Multi-spectral Pedestrian Detection via Image Fusion and Deep Neural Networks

Geoff French, Graham Finlayson, Michal Mackiewicz; University of East Anglia; Norwich, UK

Abstract

The use of multi-spectral imaging has been found to improve the accuracy of deep neural network based pedestrian detection systems, particularly in challenging night time conditions in which pedestrians are more clearly visible in thermal long-wave infrared bands than in plain RGB. In this paper we use the Spectral Edge image fusion method to fuse visible RGB and IR imagery, prior to processing using a neural network based pedestrian detection system. The use of image fusion permits the use of a standard RGB object detection network without requiring the architectural modifications that are required to handle multi-spectral input. We contrast the performance of networks trained using fused images to those that use plain RGB images and networks that use a multi-spectral input.

Introduction

Accurate pedestrian detection is an essential component of autonomous driving and surveillance systems and is a specific and challenging use case of general object detection. The apperance of pedestrians in daylight varies considerably due to a differences in clothing, pose and distance from the camera. Night time conditions present a particularly challenging scenario as low light levels can render pedestrians nearly indistinguishable from the image background.

The work of Hwang *et al.*demonstrates that multi-spectral imagery consisting of RGB and long-wave infrared thermal images improves the accuracy of pedestrian detection systems [11] as the thermal images contain additional visual cues that a pedestrian detetion system can utilise. The benefits of multi-spectral imagery are particularly apparent at night time, as poor lighting conditions have far less effect in long-wave infrared bands.

Multi-spectral pedestrian detection systems developed thus far use either a tetrachromatic input that combines RGB and thermal channels, or perform feature extraction on the visible and thermal bands separately and combine features drawn from RGB and infrared bands later in the pipeline. Neural network based pedestrian detection systems utilise similar structures. Liu *et al.* [17] developed a Faster-RCNN neural network based object detection network that accepts multi-spectral (RGB and thermal) input and confirm the clear advantages of using multi-spectral imagery over RGB only.

In this paper, we explore the use of the specral edge image fusion [2] algorithm to fuse the thermal band of a multi-spectral image into the RGB band as a pre-process, prior to use as an input to an object detection neural network. The use of image fusion permits the use of a standard pre-trained RGB object detection network without requiring the architectural modifications that are required to handle multi-spectral input. While we do not present earth shattering results, we demonstrate that fused RGB images produced with spectral edge image fusion retain useful visual cues from the thermal band, helping the network achieve higher performance than can be achieved using plain RGB images.

Related Work

Object detection

Within the last several years deep neural networks have enabled significant strides forward in the field of computer vision, yielding state of the art image classification results [14, 20, 10]. These advancements spurred the development of new approaches for object detection [23] and segmentation [1] among other computer vision tasks.

Modern neural network based object detection systems utilise transfer learning, in which an image classification network (normally one trained using the ImageNet[3] dataset) is repurposed for object detection. The final layers of the classifier are removed and replaced with layers designed for object detection, after which the new network is fine-tuned. Normally this is done by using a lower learning rate for the pre-existing layers during training.

Girshick *et al.* [9] proposed R-CNN, a technique that applies a neural network classifier and bounding box regressor in a sliding window fashion across an image. Selective search [24] is used to generate initial region proposals for a given image. These regions are warped to 227 × 227 pixels (the fixed input size of the neural network) after which the network predicts if the region contains an object – and if so which class of object – and generates an improved bounding box. This approach is computationally expensive as the neural network must process approximately 2,000 region images to process a single input image.

Fast R-CNN [8] improved on R-CNN by applying the convolutional layers of a pre-trained network to generate high level features for a complete input image. Regions are extracted from the high level feature image rather than the low level raw pixel image as in R-CNN. As a consequence, the computationally expensive convolutional layers of the network are evaluated once for the complete image rather than once per proposal. The final R-CNN layers predict object presence, class and bounding box refinements from the feature images.

Faster R-CNN [23] replaces selective search region proposal system with a region proposal network (RPN). The RPN passes the high level features from the ImageNet classifier network body to an RPN head that generates object proposals. The proposals predict the presence of an object or lack thereof within regularly space anchor boxes of differing scales and aspect ratios and predict offsets and scale factors to refine the anchor boxes to better match those of the objects. These proposed bounding boxes are filtered using non-maxmimum suppression and passed to Fast R-CNN as before to classify their content and predict bounding box refinements.

