Visible and Thermal Camera-based Jaywalking Estimation using a Hierarchical Deep Learning Framework

V. John[1], A. Boyali[2], S. Thompson[2], A. Lakshmanan[1], and S. Mita[1]

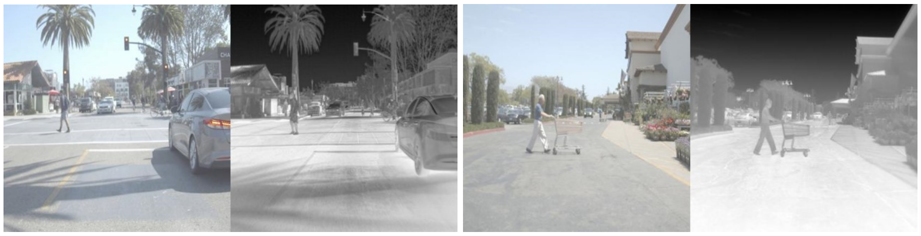
[1] Toyota Technological Institute, Japan
{vijayjohn,smita}@toyota-ti.ac.jp,
lakshmanan.annamalai@outlook.com
[2] Tier IV, Japan
{ali.boyali,simon.thompson}@tier4.jp

Abstract. Jaywalking is an abnormal pedestrian behavior which significantly increases the risk of road accidents. Owing to this risk, autonomous driving applications should robustly estimate the jaywalking pedestrians. However, the task of robustly estimating jaywalking is not trivial, especially in the case of visible camera-based estimation. In this work, a two-step hierarchical deep learning formulation using visible and thermal camera is proposed to address these challenges. The two steps are comprised of a deep learning-based scene classifier and two scene-specific semantic segmentation frameworks. The scene classifier classifies the visible-thermal image into legal pedestrian crossing and illegal pedestrian crossing scenes. The two scene-specific segmentation frameworks estimate the normal pedestrians and jaywalking pedestrians. The two segmentation frameworks are individually trained on the legal or illegal crossing scenes. The proposed framework is validated on the FLIR public dataset and compared with baseline algorithms. The experimental results show that the proposed hierarchical strategy reports better accuracy than baseline algorithms in real-time.

1 Introduction

Autonomous driving and ADAS applications, which aim to increase the safety of road users, have received significant attention from the research community [1–3]. Environment perception is a key task for autonomous driving. Examples of environment perception include pedestrian detection [4], road surface segmentation [3], pedestrian behavior estimation [5] etc. Jaywalking is an example of abnormal pedestrian behavior which occurs when pedestrians walk or cross the road at locations, disregarding traffic rules.

Owing to the risk of accidents associated with this behavior, autonomous driving applications should robustly estimate jaywalking pedestrians. However, jaywalking estimation is not a trivial task with several challenges, especially when the visible camera is used. The challenges include variations in illumination, appearance similarity between *normal* pedestrian behavior and *abnormal* pedestrian behavior (Fig 1), appearance variations in *legal* pedestrian crossing points or scenes (Fig 2]), environmental noise etc.



(a) Normal Behavior: Legal Pedestrian Crossing Scene

(b) Jaywalking: Illegal Pedestrian Crossing Scene

Fig. 1. Appearance similarity in legal and illegal crossing scene.

A naive vision-based approach to solving this problem involves segmenting or detecting pedestrians and legal-illegal crossing image regions, and using the segmentation results to estimate the pedestrian behavior. However, such an approach is limited by the appearance similarity between the *normal* and *abnormal* pedestrian behavior in certain scenes (Fig 1. In such scenes, the pedestrian crossing segmentation can be used to classify the pedestrian behavior. But as shown in Fig 2, the pedestrian crossing estimation is by itself a challenging problem owing to varying pedestrian crossing regions.

