Multistage Deep Neural Network Framework for People Detection and Localization Using Fusion of Visible and Thermal Images

Bushra Khalid(✉), Muhammad Usman Akram, and Asad Mansoor Khan

National University of Science and Technology, Islamabad, Pakistan
bkhalid.ce16ceme@gmail.com, usmakram@gmail.com, asad.m.khan12@gmail.com

Abstract. In Computer vision object detection and classification are active fields of research. Applications of object detection and classification include a diverse range of fields such as surveillance, autonomous cars and robotic vision. Many intelligent systems are built by researchers to achieve the accuracy of human perception but could not quite achieve it yet. Convolutional Neural Networks (CNN) and Deep Learning architectures are used to achieve human like perception for object detection and scene identification. We are proposing a novel method by combining previously used techniques. We are proposing a model which takes multi-spectral images, fuses them together, drops the useless images and then provides semantic segmentation for each object (person) present in the image. In our proposed methodology we are using CNN for fusion of Visible and thermal images and Deep Learning architectures for classification and localization. Fusion of visible and thermal images is carried out to combine informative features of both images into one image. For fusion we are using Encoder-decoder architecture. Fused image is then fed into Resnet-152 architecture for classification of images. Images obtained from Resnet-152 are then fed into Mask-RCNN for localization of persons. Mask-RCNN uses Resnet-101 architecture for localization of objects. From the results it can be clearly seen that Fused model for object localization outperforms the Visible model and gives promising results for person detection for surveillance purposes. Our proposed model gives the Miss Rate of 5.25% which is much better than the previous state of the art method applied on KAIST dataset.

Keywords: Object detection · Object localization · Mask-RCNN

1 Introduction

In the present age of technology, security of individuals is one of the most important concerns. Offices, schools, hospitals and organizations are provided with complete security measures to a avoid any kind of security breach. These security measures include security personnels, Close Circuit Television cameras (CCTV)

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 138–147, 2020.
https://doi.org/10.1007/978-3-030-51935-3_15

and surveillance. CCTV cameras can be used in a number of other applications as well, such as offense detection, public safety, crime prevention, quick emergency response, management and for the reduction of fear of crime in public [1]. Surveillance cameras can also be used for monitoring traffic flow and to keep an eye on the staff in regards to the complaints received [2]. While monitoring a vicinity there are many types of surveillance cameras to cater different weather and lighting conditions. During the day light visible cameras come in handy for clear observation while utilizing the bright light of the sun. Thermal or Long Wave InfraRed (LWIR) cameras on the other hand becomes useful when there is a bad lighting condition, storm, fog or dark scene. Visible cameras compliment with a source of information where thermal cameras identify the presence of persons, animals and weapons. Heat signatures along with visible information becomes more informative during surveillance. Intruders and robbers can be identified during night time and even if they have hidden themselves behind any solid object, thermal cameras nullify the effect of occlusion.

Section 2 contains a brief overview of related work. Section 3 gives an overview of datasets used in experimentation. Section 4 states the details of the proposed methodology. Experiments and results are shown in Sect. 5 and Conclusion with future work is stated in Sect. 6.

2 Related Work

In feature level classification objects are classified, but not detected and localized. Convolution neural networks are capable of object detection and excelled in this field over the past few years. Object detection caught the eyes of researchers in 2014 after the introduction of the basic object detection technique introduced by Girshick [3]. After the comparison of results of the Regional Convolutional Neural Network (RCNN) and Histogram Of Gradient (HOG) based classifiers, it was clear that CNN outperformed other techniques with a clear margin. This literature review is divided into two parts: first we discuss about visible and thermal data sets and how they are utilized in various applications and then object detection and localization is discussed.

