ZULFAQAR Journal of
Defence Science, Engineering & Technology

Journal homepage: https://zulfaqar.upnm.edu.my/

HUMAN DETECTION FOR THERMAL AND VISIBLE IMAGES

Mazlinda Ibrahim[a,*], Hoo Yann Seong[a], Siti Zulaikha Ngah Demon[a] ,Suzaimah Ramli[b] , Syed Nasir Alsagoff Syed Zakaria[b]

[a] *Center for Defence Foundation Studies, Universiti Pertahanan Nasional Malaysia, Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia*
[b] *Faculty of Defence Science and Technology, Universiti Pertahanan Nasional Malaysia, Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia*

*Corresponding author: *mazlinda@upnm.edu.my*

ARTICLE INFO

Article history:
Received
29-11-2019
Received in revised
18-06-2020
Accepted
18-02-2021
Available online
30-06-2021

Keywords:
aggregate of channel features, histogram of oriented gradient, human detection, integral image

e-ISSN: 2773-5281
Type: Article

ABSTRACT

Human detection and localization is one of the importance aspects in computer vision. It has broad applications in surveillance, robotic, driver assistance system, and for the military applications. The task is difficult because it depends on various conditions such as illumination, distance, human pose and weather condition. This study aimed to investigate human detection methods for thermal and visible images. We have explored three methods which are histogram of oriented gradient, integral image and aggregate of channel features. Our result showed that histogram of oriented gradient outperformed the other two using the tested images. However, the method is only applicable when the human is on the standing or upright position and limited to a certain distance between the scene and the camera position.

Introduction

Thermal cameras are being used widely in the surveillance system due to the ability of the camera to operate without additional light source. However, the thermal camera must be equipped with the ability to distinguish humans and animals from other warm objects for the surveillance and military applications. Besides, human detection in thermal images is a very tedious and difficult task due to the low image resolution, noise, and low texture information as mentioned in Riaz et al. (2003). Generally, the human detection process in images requires three stages as shown in Fig.1. First, the pre-processing stage where the images are filtered with median filtering to remove noise. Second, the images are segmented into foreground and background for feature extraction. Third is the decision making stage where the objects in the images are classified as human or not.

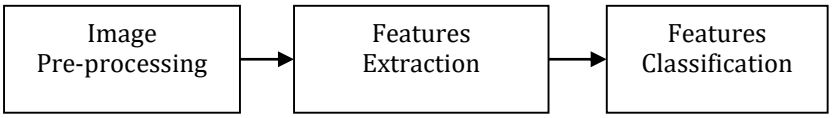


Fig. 1: Flow chart for object detection

In this paper, we investigated three methods for the human detection in the visible and thermal images. There are histogram of oriented gradients (HOG) Dalal & Triggs (2015), integral image (VJ) Viola & Jones (2001), and aggregate of channel features (ACF) Dollar et al. (2014). Experiment and results showed that HOG is the best method in terms of correct detection where the method managed to detect human in half of the data set followed by VJ and ACF.

The literature on the methods in human detection are mostly developed for the visible images. According to Jeon et al. (2015) the method of human detection can be divided into two: based on background subtraction and without background subtraction. However, the method can also be categorized based on the static and moving images. It is because the task of human detection in moving images can employ the fact that human is moving objects. Based on Budzan (2015), the methods can also be categorized into two: single image processing (static) and sequence of images (video). The single image processing for human detection is more difficult compared to the sequence of images due to the limited information about the environment involved. For instance, Surasak et al. (2018) used HOG for human detection in the video. The authors in Budzan (2015) also highlighted that human legs are the part of human body which are mostly wrong in the identification and detection process.

Zhang et al. (2007) investigated methods derived from visible images for human detection. They get comparable results for the detection of human in infrared and visible images. They also showed that the two images share many features such as the human silhouettes. In addition, Negied et al. (2015) stated that when dealing with thermal images, we do not need special processing techniques than the visible images. The effectiveness of the methods which use machine learning algorithm in feature classification and extraction are highly depend on the set of the images used in the training. Thus, there is no universal method for human detection. It should be tailor accordingly to the specific application as mentioned in Budzan (2015).

