; only their faces and status are. For representing state of the moving objects, the model will employ a map-ping approach to present the characteristics during the training database as keys of descriptions [22].

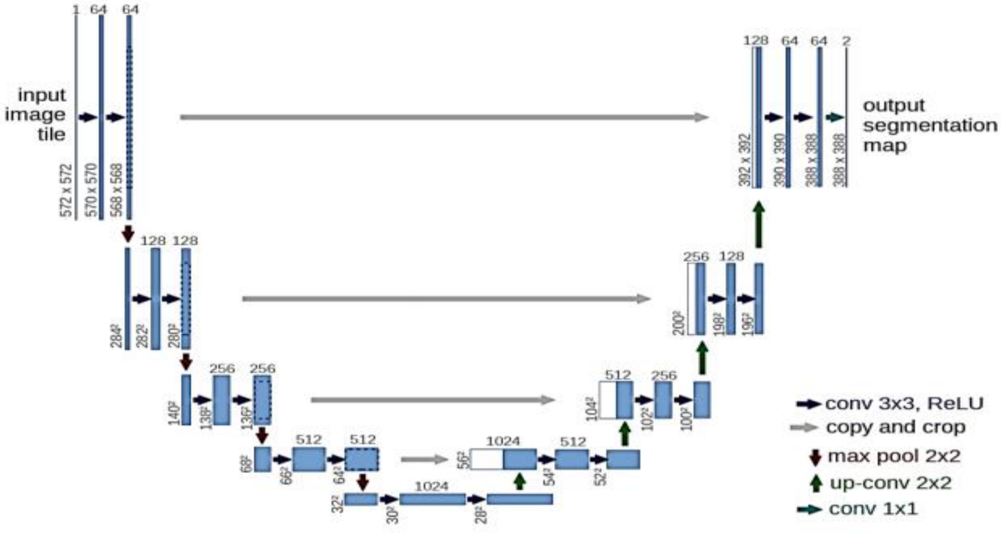

Figure 1. The U-Net architecture [19]

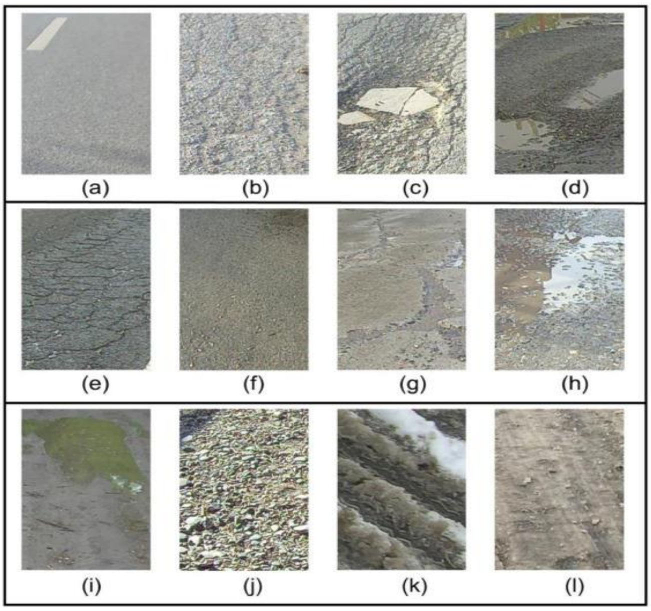

Figure (3) . Sample images of part of the classes. (a). dry-asphalt-smooth (b). dry-asphalt-slight (c). dry-asphalt-severe (d). water-asphalt-severe (e) wet-asphalt-slight (f). wet-concrete-smooth (g). wet-concrete-severe (h). water-concrete-slight (i). water-mud (j). dry-gravel (k). melted snow (l). ice. [23]

4. Empirical evaluation

The success of work suggested approach will be shown in this part through the presentation of experimental findings. The specifics of experimental setup will be covered first. The authors next go into further detail about how they got information about possible moving objects in road surfaces setting. Third, they assessed how well their chosen instance segmentation model performs segmentation. Then they went into detail about how they carried out the dense mapping and dynamic tracking. Finally, they assessed how well their suggested technique performs in terms of localization drifts in dynamic situations.

