







# RLSTM: A Novel Residual and Recurrent Network for Pedestrian Action Classification

Soulayma Gazzeh<sup>1,2</sup>(✉) , Liliana Lo Presti<sup>2</sup> , Ali Douik<sup>1</sup> ,  
and Marco La Cascia<sup>2</sup> 

<sup>1</sup> NOCCS Laboratory, ENISO, University of Sousse, Sousse, Tunisia  
soulayma.gazzeh@eniso.u-sousse.tn

<sup>2</sup> Department of Engineering, University of Palermo, Palermo, Italy

**Abstract.** Properly training LSTMs requires long time and extensive amount of data. To improve the training of these models, this paper proposes a novel residual and recurrent neural network, Resnet-LSTM, for spatio-temporal pedestrian action recognition from image sequences. The model includes a novel layer, called MapGrad, whose goal is improving stationarity of the feature map sequences processed by the ConvLSTM. The paper demonstrates the effectiveness of the proposed model and the MapGrad layer in the spatio-temporal classification of pedestrian actions through an ablation study and comparison with state-of-the-art methods. Overall, RLSTM achieves an accuracy value of 88% and an average precision of 94% on the JAAD dataset, which is a widely used benchmark in the field. Finally, the paper empirically analyzes the effect of increasing input sequence length on standing action recognition, showing that the proposed method yields a recall of 93%.

**Keywords:** Pedestrian action recognition · Time series data · LSTM · Spatio-Temporal features

## 1 Introduction

Autonomous driving (AD) is a rapidly evolving field in computer vision whose primary focus is to ensure the safety of pedestrians, who often interact with vehicles in complex and unpredictable ways [12, 13]. A crucial task for autonomous vehicles is to recognize whether or not a pedestrian is crossing the road. Preliminary steps to achieve this, involve detecting and tracking pedestrians and identifying *walking* and *standing* actions. The latter task, pedestrian action recognition (PAR), is challenging when using mobile cameras. Indeed, motion blur, the dynamic background of the street scene, variations in the pedestrians' visual appearance, and frequent occlusions complicates the action classification task. To address the problem, techniques derived from time series analysis are often employed, which allow for the processing of frame sequences to extract motion and changes in the scene over time [4–6]. However, meaningful motion patterns

are difficult to model due to the complex interference between pedestrians' and vehicle's movements. Indeed, changes in vehicle speed and direction can lead to changes in the apparent motion of pedestrians. To model such complex temporal dependencies, Long-Short Term Model (LSTM) [7] is often used with time series due to its ability to capture both short- and long-term dependencies over time. ConvLSTM [8] has instead been used to process image sequences. In [13], LSTMs are used for action recognition despite these models are difficult to train, in the sense that training requires long time and extensive training data.

In this paper, we propose a novel end-to-end trainable deep architecture that leverages residual layers [9] and ConvLSTM for PAR. Our architecture takes advantage of a novel layer, *MapGrad*, that improves the extraction of temporal features. MapGrad builds on preprocessing techniques adopted in time series analysis and, in the context of PAR, helps improve learning of an LSTM and reduce the negative effect of camera motion on feature maps without increasing the number of model parameters. To achieve this goal, MapGrad computes the forward difference of the feature maps extracted over time from a convolutional network, thus improving the stationarity of that sequence while, at the same time, highlighting temporal feature changes.

In addition, we emphasize that the length of the input sequence (SL) should be carefully selected when designing a PAR system, as it directly impacts the real-time performance of the model and the recall of the standing action.

In summary, our contributions in this paper are:

- A novel residual and recurrent architecture (RLSTM) for PAR;
- A novel layer, MapGrad, that pre-processes feature maps before feeding a LSTM. Our ablation study shows that, in our experiments, MapGrad contributes to increase the accuracy in classification of more than 17% when processing input sequences of 7 frames;
- A study on the effect of increasing the SL on the recognition of standing pedestrians with respect to the real-time constraints of the PAR system.

The plan of the paper is as follows. Section 2 reviews works on action recognition with mobile cameras. Section 3 describes in detail the proposed architecture and the MapGrad layer. Section 4 presents experimental results on a public available benchmark and the comparison to the existing state-of-the-art techniques. Finally, Sect. 5 summarizes our main findings and describes future works.

## 2 Related Work

Action recognition is a widely studied field in computer vision that aims to automatically recognize human actions from image sequences. These approaches have been applied to different domains such as sports analysis [14], surveillance [15] and AD [16].

