A Robust Deep Learning Approach to Pedestrian Detection via Thermal Imaging

A Technical Report submitted to the Department of Engineering Systems and Environment

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Navya Annapareddy
Spring, 2021

Technical Project Team Members
Emir Sahin
Sander Abraham

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Abstract

Pedestrian detection has been explored for application in fields such as autonomous vehicles, safety systems, and robotics. Computer vision has been frequently applied to pedestrian and bicycle detection for the purpose of augmenting detection capabilities of autonomous vehicles among other use cases. The use of thermal imaging input in computer vision models is beneficial due to its quality being independent of time of day or lighting conditions. Aside from multimodal sensor systems, Long Wave Infrared (LWIR) thermal cameras can detect pedestrians in illuminating conditions that would pose challenges for color-only image detection systems. The KAIST Multispectral Pedestrian Benchmark allows for multispectral pedestrian detection via a combination of paired LWIR and color data. A method to detect both pedestrians and cyclists via thermal images was developed using the Faster-R-CNN base network. The model was trained for 20 epochs and reached a final train accuracy of 65.16% and an F1 score of 81.34%. Subsequent adaptations to the models will implement methods to reduce overfitting and to include both color and thermal modalities, potentially creating an illumination-aware architecture if needed.

Introduction

Historically, the accuracy of color image based pedestrian detection systems have relied heavily on the lighting and weather conditions at the time of data collection, as well as camera-object distance and object-background contrast. (Konig et al., 2017) Thermal modalities such as LWIR imaging reduce this dependence. Combining both visible light and LWIR data allows for more accurate and improved pedestrian detection performance. While visible light images consist of three channels (RGB), thermal images consist of one infrared (IR) channel.

The KAIST Multispectral Pedestrian detection dataset includes 95,000 paired LWIR and visible light images in both day and night scenarios. (Konig et al., 2017) The infrared images were collected using long-wavelength IR Cameras which are more robust to visual interference, such as headlights in vehicular contexts. The ground truth annotations were manually collected and tagged by human annotators. Data is organized into single image frames of videos in 12 distinct sets. Not all frames include pedestrians and sets have variable amounts of videos, annotations, and class distributions. Of the 45,000 annotations referencing a type of pedestrian, 71.2% were single persons, 19.4% were multiple people, 7.6% were cyclists, and 1.8% were of an uncertain label class. In the implemented model, people and persons were combined into a single label class constituting more than 90% of all annotations.

Previous Work

As our work is based on using thermal inputs for biker and pedestrian detection, it is crucial to explore previous works based on thermal data recognition. Thermal images tend to have significantly lower resolution than RGB images due to their nature. This tends to cause issues while training as there isn't as much information available for the model to utilize. However, in situations where visibility is low, thermal sensors do provide increased information in comparison to RGB images. An example of such a case is indicated in Bhattarai and Martinez-Ramon's paper on utilizing thermal input for target detection for firefighting (Bhattarai & Martinez-Ramon, 2020). In their model, they've utilized 320x240 pixel input videos through a CNN based on the VGG16 neural network. This indicates that 2dCNNs do provide actionable results from thermal images - though, it should be noted that in their use case, targets are close to the camera unlike the situation present in automated vehicles where pedestrians can be over 30 meters away. While 2dCNNs can be used for pedestrian detection, it is likely that different

approaches will result in more robust models. The mentioned multi-scale convolutional networks utilize sliding windows to generate feature maps of different scales which aims to solve size discrepancy. The convolutional-deconvolutional model connects a deconvolutional layer to the base network to output a segmentation map of the whole image, letting the model determine the feature scales it will use through training - much like using an auto encoder for unsupervised learning.

Either of the models state that the thermal imaging assists in low visibility detection. As vehicles may be subject to low-visibility situations, utilizing the proposed models in creating a novel methodology aimed at detecting pedestrians and bikers will prove useful.

Dataset

There is a lack of publicly available databases for thermal imaging - especially regarding roads and pedestrians. Currently available datasets include the OSU Thermal Pedestrian Database from a pedestrian intersection in the Ohio State University campus (OHIOP) as well as the Color-Thermal database from the Ohio State University (OHIOT). Another dataset for moving thermal objects is the CSIO database (CSIO) where thermal objects are recorded from a stationary tripod that is 4ft high. This dataset provides both bikers, cars, pedestrians, and more in a moving format at a perspective closer to one that of a car. However, as the tripod is stationary, it does not provide sufficient information as a vehicle mounted data stream would.