The YOLO [21, 22] (You Only Look Once) and SSD [18, 7] (Single Shot Detector) models are one stage detectors. Their structure – that of SSD in particular – is similar to that of the RPN component of Faster R-CNN. The elimination of the Fast R-CNN refinement stage simplifies the overall algorithm and provides signficant speed ups, at the expense of accuracy in comparison to Faster R-CNN. Lin *et al.*discovered that the inferior accuracy of one-stage detectors was due to the class imbalance between foreground and background samples and proposed to adress this focal loss [16], resulting on one-stage detectors whose accuracy can surpass that of Faster R-CNN.

Image fusion

Image fusion is a process in which the channels in a multichannel image (the channels can differ in resolution in some situations) are combined, reducing the number of channels. Examples include colour to greyscale conversion and converting multi-band satellite imagery to RGB images. An effective image fusion algorithm attempts to preserve as much salient visual information from the original high-channel image during the fusion process.

Spectral edge image fusion [2] is an approach for transforming an N-channel image to an M-channel image where $N > M$ while preserving contrast and gradient information such that the lower dimensional image can preserve RGB appearance while incorporating gradient and contrast information from the other channels. Frequent use cases involve enhancing an RGB image with information from other channels within a hyperspectral image, such that details from invisible parts of the spectrum can be perceived by the viewer, while retaining much of the original colour of the RGB image so that it can be easily interpreted.

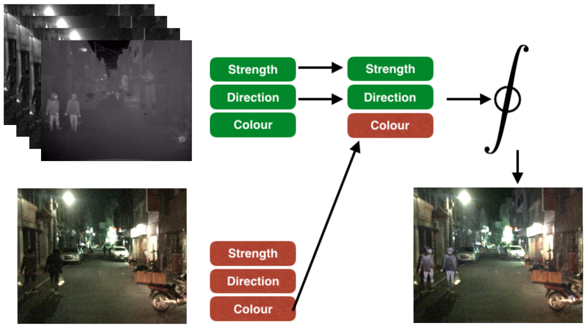

Figure 1. *An example of Spectral Edge Image Fusion being used to fuse RGB+NIR to make a new RGB image*

In Figure 1 top left we show (coded as grey scale images) a 4 channel imge. The RGB channels are drawn from the picture shown bottom left. The front grey-scale shows the Near Infrared information. In Spectral Edge Fusion we represent the 4 channel image and the RGB original (which is called the 'guide'). Without presenting the detail, the colour gradient was shown to have what was called the 'spectral edge' structure. Specifically, each colour edge can be decomposed into 3 parts: colour, direction and strength, see Figure 1. It was proposed a good fusion would take the colour edge information from the guide (in the hope of preserving the colour look of the original) but the strength and direction of the edges were taken from the multispectral (in this case 4 channel) image.

This process results in 3 gradient images (one for the red-, green- and blue- channels). And, these need to be reintegrated to form the final image. The final result is shown bottom right of Figure 1. The bright appearance of the pedestrians that can be seen in the NIR image has been incorporated into the final image.

Multi-spectral pedestrian detection

The KAIST multi-spectral pedestrian detection dataset was introduced by Hwang *et al.* [11], along with a baseline detector based on aggregated channel features [4]. The results obtained from their baseline approach clearly illustrated the effectiveness of multi-spectral imagery for pedestrian detection.

Liu *et al.* [17] implemented a multi-spectral pedestrian detector based on a Faster R-CNN deep neural network object detector. They explored the effect of fusing RGB and thermal imagery at four different points within the network. Like the original Faster R-CNN model, they used the VGG-16 [20] ImageNet classifier as a backbone. Their early and halfway fusion models fused the RGB and thermal paths within the convolutional stages of the network using Network-in-Network (NIN) [15] layers to blend the RGB and thermal features. The early and halfway models fused after the first and fourth block of convolutional layers respectively. Their late fusion model fused after the last feature generation layers, while their score fusion model averaged score predictions. Their halfway fusion model yielded the best performance.

Zhang *et al.* [25] report that the R-CNN classification and bounding box refinement stage of a Faster R-CNN network hampered pedestrian detection performance in comparison to the underlying region proposal network (RPN) whose predictions it refines. They proposed a model that combines a RPN with a boosted forest based classifier instead of an additional neural network head. König [13] adopted this approach and proposed a multi-spectral RPN for pedestrian detection.

Method

Network structure

Following prior approaches for pedestrian detection [17, 25, 13], we employ transfer learning, utilising the VGG-16 [20] image classification network as the feature extraction backbone.

Similar to the work of König *et al.* [13] we base our pedestrian detection system on region proposal networks (RPN); the first stage of a Faster-RCNN object detector. We note the similarity between our RPN-based model and that of a single-shot detector (SSD) [18].