In the context of AD, the main challenges to address concern the dynamic camera motion and the complex motion patterns of pedestrians. Several deep learning (DL) architectures have been proposed for PAR using mobile cameras,

including 2D/3D convolutional networks [6], recurrent networks [17], and hybrid models combining both approaches [5].

To improve pedestrian safety, several studies have investigated different approaches to detect crossing intention by considering environmental factors [18,20] and visual cues, which include analyzing the body posture [18,19] and pedestrians’ motion patterns [21].

Recognizing atomic actions, such as walking and standing, is an important step towards more complex pedestrian activities recognition. For instance, the posture and motion features of pedestrians while walking or standing can provide important cues for inferring their intention to cross the road. Due to the difficulty in recognizing standing from walking, a limited number of studies [1–3] have investigated this task. This is because the visual similarity between these two actions poses a significant obstacle, especially in the presence of motion blur.

Our proposed approach differs from the previous papers as we adopt a residual and recurrent network to process a sequence of image crops of the detected pedestrians. Compared to the two-stream CNN used in [2], our approach has a simpler architecture, which is computationally more efficient. Additionally, our approach does not require pedestrian pose keypoints, as in [3], making it more robust to changes in pose and viewpoint. Finally, our use of an LSTM layer allows for the incorporation of temporal information, which is not possible in the cropbox-based AlexNet architecture employed in [1].

### 3 Proposed Method

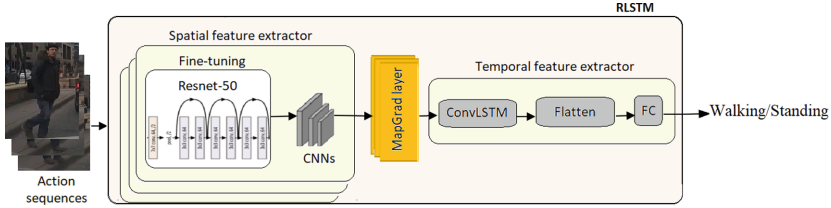
Given a video acquired by the camera mounted on the vehicle, we assume that a visual tracking algorithm, for instance DeepSort [10] or Track R-CNN [11], detects and tracks pedestrians. Our goal is to classify sequences of image crops to infer the pedestrians’ actions. These action sequences can be modeled as 4D tensors of size  $[L \times H \times W \times C]$ , with  $C$  indicating the number of channels of the  $L$  images with height  $H$  and width  $W$ .

Our model, which we refer to as Spatio-Temporal Resnet-LSTM (RLSTM), takes in input action sequences and infers if the pedestrian is walking or standing.

#### 3.1 Spatio-Temporal RLSTM

As shown in Fig. 1, RLSTM combines both spatial and temporal information of an input sequence to achieve robust PAR. It is composed of two sub-networks.

The first sub-network is the spatial feature extraction module, and focuses on time-independent spatial features extraction, since it computes convolutional features on each image crop in the action sequence. The module employs the first two residual blocks of a pre-trained ResNet50, and two additional convolutional layers before the output is passed to the temporal feature extraction module. The second sub-network models behavioral features from the action sequences by using a ConvLSTM2D layer, and uses them to classify the input action sequence.



**Fig. 1.** Our proposed RLSTM model for PAR includes a spatial feature extraction module from the input action sequences, and a temporal feature extraction module for modeling behavioral features. Our MapGrad layer is inserted between these modules to transform the feature maps to be processed by the ConvLSTM layer.

In this module, convolutional and LSTM memory cells learn spatio-temporal patterns in the input sample.

In between the two modules, the MapGrad layer has the goal of transforming the extracted spatial features in a way that is suitable for the ConvLSTM2D layer to learn the pedestrians’ behavioral patterns.

### 3.2 MapGrad Layer

In AD, the camera moves with the vehicle. Thus, the background of each frame changes dynamically, making it difficult to accurately model motion patterns when pedestrians are standing or walking. To address this problem, it is important both to extract suitable spatial features and, in the meantime, to take into account the temporal context in which the pedestrian action develops.

A common preprocessing in time series analysis is called *de-meaning*, which is to make the series zero-mean. Inspired by this, we implemented a layer to make zero-mean the sequences of feature maps that feed our ConvLSTM2D layer. In our formulation, the mean is computed only over the temporal dimension.

Given a spatial feature map  $F_t$  corresponding to the  $t$ -th frame, we element-wise subtract the mean feature map  $M_t$  in a temporal window to ensure that features are centered around zero, allowing subsequent analysis to focus on relative changes in pixel values. It helps normalizing the brightness levels and reducing the impact of lighting variations.

