Implementation Challenge and Analysis of Thermal Image Degradation on R-CNN Face Detection

Nikola Latinović, Tijana Vuković, Ranko Petrović, Miloš Pavlović, Marko Kadijević, Ilija Popadić, and Mladen Veinović

***Abstract* — Face detection systems with color cameras were rapidly evolving and have been well researched. In environments with good visibility they can reach excellent accuracy. But changes in illumination conditions can result in performance degradation, which is the one of the major limitations in visible light face detection systems. The solution to this problem could be in using thermal infrared cameras, since their operation doesn't depend on illumination. Recent studies have shown that deep learning methods can achieve an impressive performance on object detection tasks, and face detection in particular. The goal of this paper is to find an effective way to take advantages from thermal infrared spectra and provide an analysis of various image degradation influence on thermal face detection performance in a system based on R-CNN with special accent on implementation on a hardware platform for video signal processing that institute Vlatacom has developed, called vVSP.**

***Keywords* — face detection, image degradation, R-CNN, thermal images, Video Signal Processing, GPU.**

Paper received May 22, 2020; revised July 18, 2020; accepted July 31, 2020. Date of publication December 25, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Branimir Reljin.

This paper is revised and expanded version of the paper presented at the 27th Telecommunications Forum TELFOR 2019 [33].

Nikola Latinović is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: nikola.latinovic@vlatacom.com).

Tijana Vuković is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: tijanavukovic1996@gmail.com).

Ranko Petrović is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: ranko.petrovic@vlatacom.com).

Miloš Pavlović is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: milos.pavlovic@vlatacom.com).

Marko Kadijević is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: marko.kadijevic@vlatacom.com).

Dr IlijaPopadić, is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: ilija.popadic@vlatacom.com).

Prof. dr Mladen Veinovic is with the Singidunum University, Danijelova 32, 11000 Belgrade (e-mail: mveinovic@singidunum.ac.rs).

I. Introduction

In many security, surveillance or access control applications face detection algorithms are applied. In top level secure objects such as important government buildings, for access control, face detection and recognition systems have been implemented in order to verify user's identity [1], [2]. These systems are also used in security checkpoints at airports and in authentication processes for secure banking. If lighting conditions are disturbed, such as in cases of fog or heavy rain, color camera's performances are degraded and cannot provide sufficient quality in surveillance systems with a high level of security. Since a thermal camera's performance does not depend on lighting conditions, it can be used in order to improve the performances of mentioned systems [3], [4], [5], [6]. The basic principle of work of infrared sensors is that they collect emitted heat energy to form a thermogram using infrared radiation. As thermal cameras are invariant to lighting conditions, they can provide an effective view even in low-light conditions and total darkness, and in that way can be used in face detection applications.

Face detection, in general, can be considered as a special kind of operation which results in detecting objects in computer vision tasks. Moreover, deep learning methods are constantly evolving, especially in computer vision applications, while particularly the convolutional neural networks (CNN) have achieved remarkable success in terms of accuracy.

While popularity of deep learning in computer vision grows, a vast number of researchers in this area are targeting to explore deep learning methods to solve face detection tasks.

In the process to solve face detection challenges, some successful deep learning techniques have emerged for generic object detection tasks. One very useful and very successful framework for generic object detection is region-based CNN method (R-CNN). A large number of advances in solving face detection tasks usually follow this line of research with the spread of R-CNN and its enhanced variants.

The goal of this paper is to find an effective way to take advantages from thermal IR spectra and analyze the influence of various image degradations on thermal face detection performance in systems based on deep learning methods. This paper proposes an environment for testing

performance in a variety of conditions consisting of databases with images of different degrees of degradation (flipping, rotation, noise) and implementation of thermal image face detection system based on deep learning methods. Also, the implementation possibilities on a Vlatacom's Video Signal Processing Platform (vVSP) for field usage will be considered. vVSP is a hardware device capable of capturing and processing a video signal from various types of cameras, so in this case an IR thermal camera is used. Since we are in the initial phase of research in face detection field in thermal imaging, this paper presents a potential direction of algorithm application on vVSP platform. Details on this are presented in the discussion part of this paper.