2.1 Visible and Thermal Image Data

Visible images consisting of three channels Red Green Blue (RGB) when paired with thermal or infrared images are known as multi-spectral images. The infrared image can be obtained by infrared camera which represents the thermal radiations within the range of 0 to 255. The difference between visible and infrared image is that visible images are highly sensitive to lighting conditions and they contain fixed pattern information [4]. While thermal images do not contain any kind of texture information and thermal images displays the heat map of any object whose temperature is above absolute zero. Thermal images are categorized into five types. Near Infrared Image (NIR), Short Wave Infrared (SWIR), Mid Wave Infrared (MWIR), Long Wave Infrared (LWIR) and Far Infrared (FIR).

NIR ranges from 0.75 to 1.4, SWIR ranges from 1.4 to 3, MWIR 3 to 8, LWIR 8 to 15, FIR 15 to 1000 µm wavelength.

2.2 Object Detection and Localization

Object detection is the center of attention in computer vision and it is one of the basic building blocks. A number of researchers have worked on object detection more specifically human detection to avoid collision events while driving autonomous cars or during the movement of robots. Till late 90 s detection of objects was totally based on visual images and visual scenarios. But now this problem is shifting towards multi-spectral identification, i.e. object detection using images of different spectrums or modalities. Regional CNN (RCNN) [5] takes image as an input, identifies the region of interest and provides bounding box output and also a classification score. The mechanism behind RCNN is that it creates a lot of bounding boxes or square boxes of different sizes of the image by the selective search method [SSM]. The Selective search method propagates square and rectangular windows of different sizes over the image and select the boxes in which adjacent pixels show a pattern of potential object. To address the issues of RCNN, Girshick proposed an improved model in 2015 which is known a Fast-RCNN [3]. The first problem which was solved in Fast-RCNN was that all the separate models were combined as one model. RCNN calculated separate feature vectors for each proposal, but Fast-RCNN combined all the feature vectors of an image in one vector. That feature vector is then used by CNN for classification and bounding box regression. These two solutions turned out to be effective in terms of speed. Region Of Interest (ROI) pooling is the process in which an image is converted from $\mathrm{h} \times \mathrm{w}$ matrix to $\mathrm{H} \times \mathrm{W}$ matrix by applying max pooling, but the image is dealt in the form of small windows and each window is max-pooled to get the output image window and the whole image likewise. The solution to the above two problems was proposed in Faster-RCNN [6]. Faster-RCNN is a combined model which provides a classification score and bounding boxes with the help of RPN and Fast-RCNN. CNN is pre-trained on imagenet for classification. For the generation of Region proposals, Region proposal Network RPN is a fine-tuned end-to-end. Proposals having IOU overlap greater than 0.7 are positive while the ones having IOU overlap less than 0.3 are negative. Fast-RCNN is trained using RPNs from Region Proposal Network. RPN is then trained using Fast-RCNN. They have some shared convolution layers. After this Fast-RCNN is fine-tuned, but only the unique layers of it are fine-tuned.

3 Dataset

The data set we are using for training CNNs in this paper is KAIST multi-spectral data set. KAIST multi-spectral data is acquired by mounting the visible and thermal cameras on a car and an additional beam splitter is used to align the LWIR and RGB image data [7]. Different images are taken in different lighting condition to observe the effect of light on the scene and on the object detection. Figure 1 shows randomly chosen images from KAIST multi-spectral data set and it ground truth bounding box representation. KAIST consists of 95000 Visual-Thermal image pairs. The size of every image is 640 × 512 and they are aligned geometrically. The data set is divided into 60 to 40 ratio for training and testing.

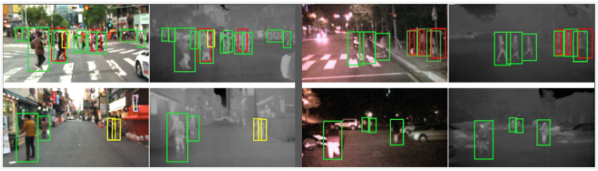

Fig. 1. Ground truth representation of KAIST dataset

Apart from KAIST multi-spectral data, we are using another local dataset for testing purpose which consisted of visible thermal pairs of 20 images. This dataset contained some images which showed a clear difference of a person's presence in the scene. In the visible image a person might be hidden behind the bush or might be wearing a camouflage which is invisible to the naked eye. In such cases Thermal images come in handy. Figure 2 shows some random images from the local dataset.