4.1. Empirical Installation

The authors installed Pycharm IDE 3.2 and Python 2.7 with almost all related modules to implement the proposed algorithms in PC computer. The project implemented by PC computer with an Core i5 CPU and display card (Nvidia GeForce RTX 2080 Ti GPU), all the trials are carried out.

4.2. Dataset

In many situations, like autonomous driving and smart warehouse operations, people are frequently viewed as dynamic objects. The HumanEva dataset has more than 600 human videos, so the authors chose them. The suggested approach is tested in the experiment in the road surfaces setting, as seen in Fig. 3. In addition to taking into account humans as dynamic objects, an advanced factory necessitates collaboration between humans and other dynamic objects making Automated viable dynamic objects. In order to train the instance segmentation network, 1,000 pictures in total are gathered. Some of these pictures are seen in Figure 3.

Table 1. Tested video's information

Video	Size	Number of pictures	Number of key frame
1.	41 M	162	4
2.	716 M	253	7
3.	126 M	104	6
4.	716 M	263	9
5.	716 M	218	9

4.3. Semantic performance

The source code (Gaussian Distributed) has been successfully utilized to detect several moving objects in several videos (of the MP4 variety). Each item consists of several qualities that may be used to classify it into various objects. Figure 3 shows the working environment and the estimated parameters that were supplied. To adjust the comparison, the figure is compared with various methods that make use of the same datasets. The strategy used in this study was built based on how many Gaussians were visible in the video to diagnose using a comparison methodology, allowing the model to be dependent on a database that nearly completely covered the state of moving objects. The error ratio is displayed by the model's estimate and analysis functions. The ratio of error either goes away toward the actual item or deviates from it. The model develops these classifications for a high degree of diagnosis by examining the frame rate, the synchronization of moving objects, and their relativity.

4.4. Dynamic monitoring and accurate labeling

In order to correct or converge to the ideal bounds and to obtain a high level of segmentation, the important discoveries that rely on GMM iterations are presented in Table 2. The videos use a ratio of sounds to explore how each algorithm works. As GMM looks for the best Gaussians for moving objects, it produced a broadly accepted result. Given that the suggested model and the Conventional Neural Network (CNN) and LBP algorithms use distinct methods, their results were compared. Finding the top results is the goal of comparison. The outcomes are extremely comparable since video1 has very little noise. Even if most videos have high noise levels, the Gaussian algorithm research to the best detection, which is why the findings are inconsistent. The suggested approach is put into practice on the example depicted in Fig. 3 to assess the performance of our multimodal semantics in dynamic situations. Different images' kinds and human operators are expected to cooperate in a smart manufacturing facility. The key technology towards industry 4.0 is the capacity of each image to locate itself under moving human operators and other images in the sequence videos. The remaining items, such as tables or running equipment, can sometimes be a static environment in environment settings. Therefore, to minimize the computational cost, the autor only take into account humans and sequence video as dynamic objects [24] [25]. In the experiment, while the human operators are strolling regularly in the road, video is manually controlled to cross a road and the environment map at the same time. The localization result is displayed in Fig. 3, where we contrast the outcomes of the proposed method, the ground truth, and alternative strategies. The suggested multi-modal semantics is more reliable and stable than previous approaches when the dynamic object occurs. The suggested approach effectively distinguishes the prospective dynamic items from the static map. Although often walking, the human operators are not at all visible on the static map. The moving human is colored in red, while the image is colored in green. All potential dynamic items are encircled by bounding boxes and added to a final semantic map to see the status of each thing in real time. Many targets may be found and located using our technology even in a complicated dynamic environment.