This paper is organized as follows. Section II describes face detection algorithm and R-CNN face detection system architecture. Section III describes a facial image database used for training R-CNN. Section IV gives an overview of a hardware platform considered for implementation. Section V illustrates the transfer learning process and training of R-CNN. In section VI statistical comparison of the tested system on different test sets and a discussion are presented. Section VII presents a conclusion, with the indication of future work in this research area.

II. System Description

A. Face detection algorithm

To detect face regions in the image a R-CNN detector [7] is used in this work. R-CNN has three main parts in the process of detection. In the first part it extracts 2000 category-independent region proposals and feeds them to the second part, which is Convolutional Neural Network (CNN). Output of CNN is a fixed-size feature vector which is used as an input for Support Vector Machine (SVM) classifier [8]. CNN input size is a fixed 32x32 image and output is a vector which consists of 64 features. SVM uses those features to classify each region; a result of that classification is either a face class or background (non-face class). Regions overlapped with other regions with the degree of overlapping bigger than a given threshold, measured by the Intersection over Union (IoU) score metrics, are removed.

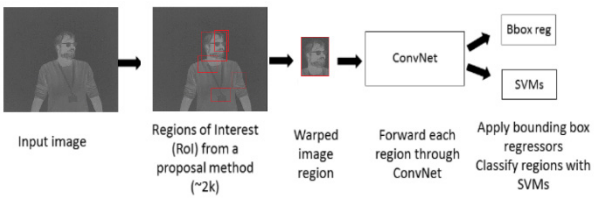

Fig. 1. Block diagram of proposed face detection system R-CNN.

The most important part of R-CNN is CNN for feature extraction. The role of the CNN is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction. The first layer of CNN is Image Input Layer, which defines the dimensions of network input. A next layer is a Medium Layer and it consists of multiple Convolutional, ReLU and Max Pooling layers. Convolutional layer consists of multiple filters (kernels) whose parameters are learned during a training process. Kernel shifts multiple times over the image and each time performs a matrix multiplication operation with a portion of the image it is hovering and result is a feature map. The purpose of ReLU layer is to introduce non-linearity in CNN. Without added non-linearity, classification would be linear and hence CNN performance would be limited. Pooling layers section would reduce the number of parameters when the image is too large. This process, also called subsampling or down-sampling reduces the dimensionality of each feature map, but retains important information. Max pooling slides the kernel across the feature map, takes the largest element from the part of the map it is hovering and passes it to the next layer.

In the Final layer there are a Fully Connected layer and a Softmax Loss layer. The Fully Connected layer flattens a matrix input into vector and feeds it to the next layer. It represents features that are later used for classification. In this CNN, they are succeeded by Softmax layer. The Softmax layer is used to calculate categorical probability distribution.

Before the training, Convolutional layer weights are first set to small random values and then updated in training process which uses Stochastic Gradient Descent with Momentum (SGDM). By computing gradient of the loss function and take small steps in the opposite direction of gradient, loss will gradually decrease until it converges to some local minima. Weight increments (Δw) are defined as:

$$\Delta w = -\eta \nabla_w E(w) = -\eta \frac{\partial E}{\partial w_j} = f\left(\eta, d^{(k)} w^T x^{(k)}\right) \quad (1)$$

where η is a learning rate, and $E(w)$ is a cost function. The learning rate defines the speed of the training process, so that large values of the learning rate can cause oscillations, while small values cause a slow training process.

With Stochastic Gradient Descent the exact derivate of the loss function is not computed, but estimated on a small batch, which means it is not always going in an optimal direction. Momentum is defined as a moving average of gradients used to accelerate gradients in the right direction and avoid oscillations when the learning rate is set to a high value. Now, weight increments at time t are defined as:

$$\Delta w(t) = -\eta \nabla_w E(w) + \alpha \Delta w(t-1) \quad (2)$$

A standard momentum value is $\alpha = 0.9$. Learning Rate Drop Period defines the period in which learning rate is reduced by the Learning Rate Drop Factor. That way, network is searching for minimum faster in the beginning of training and is gradually slowing down as the training progresses. To prevent overfitting, L2 regularization (ridge regression) is used; in loss function besides the sum of squared errors, the sum of squared weights multiplied by the λ coefficient of regularization is added. The criterion function with ridge regression has the form:

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}w_j\right)^2+\lambda\sum_{j=1}^{p}w_j^2 \qquad (3)$$

III. Facial Image Database

Images database is made with Vlatacom Multi-Sensor Imaging System (vMSIS) [12]. Database contains 287 images in total with a resolution of 640x480 pixels. Each image in the database has a different number of persons with different facial positions, expressions and orientations, captured in different illumination conditions. Since IR is opaque to glass hiding everything behind it, database has images with persons wearing glasses.