Faster R-CNN based object detection systems typically use three anchor box aspect ratios; 1:2, 1:1 and 2:1 (height:width) and four anchor box scales (32×32, 64×64, 128×128, and 256×256 pixels). Following Liu *et al.* [17] we discarded the 1:2 aspect ratio given the typical aspect ratio of pedestrian bounding boxes, but retained the four different scales.

Protocol

We followed the protocol of Liu *et al.* [17], ignoring ground truth pedestrian instances that are marked as occluded or containing multiple pedestrians, or whose height is < 50 pixels. Ignored instances are prevented from contributing to the neural network's loss function during training, thus the network is not trained to predict them as being either pedestrians or background. They also do not count during evaluation. Only frames that contain at least one postive example of a pedestrian are used for training and evaluation.

Training

At a resolution of 640 × 480 and 640 × 512 there will be 9,600 and 10,240 anchor boxes per image for the Caltech and KAIST datasets respectively (the anchor boxes are centred on grid points across the image at a stride of 16 pixels). The RPN of a Faster-RCNN network is typically trained by randomly selecting a balanced subset of (typically 256) anchor boxes and training only those boxes for each image. In contrast, we train using all of the anchor boxes in each image in a fully convolutional manner, similar to the semantic segmentation approaches of Long *et al*. [19], using 1 image per mini-batch. In early experiments we found that this accelerated training. The mean number of pedestrians in the KAIST dataset is around 2-3 per image. This results in a large class imbalance between foreground and background samples (anchor boxes that contain a pedestrian vs those that do not) that hampers training. We used α-balancing to counteract this; we scaled the RPN classification loss for anchor boxes depending to their ground truth class (foreground or background) by a weighting factor inversly proportional to the foreground to background ratio of the complete dataset.

RPN anchor boxes were considered to be contain a pedestrian if they had an intersection-over-union (IoU) overlap of $>$ 0.5 [18] and negative if they had an overlap of $<$ 0.25. Anchors whose overlap was between these thresholds are considered neutral and do not contribute to the training loss. We used these thresholds in contrast to the more stringent thresholds of 0.7 and 0.3 used in most Faster R-CNN implementations as we found that relaxing them improved detection rates.

Inference

The detections generated by the RPN are filtered using non-maximal suppression, discarding the lowest scoring detection (classification score predicting the presence of a pedestrian) of any pair of detections whose overlap IoU is $>$ 0.25. This is in contrast to the a threshold of 0.7 used to filter RPN proposals within a Faster R-CNN network and a threshold of 0.3 used to filter predictions from the Fast R-CNN network head.

Pre-train on Caltech dataset

Our networks were first trained using the Caltech pedestrian dataset [5] using the protocol described above. They were trained for 10 epochs, using the Adam [12] optimisation algorithm using a mini-batch size of 1 and a learning rate of 1×10^{-3} for the randomly initialised parameters belonging to the RPN network head and 1×10^{-4} for the pre-trained VGG-16 parameters.

Fine-tune on KAIST dataset

The weights of a network trained on the Caltech dataset as described above were used as a starting point for training using the KAIST dataset. Given that the RPN network head is already trained by this point, we used a learning rate of 1×10^{-4} to fine-tune all network parameters, effectively treating the entire network as pre-trained.

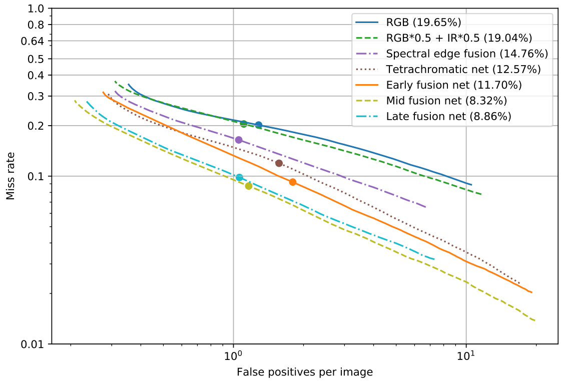

Figure 2. *Performance evaluation of various image fusion approaches. Curves represent the trade-off between miss rate and false positive rate by varying the detection threshold. The values in parentheses in the legend are the log-average-miss-rate.*

Multi-spectral fusion

We compare several methods of fusing RGB and infrared bands within the network. For each of these experiments, a network pre-trained on the Caltech dataset (as stated above) was fine tuned using the following inputs. Our early, mid and late fusion architectures are heavily inspired by those of Liu *et al*. [17].