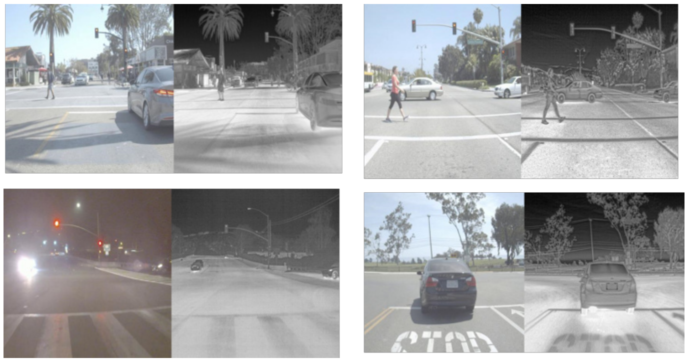
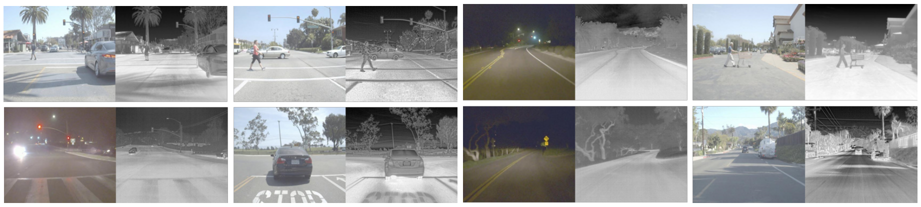


Fig. 2. Challenges in jaywalking estimation due to variations in the pedestrian crossing markers.

To address these challenges, a two-step hierarchical framework is proposed using the visible and thermal camera. The sensor fusion of the thermal and visible camera address the challenges associated with the visible camera such as illumination variations and sensor noise [6]. The other challenges are addressed by the hierarchical framework.

The hierarchical framework is comprised of a classification step and a semantic segmentation step. The classification step is formulated using a single deep learning-based scene classifier which classifies the driving scene into a *legal* pedestrian crossing



(a) Legal Pedestrian Crossing Scenes (b) Illegal Pedestrian Crossing Scenes

Fig. 3. Scene Partitions for the Hierarchical Framework.

scene or a *illegal* pedestrian crossing scene (Fig 3). The semantic segmentation step is formulated using two scene-specific semantic segmentation frameworks. The first semantic segmentation framework is trained on the *legal* pedestrian crossing scene. The second semantic segmentation framework is trained on the *illegal* pedestrian crossing scenes. These semantic segmentation frameworks estimate the normal and jaywalking pedestrians in a given image. The proposed framework is validated on the FLIR public dataset, and is compared with baseline algorithms. The experimental results show that the hierarchical fusion framework is better than the baseline algorithms (Sec 4), while reporting real-time computational complexity.

To the best of our knowledge, the main contribution of our work are as follows:

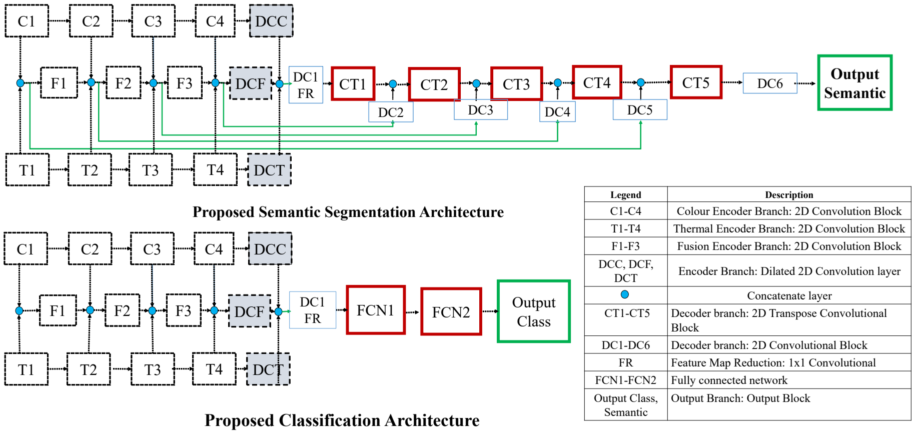
- A hierarchical learning framework for jaywalking estimation
- Visible and thermal camera fusion for jaywalking estimation

The reminder of the paper is structured as follows. In Section 2 we review the literature in jaywalking estimation. The hierarchical learning framework is presented in Section 3, and the experimental results are presented in Section 4. Finally, we summarize our work in Section 5.