To test the influence of different types of image degradation on detection performance, additional datasets are artificially created. In total there are four datasets, the first being original and three artificial ones. The second dataset contains images with Additive White Gaussian Noise (AGN), with a standard deviation in the range from 2 to 40, with a threshold chosen according to [13]. The third dataset contains images rotated by an angle in the range from -20 to 20 degrees. The last dataset is made of original one and the ones with deformed images described above. Datasets are shown in Fig. 2.

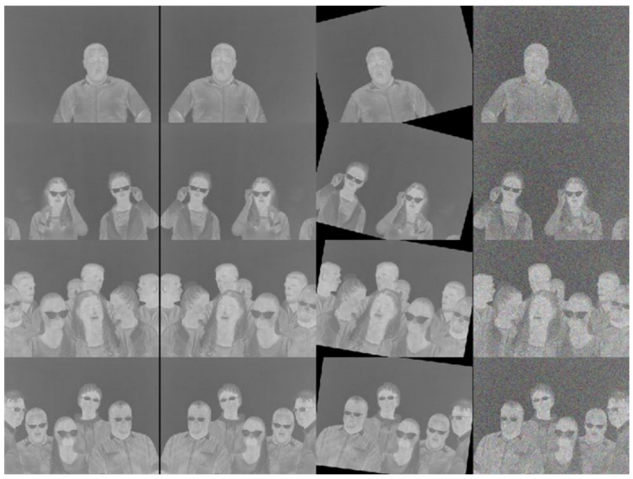

Fig. 2. Thermal image dataset with original, flipped, rotated and noisy images.

To measure the performance of system, two test datasets are created. The first test set contains 60 images from the original dataset and the second test set contains 36 images from the fourth dataset. Rest of the images are used for training sets.

IV. Hardware Platform Overview

Vlatacom Institute has been developing multi-sensor elecrooptical systems for several years [14]. A need for a specialized platform in which signal processing algorithms can be easily implemented and changed with miniature dimensions and passive cooling is emerged into developing a unified hardware platform for Video Signal Processing called vVSP. It is designed to accept any signal from various camera input interfaces [15], process the data and output it via an IP stream.

The input cameras typically are in visible light, thermal [16], [17] and short wave infrared [18]. A functional block diagram of device is shown in Fig. 3. It contains several different Printed Circuit Board (PCB) entities which are stacked to form the entire system. Those are Interface Board (IB), FPGA Carrier Board (FCB), Main Processor Unit Carrier Board (MPUCB), and Camera and Lens Control Board (CLC).

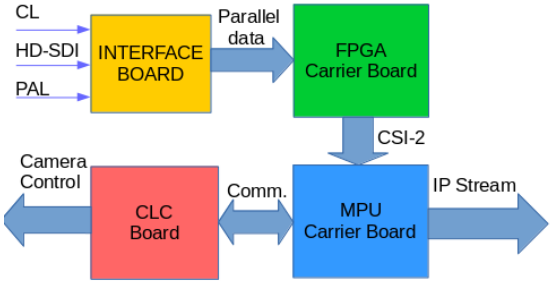

Fig. 3. System block diagram.

Interface Board (IB) is the first block in the input stage. It accepts a signal from any type of industrial camera interface (Analog PAL [19], HD-SDI [20] Base Camera Link and Medium Camera Link [21] in a format up to 1920x1080 at 30 frames per second, converts that signal into data prepared for FPGA processing and outputs it towards the second block, the FPGA Carrier Board subsystem (FCB) for further processing. FPGA board is based on Xilinx Ultrascale device [22].

FCB converts the input data into a standardized format which has the same resolution as the input signal from camera. The formatted data is provided to Main Processor Unit via a MIPI-CSI2 interface [23]. Also, the raw video output from this board can be driven directly via an output HDMI connector to a monitor or TV [24]. This feature is very useful in development phase. The board has its own precise clock generator and FPGA module which can be easily replaced to a different one, because of the board's modular design.

