The advancements and future prospects of pedestrian action recognition based on machine learning algorithms

Yi Wen

Data Science and Big Data Technology, North China University of Technology, 101300, China

21152090106@mail.ncut.edu.cn

Abstract. This paper highlights the shift from classical machine learning to deep learning models and describes the fundamental methodology and developments in pedestrian activity recognition. The four main steps of the workflow are the gathering of datasets, pre-processing, designing and training the model, and evaluating the outcome. To extract pedestrian feature vectors, data must first be collected, cleaned, and processed from public or proprietary datasets. These vectors are used to train deep learning or machine learning models, which are subsequently assessed and fine-tuned for use in practical applications such as behaviour analysis and surveillance. For action recognition, conventional machine learning techniques like Random Forests (RF) and Support Vector Machines (SVM) have been used. SVMs, despite their potential for computing complexity, identify the best hyperplanes for classification. The categorization rates for a variety of human behaviours have been enhanced by a combination strategy utilizing SVMs and decision trees. As shown in a study that uses smartphone accelerometers to accurately identify everyday activities, RFs can manage enormous datasets. Deep learning models that automatically learn complicated feature representations, such as VA-fusion, AGC-LSTM, and LC-POSEGAIT, provide improved performance. These models capture minute differences in pedestrian behaviour using CNNs, RNNs, and LSTM architectures. Interpretability, generalization to new datasets, and computing demands are some of the difficulties they encounter. Future developments could include using transfer learning to improve performance in many circumstances, combining deep learning and expert systems for improved interpretability, and utilizing distributed computing for processing in an effective manner.

Keywords: pedestrian action recognition, machine learning, deep learning

1. Introduction

The degree of urban modernization has risen rapidly as a result of China's economy modernizing at a quick pace and its transportation networks becoming more sophisticated. But this advancement has also been accompanied by a rise in pedestrian accidents due to things like people strolling in the wrong lanes. The World Health Organization has released statistics showing that 50 million people are wounded, and 1.27 million people die in traffic accidents annually. Numerous investigations on the amount of traffic fatalities in China have also been carried out by national experts. Using a grey prediction model, Zhang et al. from Chongqing Jiao Tong University discovered in 2016 that there were 55958, 53881, 51880, and 49954 road fatalities in China between 2014 and 2017 [1]. In order to predict a pedestrian's intention to cross the lane, one of the most crucial areas is pedestrian recognition and pedestrian status detection using photos and video surveillance. Conventional video surveillance systems only offer rudimentary replay and archival capabilities. Their algorithms are also unreliable and susceptible to outside influences, which makes it challenging to reliably identify unusual activity and prevents them from issuing timely alerts. Therefore, it is essential to use deep cognition in the recognition of pedestrian behaviours because of its strong extraction and prediction skills.

The subject of machine learning has advanced significantly in the last several years. Numerous domains, including finance, chemistry, and healthcare, have utilized diverse methods, such as gradient boosting machines (GBM), random forests (RF), and support vehicles (SVM). The accuracy of model projections is a crucial factor in the transportation sector. Shu et al., for instance, estimate several people's whereabouts in real time using AlphaPose representations [2].Two types of algorithms are currently used to summarize multiple human pose estimation: a top-down algorithm developed by Pishchulin et al. [3] that identifies key points and creates coordinate maps of these points in the original image; and a bottom-up algorithm developed by Chen et al. [4] that identifies all of the human body's skeletal points in an image and uses graph theory to relate these skeletal points to the human skeleton in order to estimate and recognize pedestrian poses. Furthermore, there are models like the LC-POSEGAIT pedestrian recognition model, which combines CNN with Long-Short-Term-Memory (LSTM) Qi et al. [5] and Periodic neural network (RNN) adaptive for vision with CNN and LSTM (VA-RNN) (VA-CNN). The fusion model of VA-fusion by Zhang et al. [6] illustrates how the rankings of the two networks can be combined to produce a foggy prediction, where the adaptive subnetwork visually identifies relevant observation points and the underlying Convnet network uses the skeleton in a novel way to categorize actions. Meanwhile, the new attention model enhances the convolutional LSTM network (AGC-LSTM) model by Si et al. [7], which relies on local joint characteristics to predict human action categories. Despite their individual flaws, these models have greatly improved pedestrian identification systems' performance. A thorough analysis of this discipline is required because of the relevance of this field and the recent ongoing development of increasingly sophisticated algorithms and measures.