Our experimental results show that this feature map pre-processing contributes to greatly improve the learning of the ConvLSTM2D. Probably, making the feature maps zero-mean, contributes to reducing the effects of the dynamically changing background, and allows the model to focus on the spatio-temporal patterns relevant to the classification of pedestrians’ action.

Aside from making the spatial feature maps zero-mean, our MapGrad layer uses temporal differentiation. This technique is adopted in time series analysis to improve the stationarity of the series [22]. It consists in computing the forward difference of the feature maps  $F_t$  extracted from consecutive image crops of the input sequence. In this way, MapGrad highlights the temporal changes between feature maps, which helps to isolate the pedestrian motion patterns.

Temporal differencing can be represented as:

$$D_t = F_t - F_{t-1} \quad (1)$$

where  $D_t$  is the difference between the feature maps, and the output of the MapGrad layer.

**Implementation Details.** To train our model, we use the ADAM optimizer with a binary cross-entropy loss function and a batch size of 10. To prevent overfitting, we incorporate dropout regularization with a rate of 0.5 after each convolutional layer to enhance the stability and convergence of the training process. Furthermore, we lower the initial learning rate of  $10^{-3}$  to  $10^{-6}$  for further optimization. During training, we employ early stopping to prevent overfitting.

## 4 Experimental Results

This section details the experiments conducted to demonstrate the effectiveness of our proposed model. We first describe the dataset used in the experiments, the experimental protocol and the data pre-processing. To highlight the contribution of our novel MapGrad layer and of the overall model, we conducted ablation studies. We also trained our model on sequences of varying length. Finally, we compared our best trade-off with the state-of-the-art.

### 4.1 Dataset

This work employs Joint Attention in Autonomous Driving (JAAD) dataset [1], which is widely used in pedestrian behavior recognition research. The dataset includes 346 short videos (5–20 s long), for a total of 82K frames. Videos are acquired at 30 frames per second, and each frame is annotated for pedestrian behaviors. Overall, the dataset contains annotations for 686 pedestrians.

The ground-truth annotations include the pedestrians’ bounding boxes and behavioral tags like, for instance, actions (i.e., *standing* and *walking*) or behavioral attributes (i.e., *“cross”* and *“look”*). Only action classes are used in this study. Each pedestrian may perform multiple actions within a single video, switching from standing to walking or vice versa.

JAAD dataset suffers from imbalanced classes with 974 standing action sequences and 2524 walking. Variation in visibility on the road (Fig. 2), weather conditions, and partial or full occlusions (Fig. 3) between pedestrians or due to objects in the scene can make accurate recognition of pedestrian actions difficult.

**Evaluation Metrics.** To provide a comprehensive understanding of our RLSTM performance, we report several evaluation metrics such as accuracy value, F1-score, precision, recall, and average precision (AP).

The accuracy value measures the number of correctly classified samples, while the F1-score, precision, and recall metrics provide a more detailed assessment of the model’s performance per class. We also report the AP metric, which measures the area under the precision-recall curve.



**Fig. 2.** The figure shows two images captured at different day time. As shown, this results in changes of the visibility on the road.



**Fig. 3.** The figure shows images of a pedestrian taken while the vehicle is moving. The pedestrian is severely occluded, which makes harder recognizing his/her action.



**Fig. 4.** The figure shows a sequence of images cropped around a walking pedestrian.

**Data Preparation.** We implemented a data generator to facilitate data augmentation while reducing storage and computational requirements. Our data generator leverages tracking data (pedestrians’ bounding boxes) and class labels (walking/standing), to generate a sequence of  $N$  image crops to feed our model. In our experiments,  $N$  was set to 7, 10 and 15.

To ensure the quality of the resulting image sequences, samples with full occlusion are filtered out. To maintain the aspect ratio of the pedestrian detection, the square crops of the pedestrian images also include a larger area surrounding the pedestrians (Fig. 4). Image crops are then rescaled to a  $(224 \times 224)$  size. The data generator produces balanced batches of action sequences by uniformly sampling over the time dimension. Since our model includes pre-trained residual blocks, input images were normalized by subtracting the mean RGB values and scaling the pixel values in the range  $[-1, 1]$ .

## 4.2 Ablation Study

Our model includes several components and layers. Table 1 reports the ablation study conducted to evaluate the impact of each component on PAR.

Each row of the table refers to a different model and all experiments are conducted by considering a sequence length equals to 7.

**ResLSTM** refers to our baseline model including residual blocks from the pretrained ResNet50, Conv2D layers, and a ConvLSTM.

