Metrol. Meas. Syst., Vol. 28 (2021) No. 2, pp. 307–321
DOI: 10.24425/mms.2021.136609

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS
Index 330930, ISSN 0860-8229
www.metrology.wat.edu.pl

M&MS

DETECTION OF HUMAN FACES IN THERMAL INFRARED IMAGES

Marcin Ł. Kowalski, Artur Grudzień, Wiesław Ciurapiński

Military University of Technology, Institute of Optoelectronics, gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, Poland
(✉ *marcin.kowalski@wat.edu.pl*, +48 261 839 353, *artur.grudzien@wat.edu.pl*, *wieslaw.ciurapinski@wat.edu.pl*)

Abstract

The presented study concerns development of a facial detection algorithm operating robustly in the thermal infrared spectrum. The paper presents a brief review of existing face detection algorithms, describes the experiment methodology and selected algorithms. For the comparative study of facial detection three methods presenting three different approaches were chosen, namely the Viola–Jones, YOLOv2 and Faster-RCNN. All these algorithms were investigated along with various configurations and parameters and evaluated using three publicly available thermal face datasets. The comparison of the original results of various experiments for the selected algorithms is presented.

Keywords: thermal infrared, face detection, biometrics, human detection.

1. Introduction

Face detection is a critical step in every face recognition system because at the stage of face detection, the proper region of interest is defined. Face detection aims to automatically provide the size and coordinates of facial region in an image. There are numerous methods for detecting faces and variety of detection approaches ranging from image segmentation methods to deep learning methods. Without robust facial detection algorithm operating effectively in various conditions the facial recognition system is not performing correctly.

The most widely explored area of facial detection relies on visible range images while less attention has been paid to infrared and thermal infrared face detection algorithms. Since the algorithms performing well in the visible domain do not operate well in the thermal infrared domain, the aim of this study is to compare three different approaches for thermal face detection. We compare the Viola–Jones algorithm which is a well-known machine learning algorithm, YOLOv2 which is a deep learning model that can simultaneously predict multiple bounding boxes and class probabilities for objects very quickly and the Faster R-CNN, which is a deep learning network based on *region proposal network* (RPN).

The paper is divided as follows. In Section 2, related works are presented. Section 3 presents algorithms explored during this study while Section 4 shows the methodology of experiments,

Article history: received September 7, 2020; revised November 30, 2020; accepted January 5, 2021; available online March 19, 2021.

characteristics of collected samples and processing methods. Results are presented in Section 5 with the summary provided in Section 6.

2. Related works

A variety of methods for face detection have been proposed. One of them is the *Active Shape Model* (ASM) which describes complex non-rigid features like actual physical and higher-level appearance of features. ASM should automatically locate landmark points in order to define the shape of any statistically modelled object in an image. Landmark points are matched to features such as eyes, lips, nose, mouth and eyebrows. One of the first algorithms relating to this technique was proposed by Kass *et al.* in 1987 [1].

Crowley and Coutaz in 1997 developed skin-color algorithms for detecting skin pixels in order to detect a face in an image [2]. This method uses the LowLevel Analysis technique. The consecutive stages occur when processing face images based on skin areas in an appropriate color space (RGB, YCbCr, HSV). The threshold is calculated and used to mask the skin region and to extract the face and determine the bounding box. Another method different from the Low-Level Analysis technique is based on edge analysis. Face detection based on edges was introduced by Sakai *et al.* in 1972 [3]. They analysed edges of faces in an image, aiming to locate specific facial features.

The Feature Analysis technique analyzes and finds structural features regardless of conditions such as pose, viewpoint, or lighting conditions, and then uses these to locate faces. The most known algorithm based on the Feature Analysis technique is the Viola–Jones algorithm for face detection in real time for visible images. It was introduced in 2004 by Paul Viola and Michael J. Jones [4] and is a combination of Haar features (or others) and an AdaBoost machine learning algorithm. It was introduced to provide high speed of detection with relatively good efficiency of detection. Moreover, it is reported to achieve a very low false detection rate. However, it is also suffering with long training period and limited head postures.