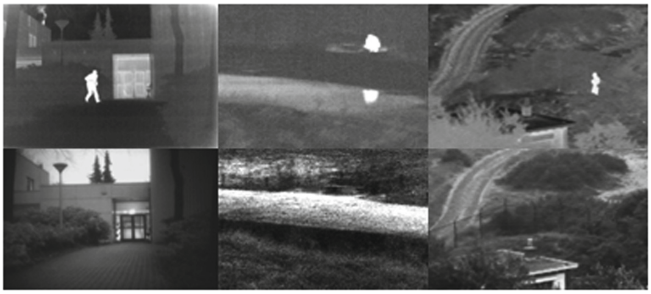

Fig. 2. Visible and thermal image pairs from local dataset

4 Proposed Methodology

Our proposed methodology consists of three main modules. The first module is a fusion module which consists of two encoders for visible and thermal image encoding. Both modules take the input images and after encoding the feature maps carry out the process of fusion. Once the features are fused the fused feature

Table 1. Dataset

Parameters	Training	Testing
Division	60%	40%
Images	50.2 k	45.1 k
Day	33399	29179
Night	16788	15962
Pedestrian	45.1 k	44.7 k

vector is transferred to the decoder block which decodes it back to image content. From this module a final fused image is obtained which is then transferred to the next module called ResNet block. This block takes the fused image and classifies it as a person class or a no person class. Images classified as no person class are discarded while the ones having person class are transferred to the third and final module of image localization. These three models are explained in detail below. Figure 3 shows the flow diagram of our proposed technique which clearly shows all the three blocks of the model (Table 1).

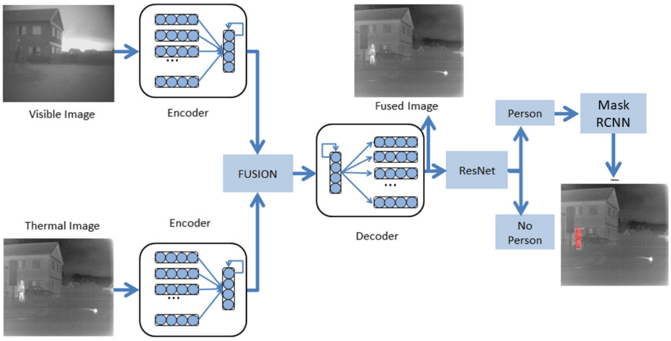

Fig. 3. Proposed image fusion and localization framework

4.1 Fusion and Classification of Visible and Thermal Images

The network architecture used in encoder part of our proposed model is a Siamese architecture [11]. Siamese architecture is a neural network, which is different from usual neural networks. It does not classify the input fed into it, rather it is designed to take decisions by comparing the similarities and dissimilarities of multiple inputs. Siamese architecture consists of two sub networks whose feature maps are compared and final encoded output is decided on the basis of that comparison. In the siamese network architecture two sister networks with same properties take same or different inputs and these inputs pass through convolution layers for feature extraction. Once the features are extracted from both inputs, then they are passed through a contrastive loss function which actually calculates the similarities and differences between the inputs.

For the classification of images we are using ResNet-152 [8] architecture. A fused image obtained from encoder-decoder architecture is passed through the convolution layers of ResNet architecture for its classification. If the network classifies the image as a person then there is a person or persons present in the frame. Images classified as persons are passed through the localization network for localization of all the objects such as persons present in the image.