2 Related Work

Pedestrian detection is an important precursor for pedestrian behavior estimation, and has been well-researched [7–10]. Pedestrian detection methods are categorized into methods based on hand-crafted features [7–9] or methods based on deep learning [10, 11]. Pedestrian spatial, contextual or temporal information obtained from pedestrian detection is used for pedestrian behavior estimation [12–14]. Pedestrian behavior estimation is based on probabilistic modeling [12, 15–17], deep learning models [18–20], and traditional frameworks that incorporate spatial contextual cues [14]. The different pedestrian behavior estimation frameworks are surveyed in the work of Santosh et al. [5].

Probabilistic and traditional frameworks model pedestrian behaviour using extracted pedestrian features. The modeled pedestrian behavior are used to identify anomalous behavior [12, 14–17]. Roshtkari et al. [14] model the pedestrian spatial-temporal information within a bag-of-words framework, which are then used to identify anomalous



Legend	Description
C1-C4	Colour Encoder Branch: 2D Convolution Block
T1-T4	Thermal Encoder Branch: 2D Convolution Block
F1-F3	Fusion Encoder Branch: 2D Convolution Block
DCC, DCF, DCT	Encoder Branch: Dilated 2D Convolution layer
●	Concatenate layer
CT1-CT5	Decoder branch: 2D Transpose Convolutional Block
DC1-DC6	Decoder branch: 2D Convolutional Block
FR	Feature Map Reduction: 1x1 Convolutional
FCN1-FCN2	Fully connected network
Output Class, Semantic	Output Branch: Output Block

Fig. 4. Architecture of the Classification and Semantic Segmentation Network.

C1: C (64, 3, 1) P (2)	**C2:** C (128, 3,1) P (2)	**C3:** C (256, 3, 1) P (2)	**C4**: C (256, 3, 1) P (2)	**DCC**: C(256, 3) D(2)
T1: C (64, 3, 1) P (2)	**T2:** C (128, 3,1) P (2)	**T3:** C (256, 3, 1) P (2)	**T4**: C (256, 3, 1) P (2)	**DCT:** C(256, 3) D(2)
F1: C (256, 3, 1) P (2)	**F2:** C (256, 3, 1) P (2)	**F3**: C (256, 3, 1) P (2)	**DCF:** C(256, 3) D(2)	
DC1: C (256, 3, 1) Pad (S), **FR:** C(64,1,1)	**DC2:** C (256, 3, 1) Pad (S)	**DC3:** C (256, 3, 1) Pad (S)	**DC4:** C (256, 3, 1) Pad (S)	**DC5:** C (256, 3, 1) Pad (S)
CT1: TC (256, 5, 1)	**CT2:** TC (256, 4, 2)	**CT3:** TC (256, 4, 2)	**CT4:** TC (256, 5, 2)	**CT5:** TC (256, 4, 2)
DC6: C (64, 1, 1)	**FN1**: 512, **FN2:** 256	**Output Semantic:** Multiclass	**Output Class:** Binary	

Fig. 5. 2D Convolutional layer parameters: C(filters, kernel size, stride), D(dilation rate), Pad(S) "same" padding; Max pooling layer parameters: P (kernel size); 2D transpose convolutional layer parameters: TC (filters, kernel size, stride).

behavior. In the work of Bera et al. [17] where the pedestrian global and local features are extracted and used within a Bayesian framework to identify anomalous behavior. In recent years, deep learning models report state-of-the-art accuracy for different perception tasks [2, 3, 21]. Medel et al. [19] use an end-to-end composite Convolutional Long Short-Term Memory (LSTM) to estimate anomalous behavior. A similar approach is proposed by Xu et al. [22] using the Resnet and LSTM.

Compared to literature, in our work, we adopt a hierarchical deep learning framework using thermal and visible cameras to estimate jaywalking.

3 Algorithm

A two-step hierarchical framework using visible and thermal camera is proposed to identify jaywalking pedestrians. The initial step is formulated using a single deep learning-based classification network, while the second step is formulated using two semantic

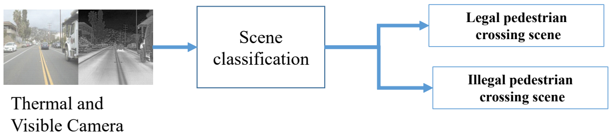

Thermal and
Visible Camera
Scene classification
Legal pedestrian crossing scene
Illegal pedestrian crossing scene

Fig. 6. Hierarchical Framework Step 1: Classification.