The authors in Wang et al. (2010) highlighted that methods based features as such Haar features (Viola & Jones, (2001)) and HOG (Dalal & Triggs, (2005)) may not be suitable for human detection in thermal images for surveillance applications because the size of the objects of interest are relatively small, weather conditions and texture information. However, based on the results obtained, HOG works well for the human detection in thermal and visible images. In Zhao et al. (2009), the authors proposed human detection via segmentation system by fusing video and thermal images. The model used background subtractions and only effective for video sequences. In Turic et al. (2008), the authors used mean shift method to segment the long distance images for their human detection model. However, the model is applied to the visible images and the success rate using their test images is 88%. Ramli et al. (2014) focused on the human motion detection in thermal and visible images sequence using background subtraction algorithm. The application system proposed in Ramli et al. (2014) managed to detect moving objects but there is no information on the variations of human posed tested.

In Teutsch et al. (2014), human detection for thermal images was performed using two stages. In the first stage, hot spot detection was utilized followed by the normalization of the images to coarse scale. Meanwhile, in the second stage, machine learning algorithms was used to distinguish between human and non-human for the classification part. However, for images with many hot spots, the method produced too many false positives. Negied et al. (2015), provided a comprehensive review on human or pedestrians detection in thermal images. The authors mentioned several issues related to human detection in visible

images such as various styles of clothing, different possible articulations, occluded accessories and pedestrian. Thus, encourage researchers and industries to migrate from visible to thermal images.

In 2018, Zaihidee et al. (2018) proposed a method based on Stationary Wavelet Transform for outdoor human detection in the fusion of the thermal and visible images. But, the proposed method did not produce the location the human in the image. For this reason, the method is applicable only for image enhancement for the fusion images. In 2019, Ivašić-Kos et al. Considered human detection in thermal videos and images where the videos are recorded at night but on different weather conditions. They used the so-called convolutional neural network for the detection process. They mentioned that the performance of their model on the thermal imagery can improve significantly with additional training on the thermal data set.

Qu et al. (2019), focused on the indoor human detection and localization using multiple thermal sensors. The sensor is installed to collect the data of human motion. Then they used interpolation method and Gaussian filter to smooth the images. Next, the threshold method is applied to remove non-human and background images for the detection processes. The outline of this article is as follows: In the next section we review the three methods: HOG, VJ, and ACF methods. We show in ‘‘Results and Discussion’’ section, some numerical tests including comparisons for our data sets. Finally, we present our conclusion and future work in the ‘‘Conclusion’’ section.

Methods

Histogram of oriented gradients (HOG)

Histogram of oriented gradient employs edge direction where the local objects or features within an image are described by the distribution of the intensity gradients. The method divides the image into smaller number of block. Within block, there are cells. For each cells, a histogram of the gradient directions is calculated for all pixels within the cell. Then, the local histogram is normalized by calculating the mean intensity across the block. This normalization step improved the effect of illumination and shadowing. The gradient computation D_x using 1D central first order finite difference is given by

$$D_x = [-1 \quad 0 \quad 1] \tag{1}$$

for the horizontal direction and similarly for the vertical direction. For an image I, the gradient is given by

$$\nabla I = [I_x \quad I_y] \tag{2}$$

and the magnitude of the gradient is $|\nabla I| = \sqrt{(I_x)^2 + \left(I_y\right)^2}$ The orientation of the gradient is calculated using

$$\theta = tan^{-1}\frac{I_x}{I_y} \tag{3}$$

Next, histogram channel is evenly spread over 0 to 180º or 0 to 360º. Dalal and Trigs, (2005) used unsigned gradients in nine histogram channel for the best implementation in the method. They also highlighted that 8 × 8-pixel cell, 2 × 2 cells per block, and 50% block overlap are performed best in their experiments. The oriented histogram for all blocks is concatenated into one big descriptor vector. The descriptor vector is compare with the classifier for the object detection. In Dalal and Trigs (2005), they used liner support vector machines (SVM) to trained the classifier. SVM is a supervised learning model that analyzed data and recognize pattern. The method is based on the human silhouette which mainly focuses on head, shoulder, and feet. Thus, it is only applicable for human detection in the standing position. Another issues with HOG is mentioned in Yıldız (2012), where the author explained that when the human is not exactly at the center of the image with respect to the left-right line, the performance of the method drops dramatically in the

context of visible images. In this project, we are using three types of images: thermal, visible and fusion of the thermal and visible images for human detection using HOG.

Integral image (VJ)

The Viola Jones object detection framework is a widely used method for real time object detection. The method consists of four main stages: Haar features selection, integral image, Adaboost training and cascade classifiers. It is limited to full view frontal upright faces. Means, in order to detect the face, the person should point towards the camera. It should not be tilted to any side. The main idea behind this model is to rescale the sub window consists of the face detector across given input images. The face detector is constructed using the integral image. The integral image is obtained by making each pixel equal to the entire sum of all pixels above and to the left of the concerned pixels. The integral image at location *x, y* is given by

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x', y') \tag{4}$$

where $i(x,y)$ is the pixel value of the original image and $ii(x,y)$ is the corresponding image integral value.