Due to the shortcomings of OSU and CSIO databases, the KAIST database described above was used. The KAIST database's baseline benchmark has been explored and built upon using many different types of models (Hwang, Park, Kim, Choi, & Kweon, 2015). The KAIST database does have a large discrepancy between the number of pedestrians and bikers, however which is likely to cause overfitting issues that will need to be addressed in the model. The

discrepancy in these numbers do mean that additional cyclists may need to be added into the data.

Faster R-CNN is an architecture built upon state-of-the-art object detection networks to speed up the process of real-time object detection with region proposal networks. In such a model, full frames are passed in and analyzed to hypothesize various object locations found within the image. The architecture of this model utilizes a CNN for feature extraction for classification, much like the current model that we are working on. However, the output provided by the model is closer to the results one would want for an object detection methodology in autonomous vehicles as it provides object classification and detection (Ren, He, Girshick, & Sun, 2017). The thermal inputs were normalized so different visibility situations provide similar features.

Due to the additional benefit of object detection introduced by Faster R-CNN, it was the base network of our approach. Several tests were conducted on the model by using individually normalized full frames as inputs for the model. Faster R-CNN model does lack a means of implementing temporal data. This can be improved upon as detection in autonomous vehicles can make use of temporal data due to the moving state of pedestrians, bikers, and the vehicle itself.

Model Implementation and Performance

Training the model is resource intensive and thus requires the use of a high powered computer (HPC). The current iteration of the model utilizes a NVIDIA P100 GPU on the Rivanna HPC from UVA Research Computing. The training was set on a job using a chosen batch size of 1 image. The batch size is low because of the large size of the thermal input. The results for 20 epochs are described.

Following model implementation, the given predictions were subject to additional clustering to eliminate background noise and eliminate proper subsets of predictions that were completely encompassed by another prediction. In every stage of clustering, the label of the prediction box with the largest area was retained and the label of the other prediction(s) disregarded.

Metrics for F1, accuracy, precision, recall, and loss were calculated for the best model before and after prediction clustering. The clustered model achieved a final train accuracy of 65.16% and an F1 score of 81.34%. F1 (which considers class imbalance), precision, and recall metrics for the unclustered model was slightly higher than the clustered model but accuracy remained 18.29% higher for the clustered model.

The loss remained relatively constant. The average loss plateaued as epoch number increases, indicating model convergence. The training accuracy reached a significant number within small amounts of iterations and exceeds validation accuracy in every epoch.

Future Work

CNN + LSTM

CNN + LSTM models utilize feature extraction capabilities of CNNs and the temporal prediction capabilities of LSTMs. This makes CNN + LSTM models more accurate than Faster-RCNN models in situations where temporal data is crucial such as location predictions of objects in autonomous vehicles. In such a model, LSTM learns temporally global features of videos while CNN captures the spatial features found in the videos. Previous work has identified that while LSTM models in long-term intent prediction do improve baseline accuracy, there is a tradeoff with mean position error rate (Ahmed et al., 2019). Taking a time series approach will require the data to be organized by sequence rather than concatenated frames by class.

ConvLSTM

Much like the CNN + LSTM model, 3dCNN + ConvLSTM are two different networks cascaded to build a single model. They utilize a 3dCNN which captures spatial and temporally local features as well as a ConvLSTM to learn short and long-term spatial and temporal features successively (3dLSTM). Such a model would overcome the lack of temporally local features in the CNN + LSTM models. In order to build such a model, the video inputs must be adjusted so that they're of the same length to fit into the 3dCNN. It should be noted that the local temporal features learned through 3dCNN + ConvLSTM models may not increase accuracy substantially from a CNN + LSTM and will require further resources to run.

3DCNN + ConvLSTM

A traditional CNN applies a one-dimensional vector for a filter. This approach can work for some datasets. The three-dimensional CNN approach (3DCNN) allows the capture of more information by using a 3D filter.