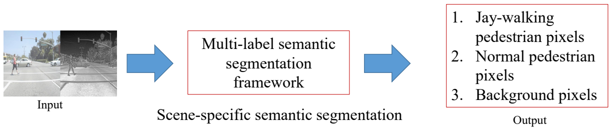

Input
Multi-label semantic segmentation framework
Scene-specific semantic segmentation
1. Jay-walking pedestrian pixels
2. Normal pedestrian pixels
3. Background pixels
Output

Fig. 7. Hierarchical Framework Step 2: Semantic Segmentation.

segmentation networks. The hierarchical framework is formulated to reduce the following pedestrian behavior estimation errors: a) appearance similarities between jaywalking pedestrians and pedestrians in legal crossing scenes (Fig 1); b) appearance variations across different *legal* pedestrians crossing scenes (Fig 2).

3.1 Classification Step

The first step in the hierarchical framework is a classification step, which categorises the driving scene into either a *legal* pedestrian crossing scene or a *illegal* pedestrian crossing scene (Fig 3). An illustration of the classification step is shown in Fig 6.

Architecture A deep learning-based visible and thermal camera fusion architecture is used for the classification step. The architecture comprises of feature extraction, classification and output layers. The feature extraction layer contains three branches with two branches for visible camera and thermal camera feature extraction and one branch for fusion.

The two feature extraction branches contain 5 blocks. The first four blocks each contain a $2D$ convolutional layer with batch-normalization followed by a max-pooling layer. The fifth block contains a 2-dilated convolutional layer with batch-normalization.

The fusion branch also has 5 blocks. The first three blocks each contain a $2D$ convolutional layer with batch-normalization followed by a max-pooling layer. The fourth block contains a 2-dilated convolutional layer with batch-normalization. The final block contains a $1x1$ convolutional layer functioning as a feature map reducing layer. The output of the feature map reducing layer is given as an input to the classification layer.

The classification layer contains two fully connected layers with 512 and 256 neurons and relu activation function. The final output layer contains 1 neuron and performs binary classification using the sigmoid activation.

3.2 Semantic Segmentation Step

This step contains two multi-class semantic segmentation networks. The first network is trained on the *legal* pedestrian crossing scenes, while the second network is trained on the *illegal* pedestrian crossing scenes. The trained semantic segmentation framework categorizes the image pixels as normal pedestrian, jaywalking pedestrian or background. An illustration of the semantic segmentation is shown in Fig 7.

Architecture A deep learning-based encoder-decoder architecture is utilized for the visible and thermal fusion and semantic segmentation. The encoder layers are the same as the feature extraction layers in the aforementioned classification network. This layer contains three feature extraction branches for visible camera feature extraction, thermal camera feature extraction and feature fusion.

The decoder layers contain 4 transpose convolutional layers and 5 convolutional layers. The transpose convolutional layers with batch-normalization upsample the encoder feature maps from the fusion branch. A skip connection is used to transfer the encoder feature maps to the decoder branches.

The output of the last decoder layer is fed into the output layer which performs the multiclass semantic segmentation using the softmax activation.The detailed architecture and the parameters are given in Fig 4 and Fig 5.

3.3 Hierarchical Framework: Training

The hierarchical framework is trained on the FLIR public dataset. Manually selected frames from the dataset are manually partitioned into *legal* pedestrian crossing and *illegal* pedestrian crossing scenes. The visible and thermal camera images for these frames are manually registered to ensure pixel-to-pixel correspondence.

During training, the single classification network is trained with all the frames (both legal and illegal pedestrian crossing scenes). On the other hand, the first semantic segmentation network is trained on the *legal* pedestrian crossing scenes, and the second semantic segmentation network is trained on the *illegal* pedestrian crossing scenes.

3.4 Hierarchical Framework: Testing

During testing, the registered thermal and visible camera images are first given as an input to the trained classification network. The classification network classifies the input images as either *legal* or *illegal* pedestrian crossing scene, assigning scene label c.

The estimated c-th scene label is used to retrieve the corresponding *trained* c-th scene-based semantic segmentation network (*illegal* or *legal*) for jaywalking estimation. Following the semantic segmentation network retrieval, the registered thermal and visible camera images are re-given as the network input and the jaywalking pedestrians are estimated. An overview of the testing is shown in Fig 8.