A set of image features similar to the Haar basic functions is used to construct the detector using the integral images. The authors in Viola & Jones (2001), found that the detector with base resolution of 24×24 pixels gives best results. After obtaining the feature values at base resolution, they used a modified version of AdaBoost algorithm for the classification. AdaBoost is a machine learning boosting algorithm develop by Freund &Schapire, (1999). Next, the detector is used to scan the input images several time and at different scale. The algorithm discard non faces window at each scale. This approach is so-called as cascaded classifier. The sub window which are not discarded by the initial classifier are further processed by more complex classifier than the last one. The model by Viola & Jones (2001) is based on supervised machine learning and mainly for the frontal face detection. Thus, the model is at disadvantages when the human face is occluded or not apparent.

Aggregate of channel features (ACF)

Aggregated channels features (ACF) framework is one of the supreme methods for object detection. Dollar et al. (2014) provided comprehensive description for the pedestrian detection in their work. Let say we are given any image, I . The image I is smoothed with [1 2 1]/4 (mean filter) to remove noise. Others filter such as Gaussian and median filter can also be used. Then, 10 channels are computed from the smooth image. There are 6 channels for histogram of oriented gradients, one channel for normalized gradient magnitude and three channels for LUV color. For each channels, the image is partition into 4x4 blocks. Then, the 16 pixels in each blocks are summed which producing a lower resolution channels. The lower resolution channels are called as the aggregated channels and the channels are smoothed again using the mean filtering techniques. The channels are computed for four scale: 1, 0.5, 0.25 and 0.125 using multiscale image representation techniques. Next, using the four scaled channels, multiresolution image features for 7 of 8 are constructed using approximation by extrapolation which producing the so-called fast features pyramids for object detection in Dollar et al. (2014).

After post smoothing, the channels in every scale are arranged into features vector to produce pixel look-up table. This process is called vectorization. For the features classification, Dollar et al. (2014) used AdaBoost. AdaBoost is an adaptive boosting to combine multiple weak classifier into a strong classifier where the author in Dollar et al. (2014) used two level trees. The final The final strong classifier $H(\cdot)$ is defined as

$$H(a) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(a)\right) \tag{5}$$

where α_t is a weight coefficient, $h_t(\cdot)$ a weak learner and T the number of weak classifiers.

The decision trees are learned over the pixels in order to separate human from background using two databases. The database used for the training and combining the features for pedestrian detection are INRIA and Caltech databases. The AdaBoost classifier in ACF framework by Dollar et al. (2014) combined 2048 depth trees over 5120 candidate features which are basically the channel pixel lookup table in each 128x64 window. The windows are formed from the 4x4 blocks. The process for training the classifier are performed using multiple rounds of bootstrapping. The exact parameters setting and training framework for ACF human detection are available online. See Dollar et al. (2014) for more details on the ACF method.

Results and Discussion

The experiment was run using Matlab R2018b on Windows 10 and functions in Matlab Image Processing Toolbox were used. There are 10 images from thermal, visual and fusion of thermal and visual images. Details of the data set are shown in Table 1 where the size of images are fixed at 240X320 pixels.

Table 1: Data set, number of human in the image and types of images. Fusion refers to images obtainted from thermal and visible sources.

No	No. of human	Source
1	1	Fusion
2	4	Fusion
3	4	Fusion
4	2	Fusion
5	1	Thermal
6	1	Visible
7	1	Visible
8	1	Visible
9	4	Visible
10	4	Visible