RGB: A vanilla network uses only the RGB band as input. It was trained using the images and ground truths in the KAIST training set and evaluated on images in the test set.

$RGB*0.5+IR*0.5$: The infrared image is converted to RGB via channel replication and blended equally with the RGB image and passed to a vanilla RGB network.

Spectral edge fusion: The RGB and infrared images are first combined using spectral edge fusion. These fused RGB images are used as input for a vanilla RGB network.

Tetrachromatic net: The first convolutional layer from the VGG-16 backbone is modified in order to accept a 4-channel input rather than a 3-channel input. The RGB part of the convolutional kernel is copied from the original network, while the weights connected to infrared channel are randomly initialised with a Gaussian distribution with the same mean and variance as that of the weights connected to the RGB channels. This partially pre-trained layer was trained using a learning rate of 3×10^{-4}.

Early, mid and late fusion nets: The network is bifurcated, with layers after the split remaining unchanged and layers prior to the split being duplicated in order to make two paths; one for the RGB band and one for the infrared. The two paths are joined at the split point by concatenting the per-pixel features from each incoming branch and fused using a network-in-network [15] layer. The RGB images are passed as is as input to the RGB path while the infrared images are converted to RGB by channel replication and passed to the infrared path. The early, mid and late fusion nets are split after the `pool1` (first max-pooling layer after the 1st block of 3 convolutional layers), `pool4` (max-pooling layer after the 4th block) or `conv5_3` layers of the VGG-16 backbone respectively.

Fusion / input	Log average miss rate
RGB	19.65%
$RGB \times 0.5 + IR \times 0.5$	19.04%
Spectral edge fusion	14.76%
Tetrachromatic network	12.57%
Early fusion network	11.70%
Mid fusion network	8.32%
Late fusion network	8.86%

Table 1. Log average miss rates

Evaluation

Following the evaluation protocols established by Dollar *et al*. [5, 6], we summarise the effect of a variety of approaches to RGB-infrared image fusion on our pedestrian detector in Figure 2 and Table 1. Our log average miss rates are computed by averaging the miss rates corresponding to false positive rates logarithmically spaced between the values of 0.35 and 6.6 false positives per image; the range over which data is available for all approaches. This is in contrast to the range of 0.01 to 1.0 that is typically used.

The curves in Figure 2 fall roughly into three clusters: RGB and 50% blend; spectral edge fusion, tetrachromatic network and early fusion; and mid and late fusion. Fusing the infrared band into the RGB band with a 50% blend offers very little additional performance in contrast to plain RGB. Using spectral edge fusion yeilds a considerable improvement, reducing the miss rate from 19.04% to 14.76%. When using an un-modified network that accepts an RGB input, spectral edge fusion gives by far the best results.

Further improvements can be obtained by modifying the network structure to utilise the infrared band as well as RGB. The miss rate is reduced to 12.57% by using a tetrachromatic net that uses inputs with 4-channels per pixel. A slight additional improvement with a miss rate of 11.70% can be obtained by fusing after the first convolutional block. The best performance is obtained by performing fusion later in the network.

We hypothesize that the superior performance of mid and late fusion networks is in part due to the displacement between the RGB and infrared bands that is present in much of the KAIST dataset, an example of which can be seen in Figure 3. Feature representations generated by later layers in the network contain higher level semantic information at lower resolution. Feature image pixels will also draw information from a larger receptive field on the original image. This could simplify the process of counteracting the effects of displacement between the RGB and infrared channels.

Conclusions and future work

We have evaluated the effect of using spectral edge image fusion to convert multi-spectral images to visible RGB images prior to processing with a neural network based pedestrian detector. Spectral edge fusion improves the performance of a standard RGB pedestrian detection network when measured using the KAIST benchmark, although less so than modifying the network to use multi-spectral imagery directly.

Our use of spectral edge fusion brought to our attention the displacement between the RGB and infrared bands that can be seen throughout much of the KAIST dataset. We believe that this offers an explanation for the superior performance of late fusion architectures. We would like to test this hypothesis in the future by evaluating the fusion approaches discussed previously using a dataset in which this displacement is not present.