The advancements in machine learning technology in this field are the main topic of this essay. Three sections make up the remainder of the paper. The integration and enhancement of the synthesis model VA-fusion are covered in the second part, Method, which also primarily explains how current approaches are used in this sector. The final section, titled "Discussion," examines and evaluates the drawbacks of the approaches used thus far, as well as future directions and individual viewpoints. The paper's analysis and findings about the application of integrated models are presented in the concluding "Conclusion" section.

2. Methods

2.1. Basic workflow of prediction

This article finds that most pedestrian action recognition studies generally follow the workflow outlined below shown in Figure 1. First, data in the form of images or videos enters the automatic generation stage, where data collection and pre-processing occur. Key features of pedestrians in the data stream are extracted and aggregated into feature vectors, which are then used to train the specified neural network model. In the second stage, pedestrian actions are analysed and recognized, and finally, results are output and optimized to complete the prediction process.

1. Dataset Collection: Obtain data for training, either by downloading publicly available datasets from websites or by creating custom datasets for specific purposes.

2. Dataset Pre-processing: Data pre-processing is a crucial initial step in machine learning, data analysis, and mining processes. This involves cleaning, transforming, and organizing the dataset to extract pedestrian feature vectors, thereby facilitating better subsequent analysis, training, and modeling.

3. Model Design and Training: The aggregated feature vectors are used to train various machine learning or deep learning models developed through research. Continuous iteration and updates are performed to enhance the model's accuracy and real-time performance.

4. Result Evaluation: Conduct a detailed analysis of the recognition results to identify the causes of misclassification and improve the model. Apply the optimized model to real-world scenarios, such as intelligent surveillance, behaviour analysis, and motion capture.

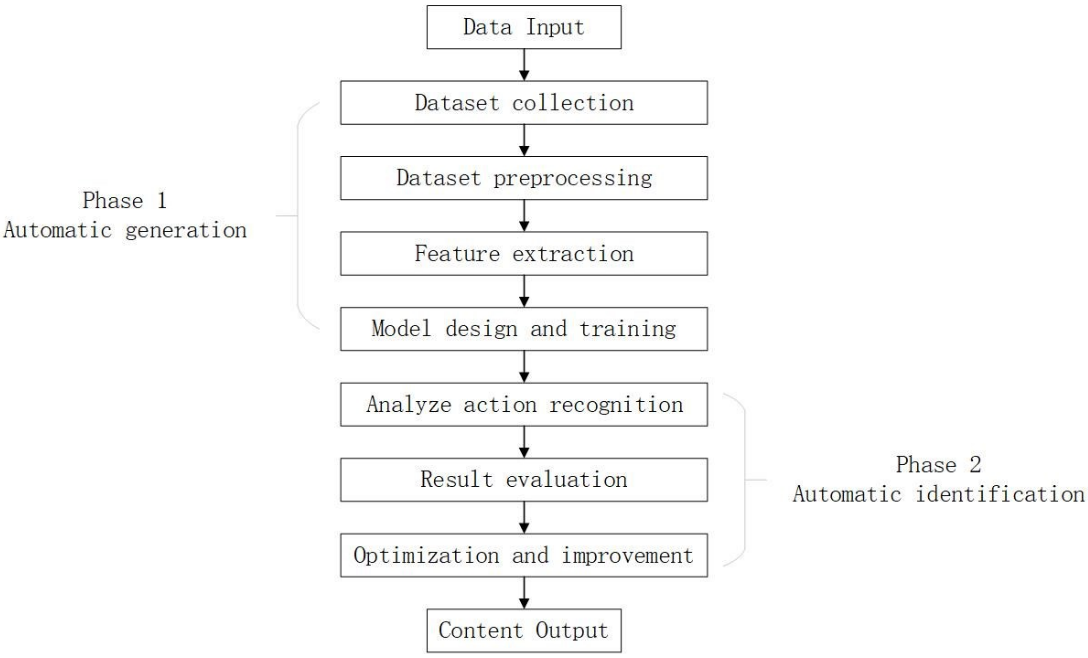

Figure 1. The basic workflow of pedestrian motion predictio (Photo/Picture credit: Original).

2.2. Traditional machine learning algorithms-based prediction

2.2.1. SVM

SVM is based on locating the most effective linear hyperplane in the feature space that distinguishes at least two basic classes. The linear training method of traditional SVM results in a complicated, time-consuming, and costly learning process for a carrier-supported machine classifier. Hong et al. solved multi-layer classification problems by combining machines with non-linear support vectors and decision trees to address this problem. The hierarchical structure of decision schemes constructed in this manner requires only one sub-m at each level. The greater the proximity of the sub SVM to the root node, the more significant it becomes, resulting in a higher number of elements in the training set. A dataset based on human movements was used to test six different types of human behaviours, including walking, jogging, waving, running, boxing, and clapping. The application of fuzzy c-means clustering technique using six cluster centers identified six behaviours for each sequence, which was then used to construct a multi-compartment SVM separator. In the end, this approach achieved a higher recognition rate for straightforward daily actions, a higher recognition rate for folding tables, and a lower recognition rate for walking, running, and jumping with one leg.