Much like the CNN + LSTM model, 3DCNN + ConvLSTM are two different networks cascaded to build a single model. This model utilizes a 3DCNN which captures spatial and temporally local features as well as a ConvLSTM to learn short and long-term spatial and temporal features successively (3dLSTM). Such a model would overcome the lack of temporally local features in the CNN + LSTM models. In order to build such a model, the video inputs must be adjusted so that they're of the same length to fit into the 3DCNN. It should be noted that the local temporal features learned through 3DCNN + ConvLSTM models may not increase accuracy substantially from a CNN + LSTM and will require further resources to run but capturing long-term spatiotemporal features has been previously successful, with early fusion being preferred to later fusion (Akula, Ghosh, Kumar, & Sardana, 2013).

Multi Scale CNN

Finally, the Multi-Scale CNNC (MS-CNN) model can handle input images containing variable object sizes. The model utilizes several subnetworks of layers optimized for different sizes and then combines the outputs at the end.

Multispectral Detection

For multispectral pedestrian datasets, an architecture consisting of both color and thermal networks will have to be created, with the potential of weighting the networks depending on illumination or other factors and testing different fusion approaches remaining a possibility. Future work can also consist of data collection and validation on the chosen models using manual data collection from a LWIR Camera.

Conclusion

The model implemented in this paper currently displays overfitting, likely due to the large imbalance in data. Before implementing further different models, it is necessary to reduce overfitting. While there are methods utilized to reduce overfitting, data augmentation seems to be necessary. The results with overfitting do show possible promising results in case there is not overfitting as the model does seem to be able to differentiate between cyclists and pedestrians. Future work consists of developing and testing one or a combination of the current and previously explored model types to provide an optimal detection scheme in conjunction with multimodal inputs. Specifically, LSTM, 3DCNN, and fusion models will be explored heavily in the future. Methods of implementing fusion will also be explored. As discussed, multimodal pedestrian detection is also an area of interest.

References

Ahmed, S., Huda, M. N., Rajbhandari, S., Saha, C., Elshaw, M., & Kanarachos, S. (2019). Pedestrian and cyclist detection and intent estimation for autonomous vehicles: A survey. *Applied Sciences, 9*(11), 2335. doi:10.3390/app9112335

Akula, A., Ghosh, R., Kumar, S., & Sardana, H. K. (2013). Moving target detection in thermal infrared imagery using spatiotemporal information. *Journal of the Optical Society of America A, 30*(8), 1492. doi:10.1364/josaa.30.001492

Bhattarai, M., & Martinez-Ramon, M. (2020). A deep learning framework for detection of targets in thermal images to improve firefighting. *IEEE Access, 8*, 88308-88321. doi:10.1109/access.2020.2993767

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J. S., An, K., & Kweon, I. S. (2018). KAIST multi-spectral Day/Night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems, 19*(3), 934-948. doi:10.1109/tits.2018.2791533

Davis, J., & Keck, M. (2005). A two-stage template approach to person detection in thermal imagery. *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*. doi:10.1109/acvmot.2005.14

Ding, L., Wang, Y., Laganière, R., Huang, D., & Fu, S. (2020). Convolutional neural networks for MULTISPECTRAL pedestrian detection. *Signal Processing: Image Communication, 82*, 115764. doi:10.1016/j.image.2019.115764

Hwang, S., Park, J., Kim, N., Choi, Y., & Kweon, I. S. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2015.7298706

Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., & Teutsch, M. (2017). Fully convolutional region proposal networks for multispectral person detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi:10.1109/cvprw.2017.36

Ran, Y., Leykin, A., & Hammoud, R. (n.d.). Thermal-visible video fusion for moving target tracking and pedestrian motion analysis and classification. *Augmented Vision Perception in Infrared,* 349-369. doi:10.1007/978-1-84800-277-7_15

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137-1149. doi:10.1109/tpami.2016.2577031

Tewary, S., Akula, A., Ghosh, R., Kumar, S., & Sardana, H. (2014). Hybrid multi-resolution detection of moving targets in infrared imagery. *Infrared Physics & Technology, 67*, 173-183. doi:10.1016/j.infrared.2014.07.022