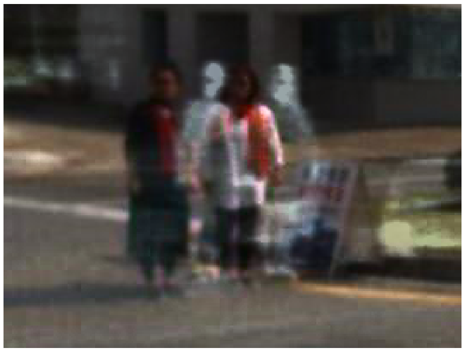

Figure 3. *Example of the displacement between RGB and infrared bands that can be seen in the KAIST dataset, in this case illustrated using spectral edge fusion. The horizontal disparity between the RGB image of the two pedestrians and the infrared band can be seen in the form of the light coloured ghosting of the pedestrians offset to right of their RGB image.*

We would like to explore the use of more modern techniques, such a residual network [10] based classification backbone and the use of focal loss [16] rather than α-balancing as it seems well suited to this problem.

The use of standard RGB networks opens interesting avenues for further research. The majority of computer vision training data consists of plain RGB images, so using additional bands – such as thermal – requires new datasets with corresponding ground truths. The possibility of using spectral edge fusion to fuse non-visible bands into an RGB image while retaining the ability to use standard RGB data for training is a tantalising one.

Acknowledgments

This work was funded by Spectral Edge Ltd.

References

[1] Chen, Liang-Chieh and Papandreou, George and Kokkinos, Iasonas and Murphy, Kevin and Yuille, Alan L, Semantic image segmentation with deep convolutional nets and fully connected CRFs, ICLR 2015.

[2] Connah, D., Drew, M. S., Finlayson, G. D., Spectral edge: gradient-preserving spectral mapping for image fusion, Journal of the Optical Society of America 2015.

[3] Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L., ImageNet: A Large-Scale Hierarchical Image Database, CVPR 2009.

[4] P. Dollár, R. Appel, S. Belongie, and P. Perona., Fast feature pyramids for object detection, PAMI 2014.

[5] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona Pedestrian Detection: A Benchmark, CVPR 2009.

[6] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona Pedestrian Detection: An Evaluation of the State of the Art, PAMI 2012.

[7] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, DSSD: Deconvolutional single shot detector, arXiv:1701.06659, 2017.

[8] Ross Girshick, Fast R-CNN, ICCV 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich

feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014.

[10] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, Deep residual learning for image recognition, CVPR 2016.

[11] Soonmin Hwang and Jaesik Park and Namil Kim and Yukyung Choi and In So Kweon, Multispectral Pedestrian Detection: Benchmark Dataset and Baselines, CVPR 2015.

[12] Kingma, Diederik and Ba, Jimmy, Adam: A method for stochastic optimization, ICLR 2015.

[13] König D, Adam M, Jarvers C, Layher G, Neumann H, Teutsch M, Fully convolutional region proposal networks for multispectral person detection, CVPR 2017.

[14] A. Krizhevsky and Sutskever, I. and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012.

[15] Lin, M., Chen, Q. and Yan, S., Network in network, arXiv preprint arXiv:1312.4400, 2013.

[16] Lin, T. Y., Goyal, P., Girshick, R., He, K. and Dollr, P., Focal loss for dense object detection, arXiv preprint arXiv:1708.02002, 2017.

[17] Jingjing Liu, Shaoting Zhang, Shu Wang and Dimitris Metaxas, Multispectral Deep Neural Networks for Pedestrian Detection, BMVC 2016.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, SSD: Single shot multibox detector, ECCV, 2016.

[19] Long, Jonathan and Shelhamer, Evan and Darrell, Trevor, Fully convolutional networks for semantic segmentation, CVPR 2015.

[20] Simonyan, K. and Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, CVPR, 2016.

[22] J. Redmon and A. Farhadi, YOLO9000: Better, faster, stronger, CVPR, 2017.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015.

[24] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective search for object recognition, IJCV 2013.

[25] Liliang Zhang, Liang Lin, Xiaodan Liang and Kaiming He, Is Faster R-CNN Doing Well for Pedestrian Detection? ECCV 2016.

Author Biography

Graham Finlayson is a Professor of Computer Science at the University of East Anglia. Graham trained in computer science first at the University of Strathclyde and then for his masters and doctoral degrees at Simon Fraser University. Prior to joining UEA, Graham was a lecturer at the University of York and then a founder and Reader at the Colour and Imaging Institute at the University of Derby.

Geoff French received his BSc and MSc in computing from the University of East Anglia in 2001 and 2013. He is now a PhD student at the University of East Anglia studying deep learning for computer vision.

Michal Mackiewicz is a lecturer in Computer Science at the University of East Anglia. Michal studied for his doctoral degree at the University of East Anglia. His research interests focus on computer vision and colour imaging.