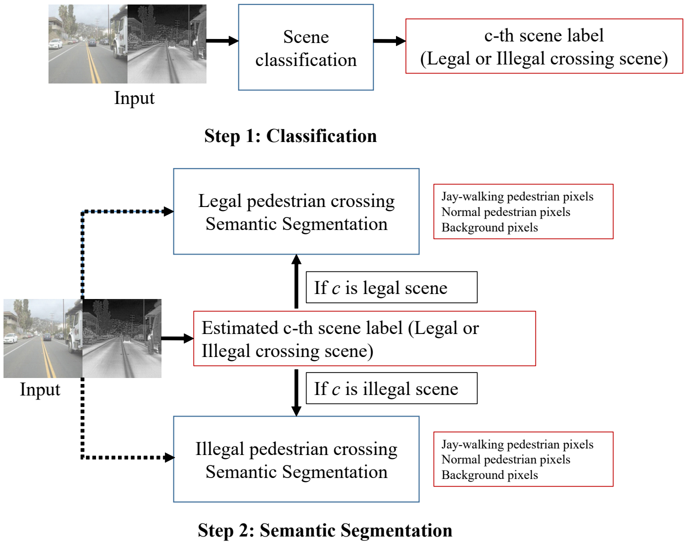

Input
Scene classification
c-th scene label
(Legal or Illegal crossing scene)
Step 1: Classification
Legal pedestrian crossing Semantic Segmentation
Jay-walking pedestrian pixels
Normal pedestrian pixels
Background pixels
If c is legal scene
Input
Estimated c-th scene label (Legal or Illegal crossing scene)
If c is illegal scene
Illegal pedestrian crossing Semantic Segmentation
Jay-walking pedestrian pixels
Normal pedestrian pixels
Background pixels
Step 2: Semantic Segmentation

Fig. 8. Hierarchical Framework.

4 Experimental Results

We validate the proposed framework on a manually selected subset from the FLIR public dataset. Our dataset contains 2799 training frames and 313 testing frames. Our algorithm was implemented using Tensorflow 2 on an Ubuntu 18.04 machine with two Nvidia 1080 GPUs. The algorithm was compared with six baseline algorithms and validated on the dataset. The proposed framework and the baseline algorithms were trained with batch size 4 for 30 epochs. The performance is measured using the pixel segmentation accuracy, the intersection-over-union (IOU) measure and classification accuracy. The IoU measure is the calculated from the overlap between the prediction and segmentation divided by the area of their union. We next briefly review the different baseline algorithms.

Hierarchical MFNet: The first baseline is a hierarchical framework based on the MFNet [23]. The MFNet is a visible-thermal camera based semantic segmentation framework with encoder-decoder architecture.

For comparative analysis, the MFNet encoder was used for feature extraction in the classification network. On the other hand, the entire MFNet was used for the two semantic segmentation networks.

Hierarchical Fusenet: The second baseline is a hierarchical framework based on the Fusenet [21]. The original Fusenet is a visible-depth based semantic segmentation framework with encoder-decoder architecture. For the comparative analysis, the original Fusenet input layers were modified for visible-thermal camera input.

The modified Fusenet's encoder was used for feature extraction in the classification network. The entire modified Fusenet was used for the two semantic segmentation networks in the second step.

Hierarchical with Single Encoder: The third baseline is a hierarchical framework where the classification network feature maps are reused for the semantic segmentation network.

The classification network is first trained on the dataset, and the feature maps for the dataset are obtained. The classification network feature maps are used for the semantic segmentation network training. The feature maps are the outputs of the $DC1-FR$ block or the final fusion encoder block. Since the feature maps are re-used, there is no encoder layer in this baseline's semantic segmentation networks. The classification feature maps are directly given as input to the decoder branches and network is trained.

Hierarchical Visible: The fourth baseline is a hierarchical framework using the visible camera input as the sole input. The encoder branch of the proposed network for both the classification and segmentation networks contain a single branch for the visible camera.

Hierarchical Thermal: The fifth baseline is a hierarchical framework using the thermal camera as the sole input. The encoder branch of the proposed network for both the classification and segmentation networks contain a single branch for the thermal camera.