The development of deep learning algorithms, including great contribution of neural networks, has resulted with significant performance improvement to object detection. Several object detection methods based on *convolutional neural networks* (CNN) are applied to face detection. Lin *et al.* proposed the MletNet method in 2016 [5]. They modified known LeNet CNN structure by changing the output nodes. This method was proposed to detect possible terrorists. Zhang et al. proposed multi-task cascade Convolutional Networks to detect faces. The main role in this algorithm is played by a cascaded multi-task framework which exploits the inherent correlation between the tasks of detection and alignment to boost up the performance of face detection [6]. However, it is not computationally efficient because uses multiple CNN subnetworks to detect faces.

Currently, the state-of-the-art CNN architectures include: (1) single pass approaches that perform detection within a single step (*Single Shot Multi-Box Detector* (SSD) [7], YOLO [8] and (2) region-based approaches that exploit a bounding box proposal mechanism prior to detection (*Faster Regional-CNN* (RCNN) [9], *Region-based Fully Convolutional Networks* (RFCN) [10], PVANET [11], Local R-CNN [12]).

In 2016 Qin *et al*. proposed a FaceCraft structure based on a *region proposal network* (RPN) and a Faster R-CNN model to improve the detection rate of the Fast R-CNN method for face detection [13]. Fast R-CNN is a time-consuming method since it extracts features from an image pyramid. Faster R-CNN computes candidate regions by a fully convolutional RPN. The adopted region proposal, and region classification approaches are aimed to achieve detection and

classification capability. It uses a bank of k2 position-sensitive score maps for each category produced by the last convolutional layer. In 2016 Jiang and Learned-Miller used Faster R-CNN algorithm for face detection in the visible light spectrum [14]. It demonstrated impressive face detection performance when retrained on a suitable training set.

A great majority of works presented in this section relate to the visible spectrum with limited number of works concerning thermal infrared. In 2010 Mekyska *et al*. presented research on face segmentation in face detection in thermal images [15]. In 2017 Kopaczka *et al*. reviewed and compared various methods for facial detection [16]. They also proposed a method where landmarks were transferred from visible face images to develop a face detector for thermal images. They used the Viola–Jones method for different methods of extraction of features and algorithms, the Eye Corner Detection algorithm and the Projection Profile Analysis algorithm. Ma *et al*. in 2017 used a combination of the Multi-Block Local Binary Pattern and a cascade classifier for face detection in thermal images [17]. While face detection algorithms operating in the visible range provide high performance, a small number of works in the thermal infrared domain shows the need to develop fast and highly efficient algorithm for this specific purpose. Moreover, this study shows that the methods operating very well in the visible range do not necessarily operate robustly in thermal infrared [18]. The aim of this paper is to compare three different object detection algorithms trained for thermal face detection. The paper concerns three methods, the machine learning Viola–Jones, and two deep learning methods YOLOv2 and Faster-RCNN.

3. Human face detection algorithms

3.1. Introduction

Current methods for object detection fall into the category of either machine-learning based approaches or deep-learning based ones. Machine learning methods utilize various image features to detect objects while deep learning techniques are able to do end-to-end object detection without specifically defined features. Current CNN-based methods are able to perform detection and classification tasks in single architecture. A combination of various approaches allows to effectively detect and classify various objects of interests with relatively high processing speed.

For this study three methods were selected, namely the Viola–Jones, YOLOv2 and Faster R-CNN. The Viola–Jones is a well-known algorithm that has been tested for facial detection in the visible domain. It is fast but has been outperformed by current deep learning methods with regard to detection rate and accuracy. As a second method, YOLOv2 was selected as an object detection method targeted for real-time processing. The third algorithm, Faster R-CNN represents the region-proposal approach and is known as highly efficient but computationally complex. All three algorithms work in various configurations using various features (Viola–Jones) or various neural networks for feature extraction (YOLOv2, Faster R-CNN). We selected the same set of feature extraction networks for YOLO and Faster-RCNN. All the algorithms have been briefly introduced in the subsections below.