MPU Carrier Board acts as a central part of the vVSP module system. Besides from providing a power supply for the whole system, its main functionality is to process the data gained from FCB and output it via an IP stream. In this case scenario, as a Main Processing Unit we chose a System with 256 parallel GPU Cuda Cores and Quad-Core ARM® Cortex®-A57 as a main processor with Linux operating system with 8GB 128-bit LPDDR4 Memory [25] [26]. The two physical chips where signal processing algorithms [27] can be implemented are either FPGA (on FCB) or MPU with GPU cores. Currently, the image stabilization, objects tracking, motion detection and image enhancement algorithms are applied on this MPU. vVSP can capture and encode to H.264 stream video signal at input channel up to resolution 1920x1080 (FullHD) at 30 frames per second. This format is supported even with enabled implemented video tracking and video stabilization algorithms which add a certain load to MPU, since MPU resources are not infinite. Higher resolution formats are not supported due to lack of processing resources on MPU. Complete hardware platform with all boards and interfaces is shown in Fig. 4.

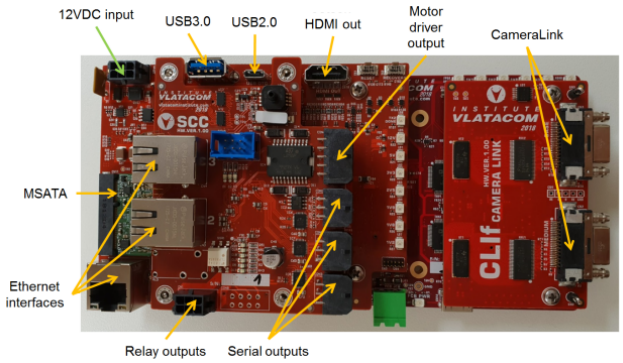

Fig. 4. vVSP module with interfaces.

Camera and Lens Control Board (CLC) is specially designed to control camera's lens functions, such as zoom and focus. It also has a motor driver implemented for moving pan-tilt platforms and network of sensors needed for some image processing methods.

vVSP platform also has an integrated 7-port Gigabit Ethernet switch so that more of this platform can be in a cascade connection for controlling a multi-sensor imaging system. One camera unit is controlled by one vVSP module.

V. Experimental Work

A test setup for described problem is shown in Fig. 5.

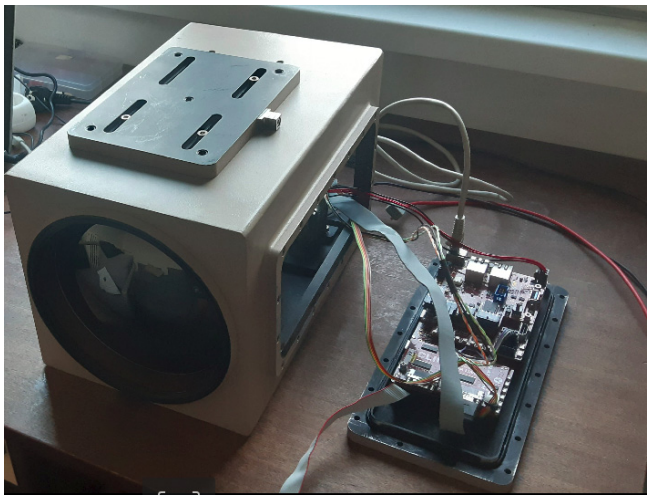

Fig. 5. vVSP module with thermal IR camera.

One of the goals of this paper was to research possibilities for face detection algorithm implementation on a vVSP module and test it in a real time scenario. In the Fig. 5. is shown a vVSP hardware platform connected to a thermal camera via Camera Link Interface Board. Camera used in this case is an uncooled thermal camera with 640x480 pixel resolution [Xenix 28] with Camera Link interface and 225mm optics [Ophir 29], range up to 8km for human detection [17].