End-to-End Semantic Segmentation: The sixth baseline is a naive end-to-end semantic segmentation framework, formulated to evaluate the hierarchical framework. The performance of the hierarchical framework is compared with the end-to-end semantic segmentation framework.

In this baseline network, the jaywalking pedestrians are directly estimated using a "single" semantic segmentation framework. Unlike, the hierarchical semantic segmentation networks, this "single" semantic segmentation framework is trained with entire dataset, both legal and illegal pedestrian crossing scenes.

4.1 Comparative Analysis

The performance of the different algorithms are tabulated in Table 1-Table 3. The results are illustrated in Fig 9 and Fig 10.

Hierarchical Framework with Varying Base Deep Learning Models: In Table 1, the results of the hierarchical framework with varying base deep learning models are tabulated. The results show that the hierarchical framework with the proposed deep learning architecture is better than the hierarchical framework with the MFNet and the Fusenet.

Hierarchical Framework with Varying Sensors: The results of the hierarchical framework with varying sensors are tabulated in Table 2. As expected, the advantages of the visible-thermal camera sensor fusion are clearly demonstrated.

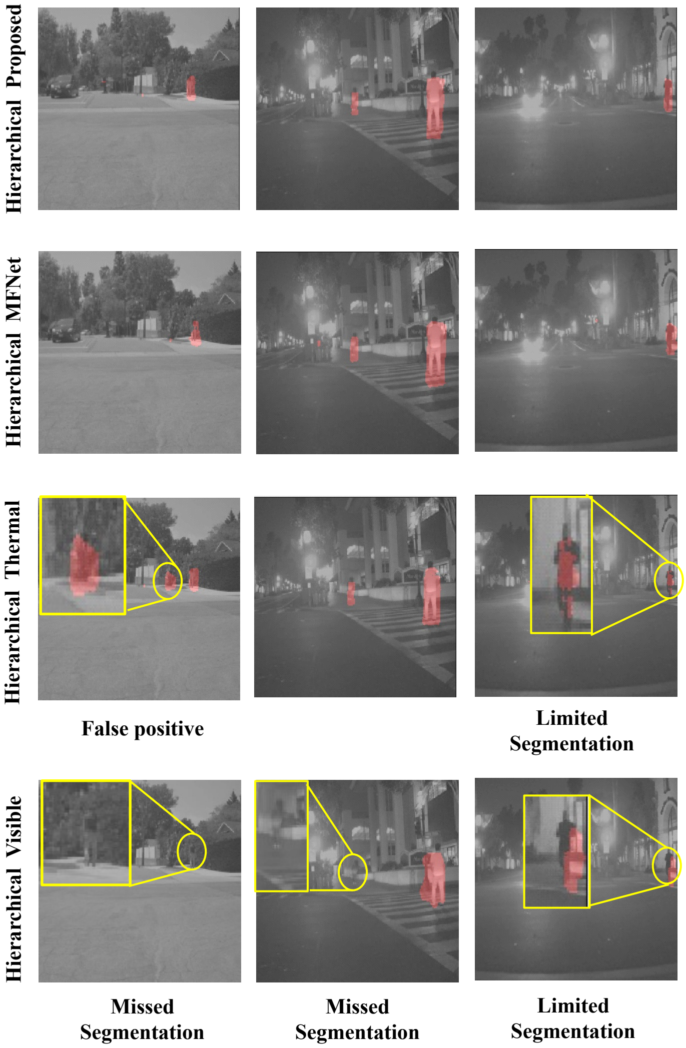
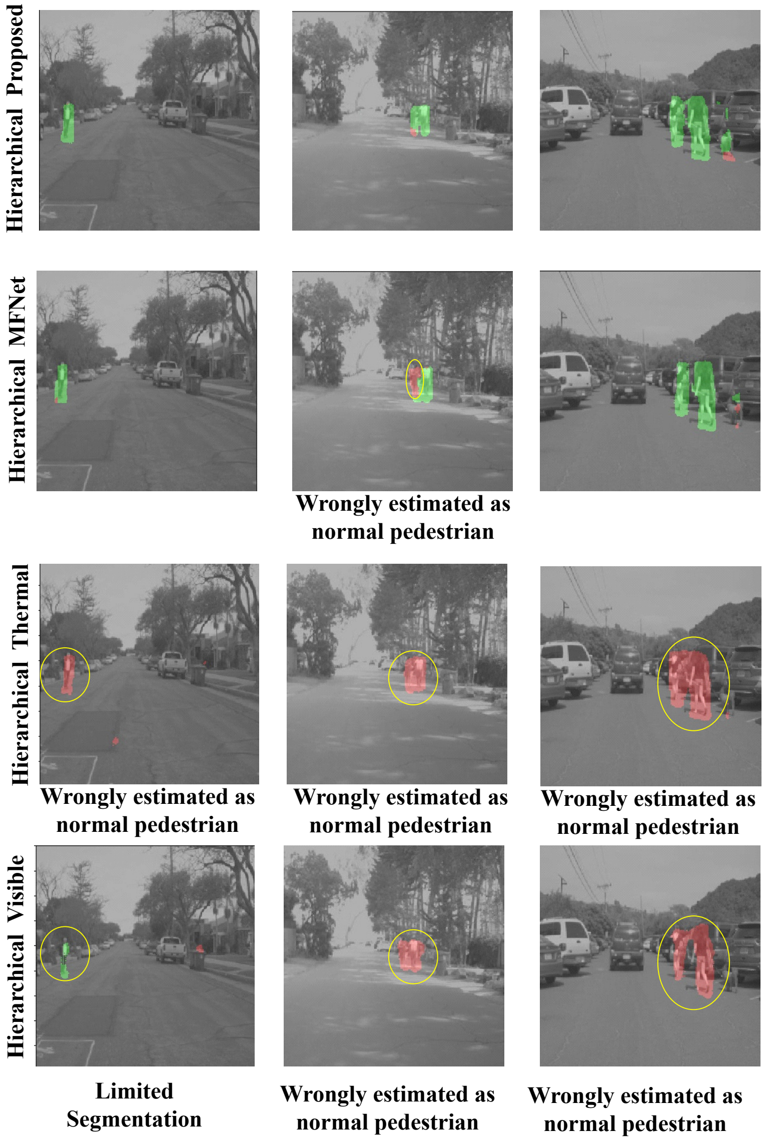