3.2. Viola–Jones

The Viola–Jones algorithm is a cascade algorithm that uses one of three different methods for feature extraction. This detector is based either on a *Histogram of Oriented Gradients* (HOG) [19], *Local binary patterns* (LBP) [20], or Haar features [4]. The cascade of classifiers is trained using an AdaBoost algorithm. The algorithm is known to be fast and robust.

The processing scheme of the Viola–Jones algorithm consists of following steps: extraction of features, creating an integral image, AdaBoost training, cascading classifiers and cascade machine learning. The algorithm was tested with all three feature extraction methods. Each method resulted in the creation of a separate face detector which was evaluated later on. We established the same training options for each of the extraction methods.

3.3. YOLOv2

You Only Look Once, also referred to as YOLO is known as very a fast object detection algorithm. For this study, a second iteration of the algorithms was considered. The YOLOv2 object detection network is composed of two subnetworks - a feature extraction network followed by a detection network. For this study, various feature extraction networks were used. They are described in Subsection 3.5.

The object detection process relies on a single convolutional network that simultaneously predicts multiple bounding boxes and class probabilities for those boxes. The YOLOv2 considers object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. The network uses Batch-Normalization to normalize the outputs of the hidden layers and Anchor-Boxes to pre-define the size of the bounding box thus improving detection performance. The classification is done with independent logistic classifiers to calculate the likeliness of the input belonging to a specific label. It predicts all bounding boxes across all classes for an image simultaneously. During the processing of an image, the image is taken globally to make predictions. During the training, YOLO learns generalizable representations of objects.

The second version of YOLO introduced anchor boxes which allowed to improve algorithms performance while maintaining the processing speed. Finding the optimal number of anchors and calculating proposal anchor boxes for training dataset is critical to achieve performance at a satisfactory level.

To estimate anchor boxes from training data we calculated mean value of *Intersection over Union* (IoU) distance metric [21]. IoU is invariant to the size of boxes, unlike the commonly used Euclidean distance metric, which produces larger errors as the box sizes increase. Moreover, while using IoU distance metric we are able to obtain anchor boxes of similar aspect ratios and sizes being clustered together.

Choosing the number of anchors for training requires calculation for each of the data clusters. We have used a k-means clustering algorithm with mean IoU distance metrics of the boxes in each cluster to calculate the overlap ratio and to estimate the optimal number of anchor boxes. This analysis was made iteratively for various numbers of anchors. The graph presenting mean IoU versus number of anchors is shown in Fig. 1. We made the assumption that the mean IoU should be greater than 0.7 which ensures that the anchor boxes overlap well with the ground-truth boxes in the training dataset. The training was made for several different feature networks described in Subsection 3.5.

3.4. Faster R-CNN

The Faster R-CNN is based on the idea of a region proposal network. The RPN outputs the objectness score for many proposed boxes which indicate whether the selected part of an image contains a background object or a foreground one. All the boxes are examined by a classifier and regressor to check the occurrence of objects. The Faster R-CNN is composed of two networks:

Metrol. Meas. Syst., Vol. 28 (2021), No. 2, pp. 307–321
DOI: 10.24425/mms.2021.136609