**ResLSTM + BN + D** refers to the regularized version of the previous ResLSTM model by using batch normalization (BN) and dropout layers. In particular, we adopted a BN layer after each convolutional layer to stabilize the network and improve learning. We noticed that including a BN just before the ConvLSTM2D layer was more effective in preserving temporal information, allowing the network to learn more robust features. Regularization improved the recall of the standing action.

**RConv3D + BN + D** refers to a regularized model including residual blocks from the pretrained ResNet50, Conv2D layers and a Conv3D layer that handles the time dimension of the feature maps. This experiment serves to highlight the contribution of the ConvLSTM to the overall accuracy of the model. As shown in the table, this model achieves similar performance to that of the regularized ResLSTM suggesting that the ConvLSTM is unable to learn the dynamics underlying to the input sequence from the extracted spatial features.

**RConv3D + MapGrad** refers to a model including residual blocks from the pretrained ResNet50, Conv2D layers and a Conv3D layer. In this case, no regularization technique is adopted. Instead, between the spatial feature extractor and the Conv3D layer we include our MapGrad layer. The MapGrad layer contributes to improve the accuracy value by about the 4.17% compared to the regularized RConv3D model. While the recall for the walking action increases, the one for the standing action decreases. This may indicate that the Conv3D has issues in discriminating between the (dynamic) background and the standing pedestrian. In our experiments, we noted that, when using the MapGrad layer, the impact of BN layers is very limited.

**Centering Sequence of Maps** refers to ResLSTM including centering the feature maps along the time dimension (i.e., making zero-mean the map sequence). As Table 1 shows, centering the map sequence improves over the ResLSTM.

**RLSTM (with MapGrad)** refers to our proposed model. It is similar to the ResLSTM model but includes the MapGrad layer between the spatial feature extractor and the ConvLSTM layer (Fig. 1). As shown in the table, MapGrad contributes to increase the accuracy in classification of more than 17% compared to the regularized ResLSTM model, and of about 21.1% compared to the simpler ResLSTM model. With respect to the RConv3D + MapGrad model, the increase in the accuracy value is of about 14.7%. While ConvLSTM and Conv3D were initially getting similar results, after the introduction of MapGrad in the model, the performance of ConvLSTM is much higher than Conv3D. Therefore, the preprocessing of feature map sequences to improve the stationarity of the series appears to have a positive effect on the training of the LSTM layer.

**Impact of the Sequence-Length.** Table 2 compares the performance achieved by our model when the SL assumes values 7, 10 and 15. As shown in the table, increasing the SL improves standing action recognition since the network receives

**Table 1.** Ablation studies

Models	Accuracy	F1-score	Precision	Recall		AP
				Standing	Walking	
ResLSTM (no preprocessing)	71	70.5	72.5	59	83	75
ResLSTM+BN+D	73	72.5	72.5	71	73	80
RConv3D+BN+D	72	71.5	72	73	70	74
RConv3D+MapGrad	75	75.5	75.5	69	82	83
ResLSTM + Centering Sequence of Maps	84	85	84	85	83	90
RLSTM (with MapGrad) (ours)	<b>86</b>	<b>87</b>	<b>87.5</b>	<b>90</b>	<b>85</b>	<b>92</b>

**Table 2.** Results achieved by RLSTM when varying the Sequence-Length

Model	Seq. Length (frames)	Observed ms	Accuracy	F1-score	Precision	Recall		AP
						Standing	Walking	
RLSTM (Ours)	7	200	86	87	87.5	90	85	92
	10	300	88	88.5	88	93	84	94
	15	500	90	90.5	90.5	94	87	96

more temporal context, and can capture the nuances of the standing action. The column *Observed ms* shows the length in milliseconds of the observed sequences. While the best performance is obtained when using 15 frames, the observed ms equals half a second, which may not ensure a quick and accurate decision in the context of AD. We note that 10-frame sequences are used in previous works [1, 2].

### 4.3 Comparison with the State-of-the-Art

Table 3 reports the comparison of the results achieved by our model and works at the state-of-the-art on the JAAD dataset.

Our approach outperformed other methods such as the Two-Stream CNN approach in [2], the AlexNet model in [1], and the recurrent architecture in [3]. We note here that the methods in [2] uses multiple inputs. Similarly, the work in [3] takes in input also the pedestrian’s pose keypoints. On the contrary, our method only takes in input sequences of image crops of the pedestrian. Despite the input of our model is simpler, it achieves superior results in all metrics with respect to the work in [3], and a comparable accuracy value with respect to [2]. Whilst it is known that recognizing the standing action is difficult [1], our approach achieves a 93% of recall value for this action that, at the best of our knowledge, is the highest value achieved on the JAAD dataset.