2.2.2. RF

Human behaviour can be identified through a compact representation using a Random Forest classifier. Random forests possess the capacity to handle thousands of input variables and large data sets in an efficient manner. Haq et al. constructed a smartphone accelerometer model that can accurately identify four daily living activities (lying, standing, sitting, and walking) and their contexts with a random forest classifier [8]. The study's results indicate that identifying activity contexts related to walking and sitting is more challenging than identifying those related to lying and standing. Each activity's context recognition performance is best achieved by using the RF classifier. The accuracy of context recognition using RF classifiers is 97.76% for lying, 86.90% for sitting, 95.28% for standing, and 71.4% for walking.

2.3. Deep learning algorithms-based prediction

2.3.1. LC-POSEGAIT

Two LSTM layers and two Flatten layers make up the LSTM branch LC-POSEGAIT shown in Figure 2. After going through two LSTM layers and Flatten layers, the displacement constraint matrix is transformed into a vector of one-dimensional displacement constraints. Four layers, four merging layers, and one Flatten layer make up the CNN branch. Four cycles of displacement and merging through the CNN are used to transform the action feature matrix into a vector of one-dimensional action features, then the Flatten layer is used to convert it. To obtain the hiking feature vector, the two univariate vectors are merged and then passed through FC-1 and FC-2's fully connected layers. The 3D gait feature of layer FC-1 is employed to recognize pedestrian gait [5].

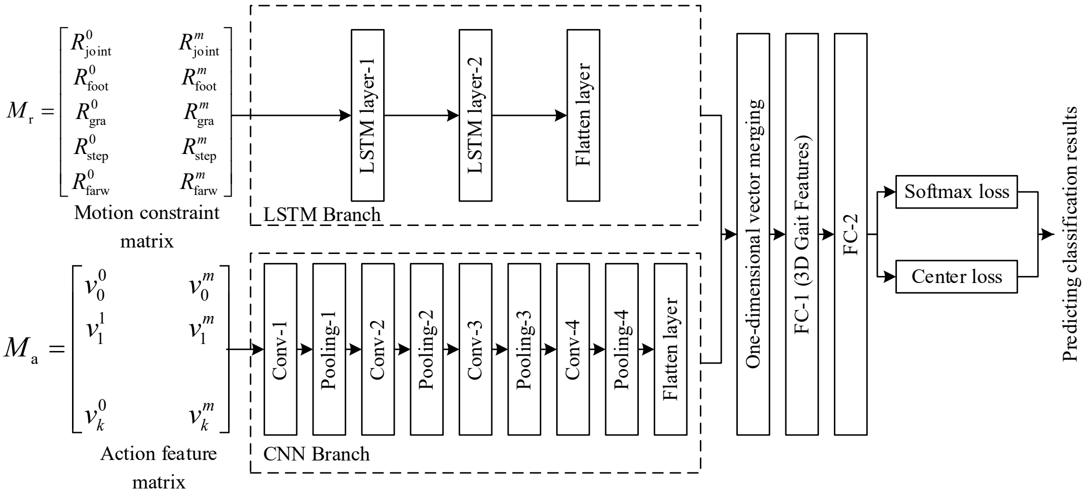

Figure 2. LC-POSEGAIT network model [5].

2.3.2. VA-fusion

LSTM is used to construct the architecture of the vision adaptive neural network that comprises a Vision-Adaptive RNN (VA-RNN) and a View-Adaptive CNN (VA-CNN). An adaptive vision subnetwork and a primary LSTM network are the components of VA-RNN, as shown in Figure 3. An observation point is present in each slot. The primary LSTM network is responsible for determining the action class by applying the skeleton representation of the new observation point. VA-CNN is composed of a subnet that adapts to vision and a primary convoluted network (Convnet). In order to determine the proper observation point, the vision-adaptive subnet works in sequence. The primary convnet is used to determine the action class using the skeletal representation of the new observation point. A combined prediction can be achieved by combining the classification scores of both networks [6].