For R-CNN training, MATLAB R2020a is used. We used CNN pretrained on CIFAR-10 dataset [30]. CIFAR-10 dataset has 50000 visible images spread out in 10 classes (airplane, bird, car, cat, deer, dog, frog, horse, ship and truck) and image resolution of 32x32 pixels. A pretrained CNN has learned filter values which are useful for feature extraction in visible images. To train R-CNN, transfer learning is implemented; CNN pretrained on CIFAR-10 dataset is fine-tuned in R-CNN training process [31]. Network is trained with mini batches of size 64 and takes 250 epochs. Weights updates are made after all the images from the mini batch have been passed through the network. Since R-CNN is used for detection, we need positive and negative regions from images used in training. Positive and negative samples are defined by the bounding box IoU metric. If image region overlaps with ground truth boxes in a positive overlap range, then that region is used as a positive sample; if they overlap in a negative range then that region is defined as a negative sample. A positive overlap range is set to be between 0.8 and 1 while a negative overlap range is less than 0.5.

Each of training datasets described in Section III is used in the training of R-CNNs, so there are 4 R-CNNs in total. Testing is performed on both test sets described in Section VI.

VI. Results and Discussion

Detection performances in this paper are presented with the following evaluation metrics: *tp* (true positive), *fn* (false negative), *fp* (false positive). The number of successfully detected faces is a true positive value; the number of faces that exist in the image but are not detected is false negative; the number of regions that R-CNN labels as face region, but do not contain face in them is called false positive. That the detection can be said to be true positive, the IoU between detected and the ground truth bounding box was adopted to be greater than 0.7. In table 1, *tp*, *fn* and *fp* values for four different R-CNNs trained on four different datasets and tested on two test datasets are shown.

TABLE 1: TP, FN AND FP VALUES OF R-CNN TRAINED ON TRAINING SETS WITH ORIGINAL AND DEFORMED IMAGES.

Training set	***Test set with original images***			***Test set with flipped, rotated and images with AGN***		
	tp	***fp***	***fn***	***tp***	***fp***	***fn***
Original images	183	11	19	109	18	15
Original, flipped and images with AGN	175	19	27	111	16	13
Original, flipped and rotated	174	15	28	110	17	14
Original, flipped, rotated, with AGN	177	15	25	111	20	13

Based on results from table 1 precision and recall can be calculated. Precision is the number of correctly detected faces divided by the total number of regions R-CNN labeled as face region in images ($p = tp / (tp + fp)$). Recall is the number of correctly detected faces divided by the total number of faces in the images ($r = tp / (tp + fn)$). As we tested algorithm on our, not publicly available test set, results are presented only by precision and recall, and not by the

average precision. High precision means that most of the regions R-CNN labels as face region are in fact a face region, while high recall means that most of the regions that are faces in images R-CNN labeled as face regions. Precision and recall values are shown in table 2.

TABLE 2: PRECISION AND RECALL VALUES OF R-CNN TRAINED ON TRAINING SETS WITH ORIGINAL AND DEFORMED IMAGES.

Training set	*Test set with original images*		*Test set with flipped, rotated and images with AGN*	
	Precision	*Recall*	*Precision*	*Recall*
Original images	**94.33**	**90.59**	85.83	87.9
Original, flipped and images with AGN	90.21	86.63	**87.4**	**89.52**
Original, flipped and rotated	92.06	86.14	86.61	88.71
Original, flipped, rotated, with AGN	92.19	87.62	84.73	**89.52**

For the first test set that only has original images the best results are achieved by R-CNN trained on just the original images. In practice however, infrared sensors often suffer from various disturbances - blur, noise originating from sensors and their environment, sensor dead pixels. Faces could also happen to be in different orientation and poses. For those reasons, results on the second test dataset are more important. On the second test dataset, the best results are achieved by the R-CNN trained on original, flipped and images with AGN.

For a better detection performance on images with diverse types of degradations, a training dataset should have images with degradation (noisy images with lower values of standard deviation of AGN) in addition to original images.

R-CNN face detection algorithm execution is fully tested on Nvidia RTX 2060. Since RTX 2060 has about 10 times more computational resources than vVSP we made a basic assumption that vVSP is up to 10 times slower than RTX 2060. Since we haven't yet developed the port of R-CNN for vVSP, as one of the goals of this paper is to research possibilities of implementation and to evaluate practicality of our solution, we compared execution of another algorithm for object detection that we had previously implemented on both platforms (Nvidia RTX 2060 and vVSP). We measured the time it takes for the YOLOv3 [32] algorithm to process the image. That object detection algorithm takes about 33ms per frame on RTX 2060 and about 312ms on vVSP. Since the results of YOLOv3 comparative test confirmed our basic assumption, we accepted the assumption that vVSP will also be 10 times slower than RTX 2060 for R-CNN execution.