**Table 3.** Comparison to the state of the art models

Model	Input	AP	Accuracy	Recall	
				Standing	Walking
Rasouli et al. [1]	Cropboxes	83	-	-	-
Marginean et al. [3]	Pose keypoints	-	77	76	76
Park et al. [2]	First frame	-	<b>88</b>	72	<b>91</b>
	Flow images Position information				
<b>RLSTM (Ours)</b>	Cropboxes	<b>94</b>	<b>88</b>	<b>93</b>	84

## 5 Conclusions and Future Work

In this paper, we proposed RLSTM, a spatio-temporal neural network for the classification of walking and standing actions in AD. RLSTM includes residual blocks, convolutional and ConvLSTM layers, and MapGrad, namely a novel feature map preprocessing layer. Our experiments show that the introduction of MapGrad to the model improves the learning of ConvLSTM without increasing the number of parameters of the model. On the JAAD dataset, MapGrad contributes to increase the accuracy in classification of more than 17%. Our experiments also show that increasing the SL significantly improves our model’s ability to recognize standing actions in real time, achieving a recall of 93%.

In future work, we plan to explore more information, such as the velocity and ego-motion of the vehicle, to improve our model performance. Our final goal will be the recognition of the pedestrians’ crossing intentions and we will study if it is possible to extend RLSTM to predict the pedestrians’ intentions.

## References

1. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 206–213 (2017)
2. Park, S.K., Chung, J.H., Pae, D.S., Lim, M.T.: Binary dense SIFT flow based position-information added two-stream CNN for pedestrian action recognition. *Appl. Sci.* **12**(20), 10445 (2022)
3. Marginean, A., Brehar, R., Negru, M.: Understanding pedestrian behaviour with pose estimation and recurrent networks. In: 2019 6th International Symposium on Electrical and Electronics Engineering (ISEEE), pp. 1–6. IEEE (2019)
4. Yang, B., Zhan, W., Wang, P., Chan, C., Cai, Y., Wang, N.: Crossing or not? Context-based recognition of pedestrian crossing intention in the urban environment. *IEEE Trans. Intell. Transp. Syst.* **23**(6), 5338–5349 (2021)
5. Yang, D., Zhang, H., Yurtsever, E., Redmill, K.A., Özgüner, Ü.: Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Trans. Intell. Veh.* **7**(2), 221–230 (2022)

6. Chen, T., Tian, R., Ding, Z.: Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3103–3109 (2021)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
8. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649. IEEE (2017)
11. He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., Sun, J.: Track R-CNN: multiple object tracking with track-RCNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10838–10847 (2020)
12. Liu, B., et al.: Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robot. Autom. Lett.* **5**(2), 3485–3492 (2020)
13. Guo, D., Mordan, T., Alahi, A.: Pedestrian stop and go forecasting with hybrid feature fusion. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 940–947. IEEE (2022)
14. Qi, M., Qin, J., Wu, Y., Yang, Y.: Imitative non-autoregressive modeling for trajectory forecasting and imputation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12736–12745 (2020)
15. Mangalam, K., et al.: It is not the journey but the destination: endpoint conditioned trajectory prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 759–776. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-58536-5\\_45](https://doi.org/10.1007/978-3-030-58536-5_45)
16. Noguchi, C., Tanizawa, T.: Ego-vehicle action recognition based on semi-supervised contrastive learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5988–5998 (2023)
17. Lian, J., Yu, F., Li, L., Zhou, Y.: Early intention prediction of pedestrians using contextual attention-based LSTM. *Multimedia Tools Appl.* **82**(10), 14713–14729 (2023)
18. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs (2020)
19. Cadena, P.R.G., Yang, M., Qian, Y., Wang, C.: Pedestrian graph: pedestrian crossing prediction based on 2D pose estimation and graph convolutional networks. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 2000–2005. IEEE (2019)
20. Moreno, E., et al.: Pedestrian crossing intention forecasting at unsignalized intersections using naturalistic trajectories. *Sensors* **23**(5), 2773 (2023)
21. Yang, C., Pei, Z.: Long-short term spatio-temporal aggregation for trajectory prediction. *IEEE Trans. Intell. Transp. Syst.* **24**(4), 4114–4126 (2023)
22. <https://www.otexts.org/fpp/8/1>