Fig. 9. Results of the different algorithms illustrating normal pedestrian behavior.

Varying Estimation Approaches: Apart from the hierarchical formulation, we also investigate two different jaywalking estimation approaches namely end-to-end semantic segmentation and hierarchical with single encoder. The results of the different approaches are tabulated in Table 3. The results show that the proposed hierarchical approach is better than the other two approaches. The reuse of the classification feature maps (hierarchical with single encoder) or the use of a naive end-to-end segmentation network doesn't improve the accuracy.



Wrongly estimated as normal pedestrian

Wrongly estimated as normal pedestrian

Wrongly estimated as normal pedestrian

Wrongly estimated as normal pedestrian

Limited Segmentation

Wrongly estimated as normal pedestrian

Wrongly estimated as normal pedestrian

Fig. 10. Results of the different algorithms illustrating abnormal or jaywalking pedestrian behavior.

5 Summary

A hierarchical deep learning framework is proposed for jaywalking estimation using thermal and visible cameras. The hierarchical framework is proposed to address the challenges in this perception task. The hierarchical framework contains two steps, an

Table 1. Comparative analysis of the proposed hierarchical framework with different hierarchical frameworks.

Proposed Algo.	Pixel Acc.%		IOU (Unit scale)		Class. Acc %	Time (ms)
	Normal Ped	Jaywalk Ped.	Normal Ped	Jaywalk Ped.		
Proposed Hier.	80.05	**82.75**	**0.71**	**0.73**	99	47
Hier. MFNet	**84.42**	68.10	0.69	0.59	96	38
Hier. Fusenet	51.40	59.08	0.46	0.52	96	35

Table 2. Comparative analysis of the sensor fusion of the proposed hierarchical framework.

Proposed Algo.	Pixel Acc.%		IOU (Unit Scale)		Class. Acc %	Time (ms)
	Normal Ped	Jaywalk Ped.	Normal Ped	Jaywalk Ped.		
Hier. Fusion	**80.05**	**82.75**	**0.71**	**0.73**	99	47
Hier. Visible	41.32	60.43	0.37	0.48	93	32
Hier. Thermal	55.18	40.90	0.47	0.35	89	31

Table 3. Comparative analysis of the proposed hierarchical framework with different estimation approaches.