The computational performance of R-CNN detection algorithm on RTX 2060 is shown in Fig. 6.

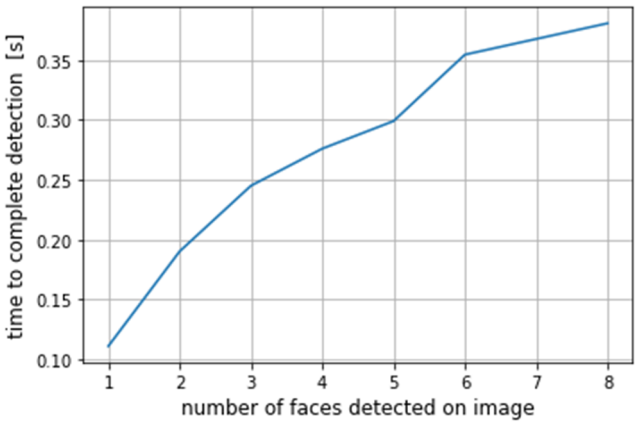

Fig. 6. Average time it took to detect all faces by number of faces detected on image. Tested on RTX 2060.

R-CNN algorithm takes more time for detection in images with more faces because it needs to do bounding box regression for more regions than it would do in images with less faces on them. If there is only one face detected in the image, we can expect 1s on Jetson TX2, but we see that if there are more faces it takes more time.

VII. CONCLUSION

In this paper a thermal face detection performance in a real operational scenario is presented, including different image degradation influence.

The advantage in performance in face detection over standard lowlight sensors is achieved by including a thermal IR sensor and has been demonstrated and shows an excellent detection performance on images captured in different lighting conditions including total darkness.

A great detection performance on images captured in different illumination conditions is obtained even when images are degraded. Robustness is achieved by including images with different types of deformations in the training set; by including those images in the training set we get worse results on non-degraded images, but better results on degraded images which are more common in practice.

In terms to improve R-CNN detection performance and robustness to image degradation basically caused by noise and rotation, it is shown that training datasets should include images with different types of deformations.

One large problem of R-CNN face detection is processing time. This method cannot be used for real-time applications. As presented in literature, the state-of-the-art methods for object detection based on deep learning approach such as *Faster R-CNN* and *YOLO* algorithm, which can reach the same detection accuracy as R-CNN with much less processing time per frame, can be engaged in face detection systems and can be implemented on a vVSP platform. Our future work in this research area will be oriented towards these deep learning methods and their actual application using vVSP with a thermal IR camera.

REFERENCES

[1] Y. K. Cheong, V. V. Yap, H. Nisar, "A Novel Face Detection Algorithm Using Thermal Imaging," *IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 208-213, 2014.

[2] Y. Zheng, "Face detection and eyeglasses detection for thermal face recognition," *Image Processing: Machine Vision Applications V*. International Society for Optics and Photonics, vol. 8300., pp. 83000C, 2012.

[3] A. Kwaśniewska, J. Rumiński, P. Rad, "Deep Features Class Activation Map for Thermal Face Detection and Tracking," *10th International Conference on Human System Interactions (HSI)*, pp. 41-47, 2017.

[4] C. Ma, N. T. Trung, H.Uchiyama, H. Nagahara, A. Shimada and R. Taniguchi, "Adapting Local Features for Face Detection in Thermal Image," *Sensors*, vol. 17, no. 12, pp. 2741, 2017.

[5] C. Herrmann, M. Ruf and J. Beyerer, "CNN-based thermal infrared person detection by domain adaptation," *Autonomous Systems: Sensors, Vehicles, Security and the Internet of Everything*, vol. 10643, pp.1064308, *SPIE*, 2018.

[6] R. F. Ribeiro, J. M. Fernandes, A. J. R. Neves, "Face Detection on Infrared Thermal Image," *SIGNAL 2017 : The Second International Conference on Advances in Signal, Image and Video Processing*, pp. 38-42, Spain, 2017.

[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014.

[8] C. Junli, J. Licheng, "Classification Mechanism of Support Vector Machines," *WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, vol. 3, pp. 1556-1559, 2000.