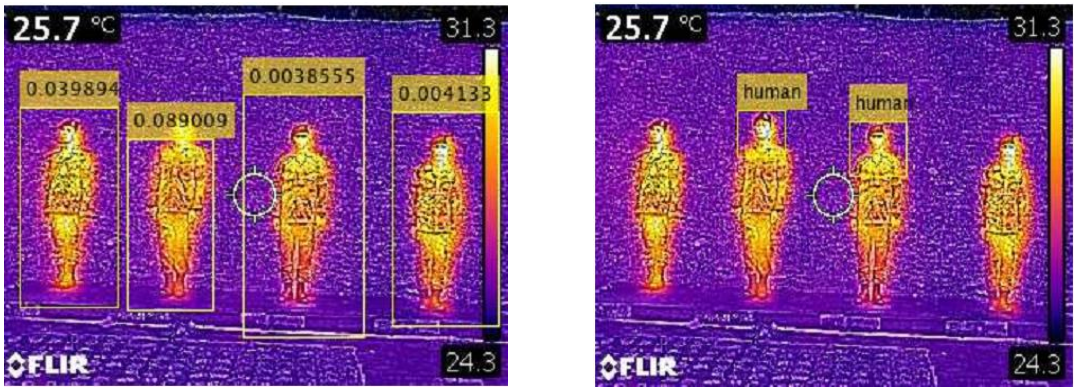
The effectiveness of the methods tested depend on the set of the images used to train the model. With an increasing amount of testing images, the probability of human detection will be increased. The HOG and VJ methods are using machine learning algorithms to train the images in human detection. These methods are highly depending on the train images and the probability of the detection are increased with increasing amount of testing images. Results obtained after processing all the images are presented in Table 2.

Table 2: Detection Results

Method	Correct Detection	Partially Correct Detection	Incorrect Detection
HOG	50%	0%	50%
VJ	40%	20%	40%
ACF	20%	0%	80%

Correct detection refers to all the human in the images are detected and neither false positives nor false negatives. If the model detects some but not all human in the images with multiple persons, results are presented as partially correct detection. Meanwhile, incorrect detection refers to either false positive or false negatives. Incorrect detection also refers to the case where no human detected in the image although they are presented in the original image.

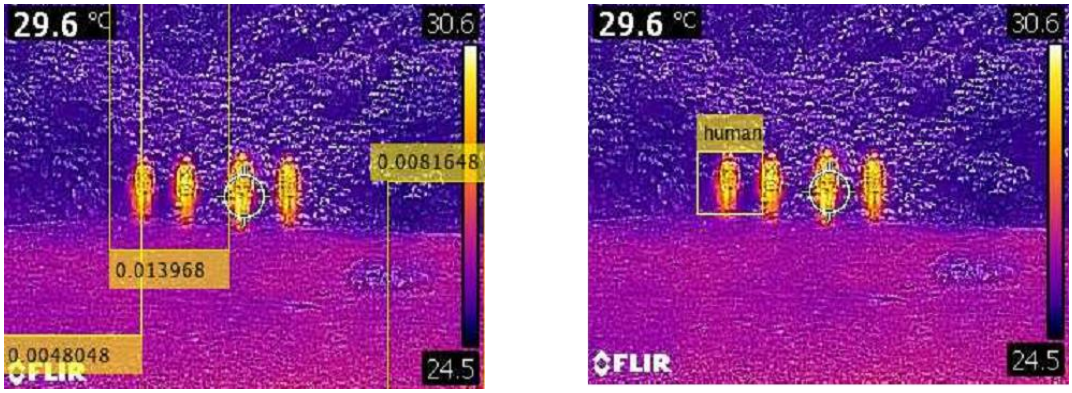
From Table 2, HOG method outperformed the others two methods. The method manages to detect human in half of the data set. Meanwhile, VJ method comes second with 40% correct detection, 20% partially correct detection and only 40% incorrect detection. ACF method only manages to detect human in 2 data set. The main difference for the performance of HOG and VJ are for data set 2 and 3.



(a) **(b)**

Fig. 2: Detection of data set 2 using a) HOG and b) VJ methods

Results for data set 2 using HOG and VJ methods. Left figure is the result using HOG method where the method manages to detect all human in the image (correct detection). Right figure is the result using VJ method where the method manages to detect some of the human in the image (partially correct detection). Meanwhile ACF method did not manage to detect any human present in the image (incorrect detection). The detection results using HOG and VJ methods for data set 2 are shown in Fig. 2. HOG manages to detect all human in the image but VJ only manages to detect two out of four humans in the image. For data set 2, size of the human silhouette in the image is approximately 110x30 pixels. Thus, the ratio of the height of the human with the image size is 0.5.



(a) **(b)**

Fig. 3: Detection of data set 3 using a) HOG and b) VJ methods

Fig. 3: Results for data set 3 using HOG and VJ methods. Left figure is the result using HOG method where the method fails to detect all human in the image (incorrect detection). Right figure is the result using VJ method where the method manages to detect some of the human in the image (partially correct detection). Meanwhile ACF method did not manage to detect any human present in the image (incorrect detection). In Fig. 3, HOG method performs incorrect detection for the four human in the image. At the same time, VJ managed to detect one out of the four human which fall into partially correct detection. HOG fails to detect human in the image because the size of the human is relatively small compared to the image size. The ratio between the human's height in the image is 0.2. Thus, the distance between the scene and the camera is far in comparison with the data set 2 in Fig. 2. We can observe that when the distance between the human and the camera is considerably close, HOG method is at advantages.