4.2 Localization of Persons

In this section we will discuss the model we are using for object detection and localization. Mask-RCNN [9] is an extension of Faster-RCNN. Apart from object detection and classification Mask-RCNN also produces instance segmentation masks. Mask-RCNN introduces a new branch for the instance segmentation which outputs the masks of detected objects. The instance segmentation branch is a fully connected network, which provides pixel to pixel segmentation on the basis of the Region of Interests. A fully connected network is a network in which every node is connected to every other node. As detecting and overlaying a mask over the object is much more complex than just drawing the bounding boxes around them, Mask-RCNN introduces a new layer called ROI-Align layer in place if ROI-pooling layer. Figure 4 represents the architecture of Mask-RCNN.

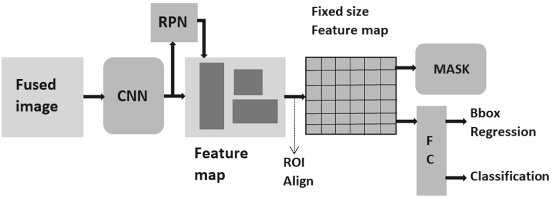

Fig. 4. Mask-RCNN architecture

RPN takes a sliding window and propagates it throughout the image. 2k anchors are formed in the result. RPN usually takes about 10 ms to skim an image, but in case of Mask-RCNN it might take longer than that as the input image to Mask-RCNN is relatively of bigger size hence it needs more anchors. Anchors are divided into two classes based on the IOU overlap. The foreground class is highly likely to contain an object while background class does not. Foreground anchors are then refined to get an exact bounding box of the object present inside the anchor. The problem of bounding box is solved, but the problem of classification still remains. To solve this problem ROI is used, the ROI takes the foreground object and classifies it into its actual class. Than ROI pooling is done to resize the image so that it can be sent into the classifier network.

The segmentation branch takes in the ROI results and gives the mask for objects in the output.

5 Experiments and Results

Our experimentation is divided into two parts. The first part explains the experimental setup used for fusion model and shows the results obtained from the fusion of thermal and visible images. The second part describes the experimental setup of detection and localization model and displays few results from localization part.

5.1 Detection and Localization Results

The image having persons present in it is then transferred into the Mask-RCNN model for localization. For training KAIST dataset provides bounding box annotations, but in case of Mask-RCNN we need binary masks for training. For this purpose, we used the VIA VGG annotator tool. Using this interface we created image masks for our data set. VIA tool saves json model for training and validation data sets separately. There is no need for extensive training by a huge bulk of data because Mask-RCNN model can utilize pre-trained weights of imagenet and MS-COCO. ResNet architectures need high computation power in the training process. NVIDIA GPU GTX-1080 Ti with 64 GB RAM is used in the

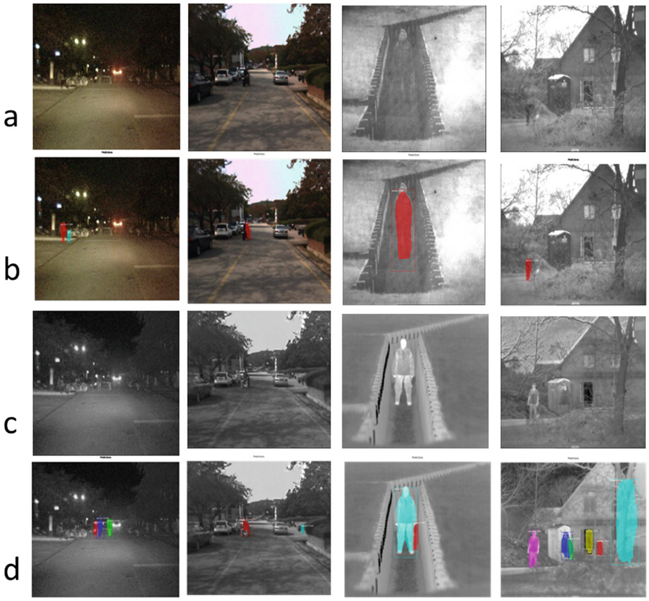

Fig. 5. KAIST: (a) Visible images (b) Localization of a (c) Fused images (d) Localization of c