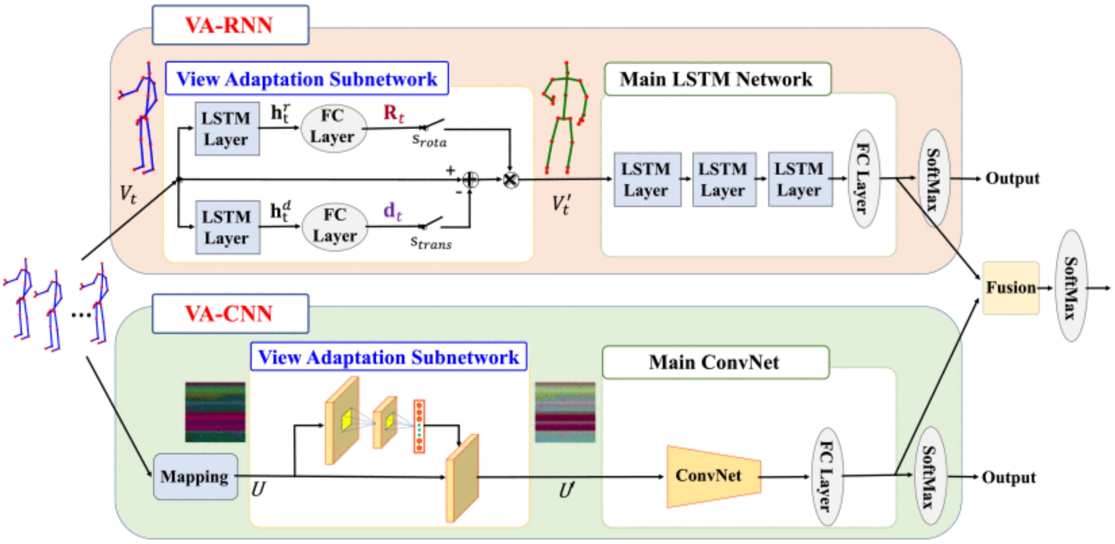

Figure 3. VA fusion network model [6].

2.3.3. *AGC-LSTM*

Attention (AGC-LSTM) shown in Figure 4 is a factor in the enhancement of the current LSTM network architecture graph. Positional features are computed and associated with feature differences by Feature augmentation (FA). Scale differences between feature differences and positional features are eliminated through the use of the LSTM network. Different spatiotemporal characteristics can be formed by the three layers of AGC-LSTM. Time-averaging aggregation is used to achieve time-averaging. The last AGC-LSTM layer is used by this architecture to predict the human action category based on the global characteristics of all joints and the local characteristics of focused joints [7].

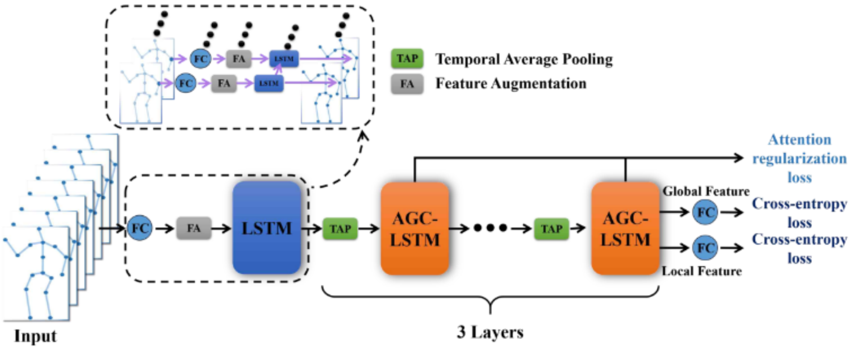

Figure 4. AGC-LSTM network model [7].

3. Discussion

Currently, pedestrian action recognition models have gradually transitioned from traditional machine learning to deep learning. This shift is primarily due to:

1) Deep learning models like CNN and RNN can automatically learn complex feature representations from raw data. These models do not require manual design of feature extractors as in machine learning; instead, they progressively extract and aggregate features through layered stacking to better capture subtle variations and patterns in pedestrian actions. 2) Deep learning models support end-to-end learning, meaning they can directly learn the final action recognition results from raw inputs (such as image sequences or sensor data). This approach simplifies the complex process in traditional machine learning that involves manually designing and adjusting multiple stages (such as feature extraction, feature selection, and classifier design). 3) Deep learning models typically require large amounts of data for training, and over time, large-scale pedestrian action datasets available for training have become richer

and more extensive. These datasets help deep learning models generalize better and recognize pedestrian actions in different environments. 4) In many pedestrian action recognition tasks, deep learning models have demonstrated higher accuracy and robustness compared to traditional machine learning methods. Especially when dealing with complex backgrounds, occlusions, and significant variations in pedestrian actions, deep learning models often provide more accurate recognition results.