Table 2. Comparison of events classification performance

Model	**Segmentation**	**Means %**	**Inference Time (ms)**
CCP [26]	0.29	38	55
LPB [27]	0.36	32	58
KNN [28]	0.38	39	56
NeRF [6]	0.34	37	61
GMM & U-Net (proposed Model)	0.28	41	62

The article employed algorithms that categorize the many types of roads, including flat roads, potholes, machine vibration, uneven surfaces, and speed bumps, which are each represented by a different color in the segmentation results. The algorithms suggested somewhat different start and finish positions in these situations. The best outcomes, which closely match the ground truth segments, were from designs modified for BiLSTM networks. In Table 3, the outcomes of the proposed model for one scenario (machine vibration) are displayed and compared to those of other models.

Table 3 the comparison of performance results for the proposed model in one case (machine vibration) with other models

Model	**Machine vibration**		
	Precision	**Recall**	**F1-score**
CCP [26]	%72.13	%50.78	%44.56
LPB [27]	%73.45	%52.28	%46.89
KNN [28]	%71.54	%55.12	%57.89
NeRF [6]	%74.15	%63.2	%61.33
GMM & U-Net (proposed Model)	%89.22	%88.52	%88.26

5. Conclusion

To classify various road surfaces according to their conditions, smart phone motion sensor data was segmented using BiLSTM networks and a distinctive U-Net architecture. Three separate cellphones' worth of data were used in experiments to show that the recommended approach works well for identifying motion sensor data segments linked to the condition of the road. The findings suggest that by including BiLSTM into pre-existing network topologies like U-Net, segmentation performance may be enhanced.The article is used methods for classifying the many sorts of roads, including flat roads, potholes, machine vibration, uneven surfaces, and speed bumps. One case from the study—machine vibration—showed how the algorithms suggested somewhat different start and stop positions. The designs changed for BiLSTM networks produced the best results, which nearly matched the ground truth segments. When compared to other models, Table 3 shows the results of the proposed model for one situation (machine vibration). In terms of F1-score ratings, recall, and accuracy, the suggested model outperforms existing models. The predictions of the Gaussian Mixture Model for autonomous driving are assessed based on feature extraction. This paper investigates the temporal consistency of semantic moving objects. The study, which was published, identified the temporal consistency issues using a single sequence of unlabeled photographs. To better understand the semantics and tagging process, we first built a GMM. The study employed an unsupervised method to assess the precision of semantic prediction based on video sequences. With the use of Python image analysis, we were able to identify and classify certain moving objects. The module uses the video-datasets.org dataset to recognize moving objects. In order to provide accurate detection and evaluations of moving objects, we tried to enhance the identification process. Future revisions should improve the study's approach. The performance of segmentation and classification will be improved by looking into more effective feature extraction (such as signal decomposition) and noise reduction (such as various filters) techniques.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] M. Hamouda and M. S. Bouhlel, "Modified Convolutional Neural Networks Architecture for Hyperspectral Image Classification (Extra-Convolutional Neural Networks)," IET Image Processing, vol. 15, no. 2, pp. 305-313, 2021.

[2] H. Li, "Automatic Detection and Analysis of Player Action in Moving Background Sports Video Sequences," in IEEE International Conference on Multimedia and Expo (ICME), pp. 351-364, 2010.

[3] M. Jasim, N. Zaghden and M. S. Bouhlel, "Identification of Collision Alert in Vehicle Ad hoc based on Machine learning," in IEEE International Conference on Computing (ICOCO), 2021.

[4] S.V.-U. Ha ORCID, N.M. Chung, H.N. Phan ORCID and C.T. Nguyen, "TensorMoG: A Tensor-Driven Gaussian Mixture Model with Dynamic Scene Adaptation for Background Modelling," Sensors, vol. 20, no. 22, pp. 1-29, 2020.

[5] S. Ammar, T. Bouwmans, M. Neji, and N. Zaghden, "Moving Objects Segmentation Based on DeepSphere in Video Surveillance," Academia, pp. 307-319, 2020.

[6] B.D. Setiawan, U. Serdult, and V. Kryssanov, "A Machine Learning Framework for Balancing Training Sets of Sensor Sequential Data," Sensors, vol. 5, pp. 20-32, 2021.