[9] J. R. R. Uijlings, K. E. A. Van de Sande, T. Gevers, A. W. M. Smeulders, "Selective Search for Object Recognition," *International Journal of Computer Vision*, pp.154-171, 2013.

[10] S. Eger, P. Youssef, I. Gurevych, "Is it Time to Swish? Comparing Deep Learning Activation Functions Across NLP tasks," *EMNLP*, pp. 4415-4424, 2018.

[11] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures," In: Montavon G., Orr G.B., Müller KR. (eds) Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol 7700, Springer, Berlin, Heidelberg, pp. 437–478, 2012.

[12] Vlatacom Institute, "Vlatacom Institute Border Protection-Land," [Online]. Available: https://www.vlatacominstitute.com/border-protection-land".

[13] S. Dodge and L. Karam, "Understanding How Image Quality Affects Deep Neural Networks," *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pp. 1-6, 2016.

[14] Vlatacom Institute, "vMSIS3-CHD-1200-T Vlatacom Multi Sensor Imaging Sysem 3 - Cooled High Definition", Available: https://docs.wixstatic.com/ugd/510d2b_ccea8cf0cd674d898c05e23ab58562a5.pdf [Accessed: Mar. 20, 2019].

[15] H. Aghajan, A. Cavallaro, *Multi-Camera Networks: Principles and Application*, Academic Press; 1 edition (April 25, 2009).

[16] H. Budzier, G. Gerlach, *Thermal Infrared Sensors: Theory, Optimisation and Practice*, Wiley; 1 edition (February 14, 2011).

[17] D. Peric, B. Livada, M. Peric, S. Vujic, "Thermal imager range: predictions, expectations and reality," *Sensors* vol. 19, no 15, pp. 3313, 2019, doi:10.3390/s19153313.

[18] D. Peric, B. Livada, "Analysis of SWIR Imagers Application in Elecrto-Optical Systems,", presented at conference *IcETRAN 2017*, at Kladovo, Serbia.

[19] ITU-R Recommendation BT.656,Interfaces for digital component video signals in 525-line and 625-line television systems operating at the 4:2:2 level of Recommendation ITU-R BT.601 (Part A), International Telecommunication Union, 2007.

[20] "SMPTE 259M- SDTV Digital Signal/Data Serial Digital Interface," ANSI/SMPTE, 2008.

[21] "Specifications of the Camera Link Interface Standard for Digital Cameras and Frame Grabbers, version 2.0.," Automated Imaging Association, Ann Arbor, 2012.

[22] Xilinx, "UltraScale Architecture and Product Data Sheet: Overview," Ultrascale Datasheet, DS890 (v3.7) February 20, 2019. Available: https://www.xilinx.com/support/documentation/data_sheets/ds890-ultrascale-overview.pdf, [Accessed: Mar. 20, 2019]

[23] MIPI Alliance,"Evolving CSI-2 Specification," Technology Brief, Available: https://www.mipi.org/sites/default/files/MIPI_CSI-2_Specification_Brief.pdf, [Accessed: Mar. 20, 2019]

[24] "High-Definition multimedia interface, Specification Version 1.4," HDMI Licensing, LLC, 2009.

[25] NVIDIA, "NVIDIA Jetson TX2 Series System-on-Module Pascal GPU + ARMv8 + LPDDR4 + eMMC," Jetson TX2 Series Datasheet 1.2, subject to change, copyright © 2014 – 2018 NVIDIA Corporation.

[26] R. Farber, *CUDA Application Design and Development*, 1st edition November 14, 2011.

[27] S. Dhanani, M. Parker, *Digital Video Processing for Engineers: A Foundation for Embedded Systems Design*, Newnes: 1st edition (October 24, 2012)

[28] https://xenics.com/files/technical_resources/XTM%20Series/xb-123_r009_xtm_640_lowres.pdf

[29] https://www.ophiropt.com/infrared/lenses/supir-25-225mm-f-1-5/

[30] http://www.cs.toronto.edu/~kriz/cifar.html

[31] https://www.mathworks.com/help/vision/examples/object-detection-using-deep-learning.html.

[32] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv preprint arXiv:1804.02767*.

[33] T. Vuković, R. Petrović, M. Pavlović and S. Stanković, "Thermal Image Degradation Influence on R-CNN Face Detection Performance," *2019 27th Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 2019, pp. 1-4.