While deep learning has shown outstanding performance in pedestrian action recognition, it also faces some limitations and challenges: 1) Deep learning models are often complex black-box models, where the internal structure and parameters determine intricate decision-making processes. Therefore, understanding how these models learn, recognize pedestrian actions from input data, infer future actions, and explain the specific reasons and basis for their decisions is challenging. 2) Deep learning models typically perform well with large-scale, high-quality training data, but their ability to generalize may be limited when faced with new datasets with different distributions. Variations in pedestrian actions due to different environments, backgrounds, or cultural habits may reduce the model's accuracy in specific scenarios. 3) Deep learning models often require significant computational resources for training and may involve long processing times during inference. In real-time pedestrian action recognition applications, especially those requiring low latency and high frame rates, deep learning models may struggle to meet real-time requirements unless pretrained models are used to reduce training time significantly or optimizations are applied to achieve more efficient inference computations.

The future outlook for deep learning models in pedestrian action recognition includes several aspects: 1) Combining deep learning with expert systems can enhance the interpretability and accuracy of models. Expert systems use domain knowledge and rules to increase transparency in the decision-making process of deep learning models. Comparing the outputs of deep learning models with rules from expert systems can enhance trust in model decisions. For example, applying SHAP or other interpretable algorithms [9] to pedestrian action recognition can help understand how models predict specific action recognition results from input data, thereby improving model interpretability and transparency. 2) Transfer learning [10] can leverage pretrained deep learning models on large-scale datasets to transfer knowledge to new tasks or datasets. For pedestrian action recognition, transfer learning can reduce the need for large amounts of annotated data and improve model performance in new environments and scenarios. Thereinto, domain adaptation techniques can help models adjust between source (training data) and target domains (test data), maintaining high recognition accuracy in different data distributions. For instance, unsupervised domain adaptation techniques can enhance model adaptability in scenarios lacking annotated target domain data. 3) Leveraging Apache Spark's powerful distributed computing capabilities can accelerate the training and inference processes of pedestrian action recognition models. Using Spark for data pre-processing, feature extraction, and distributed training can significantly enhance model training efficiency and processing capacity, especially when handling large volumes of video data.

4. Conclusion

This paper reviews the application of parallel computing with integrated deep learning models to solve pedestrian action recognition problems. For instance, AlphaPose is used for real-time multi-person pose estimation, while deep learning models like LC-POSEGAIT, VA-fusion, and AGC-LSTM significantly enhance recognition accuracy and performance through various network architectures and fusion methods. The proposed methods have been extensively evaluated through experiments. The discussion indicated that each deep learning model has its own shortcomings, finding it challenging to balance improved accuracy and robustness with model generalization ability and real-time performance. Future prospects include combining deep learning with expert systems, applying transfer learning and domain adaptation techniques to improve model performance in different environments, and utilizing distributed computing to accelerate model training and inference processes.

References

[1] Zhang Y et al 2016 National road traffic accident prediction [in Chinese] Beijing Automotive vol 2016 (03) pp 32-35

[2] Fang H -S et al. 2023 AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time IEEE Transactions on Pattern Analysis and Machine Intelligence vol 45 no 6 pp 7157-7173 1 June 2023

[3] Pishchulin L Jain A Andriluka M Thormählen T & Schiele B 2012 Articulated people detection and pose estimation: Reshaping the future Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 3178-3185

[4] Chen X & Yuille A 2015 Parsing occluded people by flexible compositions Proceedings of he IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston: IEEE) pp 3945-3954

[5] Qi Y et al. 2021 A gait recognition method combining LSTM and CNN [in Chinese] Journal of Xidian University vol 48 (05) pp 78-85

[6] Zhang P Lan C Xing J Zeng W Xue J & Zheng N 2019 View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence vol 41 no 8 pp 1963-1978 1 Aug 2019

[7] Si C Chen W Wang W Wang L & Tan T 2019 An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach: IEEE) pp 1227-1236

[8] Ehatisham-ul Haq M Azam M A Asim Y Amin Y Naeem U & Khalid A 2020 Using smartphone accelerometer for human physical activity and context recognition in-the-wild Procedia Computer Science vol 177 pp 24-31

[9] Qiu Y Chen H Dong X Lin Z Liao IY Tistarelli M Jin Z 2024 Ifvit: Interpretable fixed-length representation for fingerprint matching via vision transformer. arXiv preprint arXiv:2404.08237

[10] Yan P Abdulkadir A Luley PP Rosenthal M Schatte GA Grewe BF Stadelmann T 2024 A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions IEEE Access