For data set 3 (Fig. 3), VJ manages to detect some of the human in the image which categorize under partially correct detection. In data set 2 and 3, the faces are not apparent. Hence, we concluded that the VJ method manages to detect the human in data set 3 by coincident. It is due to the shape of the human which is look like a face. Recall that the VJ method is using the features of the face for human detection. Meanwhile ACF method did not manage to detect any of the human for data set 2 and 3. The results for the ACF method are not shown here because of no object detection in the images.

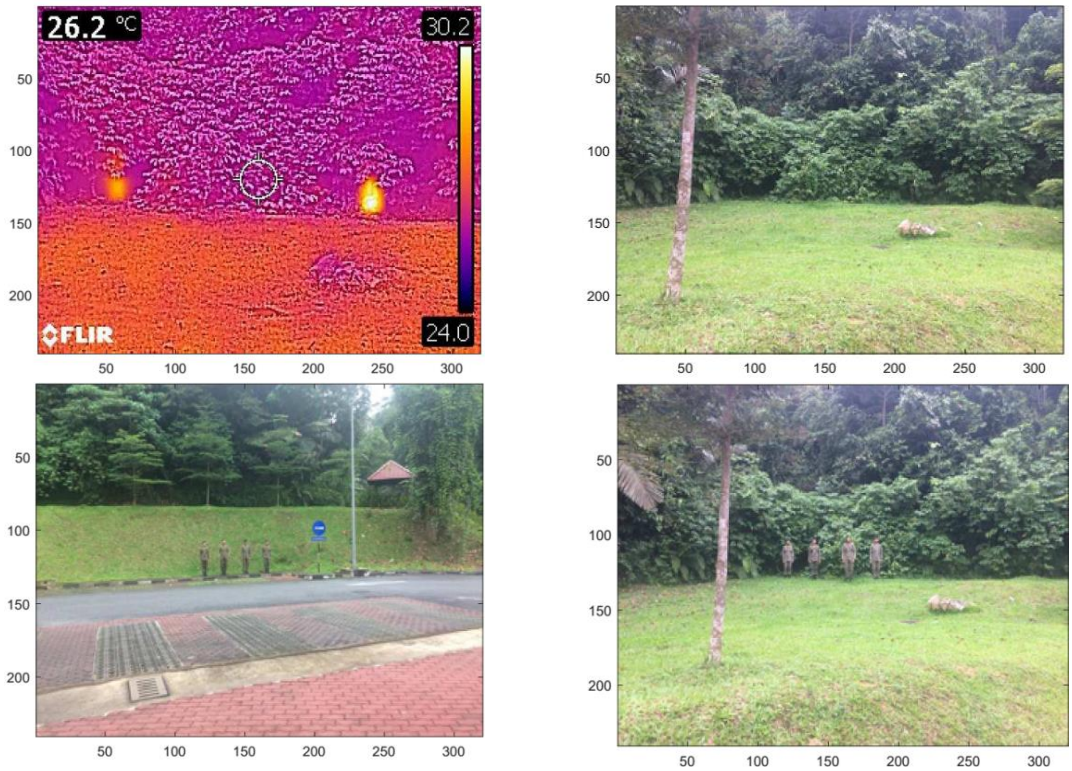


Fig. 4: Top left (data set 4) , top right(data set 8), bottom left (data set 9), bottom right (data set 10). These all the cases where none of the methods manage to perform human detection.

Data set 4 and 8 depict squatting human (not in the upright position). Meanwhile human size in data set 9 and 10 are considerably small due to the distance between the human and camera. These are the cases where all three methods fail to perform human detection due to the size of the human in the images. The ratio of the human's height in date set 9 and 10 are approximately 0.1. Meanwhile for data set 4 and 5, the humans are not clearly visible which increase the difficulty level of any human detection methods which are based on the features of the human.