training process. Training time taken by Mask-RCNN over 50.2 k images is 4 days, but once it is trained the testing timing Mask-RCNN provides promising results for localization of objects. Figure 5 shows the results of Mask-RCNN over KAIST multi-spectral dataset and local dataset. Images in the first two columns are from KAIST dataset while in images in the third and fourth column are from local dataset. Row a consists of Visible images (first column contains an image of night time while the second column contains images of day time and their respective detections are shown in row b. It can be seen that there are a lot of False detections in the visible image present in the first column of the row. While its respective fused image and its detection scan be seen in c and d. First column of row d shows that there is only one false detection in the fused image. From fourth column it is noted that the person present inside the box is hidden while in the fused image it can be seen partially. Detection results show that in the visible image the person inside the box remains undetected when passed through Mask-RCNN model while both the persons are detected in the fused image. There are some false detections in the fused image as the model is trained over KAIST and these test images are randomly chosen. For the calculation of miss rate, state of the art methodology and formula is used which can found in detail in paper proposed by Konig et al. [13]. Table 2 shows the graph of comparison of our proposed model with previous techniques. From this graph it can be seen that our proposed model outperformed previous methodologies in terms of Miss Rate. Miss Rate gives us the measure of accuracy by which an object was detected pixel by pixel.

Table 2. Comparison of MR% with previous studies

S. no	Author	Year	Technique	Data set	Miss rate
1	Wagner et al. [10]	2016	CNN	KAIST	43.80%
2	Liu et al. [12]	2016	Faster-RCNN	KAIST	37%
3	Konig et al. [13]	2017	Fusion RPN + BDT	KAIST	29.89%
4	Xu et al. [14]	2017	CMT-CNN	KAIST	10.69%
5	Ours	2019	Encoder-Decoder+ Mask-RCNN	KAIST	5.25%

6 Conclusion and Future Work

The purpose of this paper is to utilize modern technology and computer vision models for efficient surveillance and the provision of foolproof security in organizations, schools, hospitals and military zones. For this purpose, we are utilizing visible and thermal cameras to obtain images of the premises as well as heat maps of suspicious intruders. These images are then fused together to get a combined more informative output for detection of a doubtful presence. We are using Encoder-decoder CNN architecture for fusion of visible and thermal images and Resnet architecture for object detection and localization. Localization of object or persons is done using Mask-RCNN model which not only localizes the object, but also provides a mask for localized object. KAIST multi-spectral data are used for the training of CNNs and local dataset is also used in the testing process. When the results of visible detections are compared with results of Fused detections, it is clearly observed Fused model outperforms the detection and localization process by giving accurate masks for KAIST and comparatively better masks for local dataset than the visible model. The learning time of model can be improved by minimizing the layers of ResNet architecture.

References

1. Ratcliffe, J.H.: Video Surveillance of Public Places: Problem Oriented Guides for Police, Response Guides Series, No. 4, pp. 195–197. Center for Problem Oriented Policing, Washington DC (2006)
2. National Rail CCTV Steering Group and Others: National rail and underground closed circuit television (CCTV) guidance document: final version, vol. 30 (2010)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**, 142–158 (2016)
4. Gade, R., Moeslund, T.B.: Thermal cameras and applications: a survey. Mach. Vis. Appli. **25**, 245–262 (2014). https://doi.org/10.1007/s00138-013-0570-5
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
6. Shaoqing, R., Kaiming, H., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
7. Hwang, S., Park, J., Kim, N., Choi, Y., So, K.: Multispectral pedestrian detection: benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1037–1045 (2015)
8. 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)
9. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
10. Wagner, J., Fisher, V., Herman, M., Behenke, S.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2016)
11. Ram, P.K., Sai, S., Venkatesh, B.: DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
12. Liu, J., Zhang, S., Wang, S., Metaxas, D.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)

13. Konig, D., Michael, A., Christian, J., Georg, L., Heiko, N., Michael, T.: Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
14. Xu, D., Ouyang, W., Ricci, E., Wand, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)