Proposed Algo.	Pixel Acc.%		IOU (Unit Scale)		Time (ms)
	Normal Ped	Jaywalk Ped.	Normal Ped	Jaywalk Ped.	
Proposed Hier.	**80.05**	**82.75**	**0.71**	**0.73**	47
End-to-End Seg.	78.99	71.58	0.70	0.64	36
Hier. Single Encoder	60.04	32.23	0.50	0.29	50

initial step with a single classification network and second step with two semantic segmentation networks. In the first step, the classification network classifies the scene into a legal or illegal pedestrian crossing scenes. In the second step, scene-specific semantic segmentation networks are used to estimate the jaywalking pedestrians. The proposed framework is validated on the FLIR public dataset, and a comparative analysis with baseline algorithms is performed. The results show that the proposed hierarchical approach reports better accuracy than baseline algorithm, while reporting computational complexity. In our future work, we will investigate the framework with a much larger dataset in varying countries.

References

1. John, V., Guo, C., Mita, S., Kidono, K., Guo, C., Ishimaru, K.: Fast road scene segmentation using deep learning and scene-based models. In: ICPR. (2016)
2. John, V., Liu, Z., Guo, C., Mita, S., Kidono, K.: Real-time lane estimation using deep features and extra trees regression. In: PSIVT. (2015)
3. John, V., Nithilan, M.K., Mita, S., Tehrani, H., Konishi, M., Ishimaru, K., Oishi, T.: Sensor fusion of intensity and depth cues using the chinet for semantic segmentation of road scenes. In: IEEE Intelligent Vehicles Symposium. (2018) 585–590

4. John, V., Mita, S.: Rvnet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments. In Lee, C., Su, Z., Sugimoto, A., eds.: Image and Video Technology - 9th Pacific-Rim Symposium, PSIVT,. (2019)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41** (2009)
6. John, V., Tsuchizawa, S., Liu, Z., Mita, S.: Fusion of thermal and visible cameras for the application of pedestrian detection. Signal Image Video Process. **11** (2017) 517–524
7. Dollar, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **36** (2014) 1532–1545
8. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 1751–1760
9. Ohn-Bar, E., Trivedi, M.M.: To boost or not to boost? on the limits of boosted trees for object detection. CoRR **abs/1701.01692** (2017)
10. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. CoRR **abs/1706.08564** (2017)
11. Du, X., El-Khamy, M., Lee, J., Davis, L.S.: Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. CoRR **abs/1610.03466** (2016)
12. Carvalho, J.F., Vejdemo-Johansson, M., Pokorny, F.T., Kragic, D.: Long-term prediction of motion trajectories using path homology clusters. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019, IEEE (2019) 765–772
13. Yoo, Y., Yun, K., Yun, S., Hong, J., Jeong, H., Choi, J.: Visual path prediction in complex scenes with crowded moving objects. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 2668–2677
14. Javan Roshtkhari, M., Levine, M.D.: An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Comput. Vis. Image Underst. **117** (2013) 1436–1452
15. Rudenko, A., Palmieri, L., Arras, K.O.: Joint long-term prediction of human motion using a planning-based social force approach. In: 2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018, IEEE (2018) 1–7
16. Solaimanpour, S., Doshi, P.: A layered hmm for predicting motion of a leader in multi-robot settings. In: ICRA, IEEE (2017) 788–793
17. Bera, A., Kim, S., Manocha, D.: Realtime anomaly detection using trajectory-level crowd behavior learning. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (2016) 1289–1296
18. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps (2019)
19. Medel, J.R., Savakis, A.E.: Anomaly detection in video using predictive convolutional long short-term memory networks. CoRR **abs/1612.00390** (2016)
20. Wu, X., Zhao, W., Yuan, S.: Skeleton-Based Pedestrian Abnormal Behavior Detection with Spatio-Temporal Model in Public Places. In: Journal of Physics Conference Series. Volume 1518 of Journal of Physics Conference Series. (2020) 012018
21. Hazirbas, C., Ma, L., Domokos, C., Cremers, D. In: FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. (2017)
22. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. CoRR **abs/1510.01553** (2015)
23. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). (2017) 5108–5115