Conclusion

We have investigated three methods for human detection in visible, thermal and fusion of the visible and thermal images. The HOG method outperformed VJ and ACF methods based on the tested images. The HOG method uses full human figure and the VJ method uses human faces only for the detection process. Nevertheless, the method has some limitations. First, the size of the human in the images must be large enough (around half of the image size). If the human size is considerably small (around 0.2 from the image size), the method will fail. Second, the human in the images must be in the standing position. Otherwise, the method will be at disadvantages for human detection as highlighted before in Yıldız (2012).

Future works include collecting more data set for the improvement of the training algorithm for the fusion human detection and investigation of the methods under bad weather conditions. After we manage to improve the existing human detection methods in the fused images, we are going to investigate the performance of the methods under bad weather conditions and testing the effectiveness of the current military camouflage suit to shield human body heat. In addition, the method can also be used for automatic enemy detecting defense robot like the one proposed in Renuka et al. (2018).

Acknowledgement

The research and writing of this work was fully carried out under UPNM short term grants UPNM/2016/ GPJP/3/SG/3 and UPNM/2016/GPJP/3/SG/4.

References

Budzan, S. (2015). Human Detection in Thermal Images Using Low-level Features. *Measurement Automation Monitoring*, Vol. 61.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.* Vol. 1, pp. 886–893.

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 36, No. 8, pp. 1532-1545.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, Vol. 14, No. 771-780, pp. 1612.

Ivašić-Kos, M., Krišto, M., & Pobar, M. (2019). Human detection in thermal imaging using YOLO. *2019 5th International Conference on Computer and Technology Applications,* pp. 20-24.

Jeon, E., Choi, J. S., Lee, J., Shin, K., Kim, Y., Le, T., & Park, K. (2015). Human detection based on the generation of a background image by using a far-infrared light camera. *Sensors*, Vol. 15, No. 3, pp. 6763-6788.

Negied, N. K., Hemayed, E. E., & Fayek, M. B. (2015). Pedestrians' detection in thermal bands–Critical survey. *Journal of Electrical Systems and Information Technology*, Vol. 2, No. 2, pp. 141-148.

Qu, D., Yang, B., & Gu, N. (2019). Indoor multiple human targets localization and tracking using thermopile sensor. *Infrared Physics & Technology*, Vol. 97, pp. 349-359.

Ramli, S., Wahab, S. N. A., Baharon, N. A., & Zainudin, N. M. (2014). Comparison of Human Motion Detection Between Thermal and Ordinary Images. *Journal of Image and Graphics*, Vol. 2, No. 2.

Renuka, B., Sivaranjani, B., Lakshmi, A.M., & Muthukumaran, D.N. (2018). Automatic Enemy Detecting Defense Robot by using Face Detection Technique. *Asian Journal of Applied Science and Technology*, Vol. 2, No. 2, pp. 495-501.

Riaz, I., Piao, J., & Shin, H. (2013). Human detection by using centrist features for thermal images. IADIS International Journal on Computer Science and Information Systems, Vol. 8, No. 2, pp. 1-11.

Surasak, T., Takahiro, I., Cheng, C.H., Wang, C.E., & Sheng, P.Y. (2018, May). Histogram of oriented gradients for human detection in video. 2*018 5th International Conference on Business and Industrial Research (ICBIR),* pp. 172-176.

Teutsch, M., Muller, T., Huber, M., & Beyerer, J. (2014). Low resolution person detection with a moving thermal infrared camera by hot spot classification. *IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 209-216.

Turic, H., Papic, V., & Dujmic, H. (2008). A procedure for detection of humans from long distance images. *2008 50th International Symposium ELMAR,* Vol. 1, pp. 109-112.

Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001.* Vol. 1, pp. I-I.

Wang, W., Zhang, J., & Shen, C. (2010). Improved human detection and classification in thermal images. In *2010 IEEE International Conference on Image Processing.* pp. 2313-2316.

Yıldız, C. (2012). An implementation on histogram of oriented gradients for human detection. *Bilkent University*.

Zaihidee, E.M., Ghazali, K. H., Ren, J., & Salleh, M. Z. (2018). A hybrid thermal-visible fusion for outdoor human detection*. Journal of Telecommunication, Electronic, and Computer Engineering (IJTEC),* Vol. 10, No. 1-), pp. 79-83.

Zhang, L., Wu, B., & Nevatia, R. (2007). Pedestrian detection in infrared images based on local shape features. In *2007 IEEE Conference on Computer Vision and Pattern Recognition.* pp. 1-8.

Zhao, J., & Cheung, S. C. (2009). Human segmentation by fusing visible-light and thermal imaginary. *12th International Conference on Computer Vision Workshops, ICCV Workshops.* pp. 1185-1192.