[7] N. Khalid, Y.Y. Ghadi, M. Gochoo, A. Jalal, and K. Kim, "Semantic Recognition of Human-Object Interactions via Gaussian-Based Elliptical Modeling and Pixel-Level Labeling," IEEE, pp. 111249-111266, 2021.

[8] J. Sun, Y. Mao, Y. Dai, Y. Zhong and J.Wang,"Motion uncertainty-aware semi-supervised video object segmentation," Pattern Recognition vol.138,no109399 ,2023.

[9] S.Ammar,T.Bouwmans,N.Zaghden,andM.Neji,"Deep detector classifier (DeepDC) for moving objects segmentation and classification in video surveillance," surveillance ,pp1490-1501 ,2020 . [10] Luca Greco, Pierluigi Ritrovato, Mario Vento, "On the use of semantic technologies for video analysis," IOS Press and the authors, vol. 2, no. 132, pp. 1-21, 2017.

[11] B. D. Setiawana, M. Kovacs, U. Serdult, V. Kryssanov, "Semantic Segmentation on Smartphone Motion Sensor Data for Road Surface Monitoring," ScienceDirect, pp. 346-353, 2022.

[12] V. A. K. Pawar, "Deep learning based detection and localization of road accidents from traffic surveillance videos," ScienceDirect, vol. 8, pp. 379-387, 2022.

[13] O. Ronneberger, P. Fischer & T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Springers , vol. 3, no. 9351, pp. 234-241, 2015.

[14] Y. Zhang et al., "Human Activity Recognition Based on Motion Sensor Using U-Net," IEEE, vol. 7, pp. 75213-75226, 2019.

[15] L. Sigal et al., "Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," International Journal of Computer Vision, pp. 4-27, 2010.

[16] X. Li and D. W. Goldberg, "Toward a mobile crowdsensing system for road surface assessment," Computers, Environment and Urban Systems, vol. 69, pp. 51-62, 2018.

[17] A. M. Abirami and V. Gayathrii, "A survey on sentiment analysis methods and," in IEEE, pp. 72-76, 2016.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International Machine Learning Society (IMLS), pp. 448-456, 2015.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention-MICCAI, vol. 9351, pp. 234-241, 2015.

[20] L. Greco, "On the use of semantic technologies for video," Semantic Web Journal [Online]. Available: http://www.semantic-web-journal.net/system/files/swj1789.pdf., Jan. 2021.

[21] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep Video Anomaly Detection: Opportunities and Challenges," IEEE, pp. 1-5, 2021.

[22] N. Zaghden, B. Khelifi, A.M. Alimi, and R. Mullot., "Text Recognition in both ancient and cartographic documents," arXiv preprint arXiv:1308.6309, pp. 98-101, 2013.

[23] T. Zhao and Y. Wei, "A road surface image dataset with detailed annotations for driving assistance applications," Elsevier, vol. 12, pp. 23-50, 2022.

[24] M. Otani, "Video Summarization using Deep Semantic Features," in Asian Conference on Computer Vision (ACCV), Oulu, Finland, 2016.

[25] S. Ammar, T. Bouwmans, N. Zaghden, and M. Neji., "Moving objects segmentation based on deepsphere in video surveillance," in International Symposium on Visual Computing (ISVC), Cham Switzerland , 2019.

[26] S.Saad., "Semantic Analysis of Human Movements in Videos," ACM Transactions on Multimedia Computing Communications and Applications (TOMM), vol .8 , no .3 , pp .141-148 ,2012 .

[27] P.Gonçalves and M.Araújo., "Comparing and Combining Sentiment Analysis Methods," ACM Transactions on Internet Technology (TOIT), vol .14 , no .4 , pp .1-11 ,2014 .

[28] R.Morais,V.Le,T.Tran,B.Saha,M.Mansour,and S.Venkatesh., "Learning regularity in skeleton trajectories for anomaly detection in videos," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.