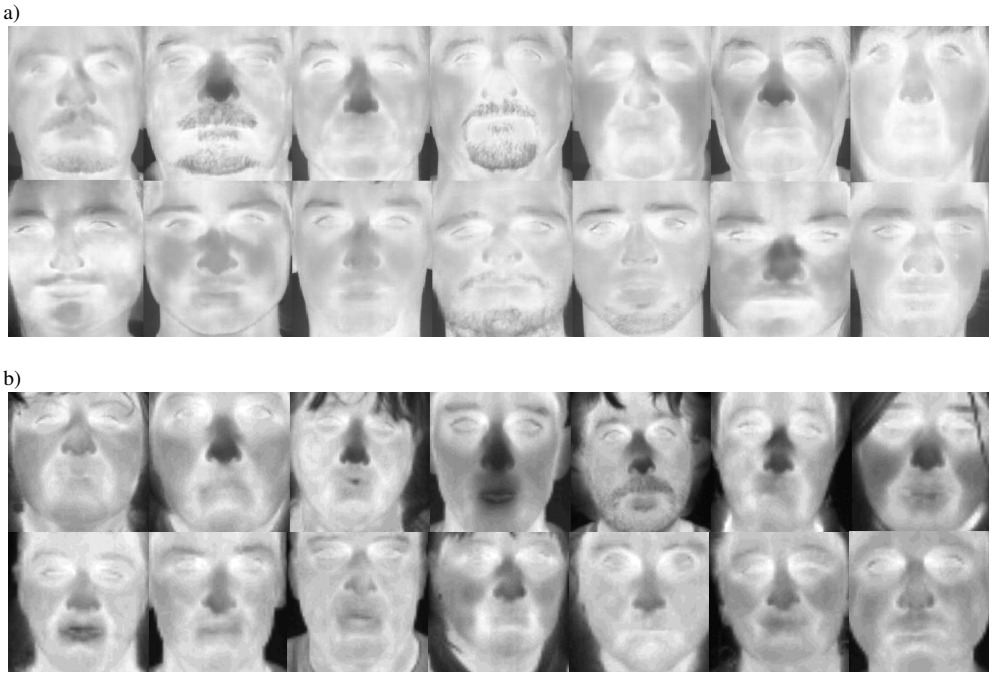


Fig. 1. Gallery of thermal facial images from (a) the in-house dataset, (b) the Carl dataset.

a region proposal network for generating region proposals and a network using these proposals to detect objects.

Similarly to YOLO, this algorithm uses anchor boxes, however, they are generated automatically instead of being pre-defined before training. It uses a bank of k^2 position-sensitive score maps for each category. Those features are calculated by the last convolutional layer. Faster R-CNN as well as YOLO can use several backbone networks for feature extraction.

Faster R-CNN ends with a position-sensitive RoI pooling layer. This layer aggregates the outputs of the last convolutional layer and generates scores for each RoI:

$$r_c\left(i,j|\Theta\right) = \sum_{(x,y)\in bin(i,j)} \frac{z_{i,j,c}\left(x+x_0, y+y_0|\,\Theta\right)}{n}, \tag{1}$$

where $r_c(i,j)$ is the pooled response in the (i,j)-th bin for the c-th category, n is the number of pixels in the bin, and Θ denotes all learnable parameters of the network. After calculating all the values for the position-sensitive ROI pool, the final class score is the average of all its elements. Each RoI produces a $(C+1)$-dimensional vector, where C is the number of object categories:

$$r_c(\Theta) = \sum_{si,j} r_c\left(i,j|\,\Theta\right). \tag{2}$$

The softmax computes responses across categories:

$$s_c(\Theta) = \frac{e^{r_c}}{\sum\limits_{c'=0}^{C} e^{r_{c'}(\Theta)}}. \tag{3}$$

3.5. Feature extraction networks

Both deep learning object detection algorithms (YOLO and Faster–RCNN) investigated in this study use different feature extraction neural networks. We had selected common set of networks that were used as feature extraction backbones for both algorithms. The selected CNNs present different approaches to feature extraction and are composed of different numbers of convolutional layers.

In this study we applied the transfer learning scheme with GoogLeNet [22], [23] ResNet18 [24], ResNet50 [24], ResNet101 [24], AlexNet [25], VGG16 [26] and VGG19 [26] and Darknet53 [27] as feature-backbone networks. All of these models were pretrained on ImageNet database. A short description of each network is provided below.

AlexNet

AlexNet is a network composed of only eight layers: five convolutional layers and three fully-connected layers. The network is focused on smaller window sizes and strides in the first convolutional layer.

VGG networks

Two VGG networks, VGG16 and VGG19 were used. Instead of using large receptive fields such as AlexNet (11×11 with a stride of 4), VGG uses very small receptive fields (3×3 with a stride of 1). Because there are now three ReLU units instead of just one, the decision function is more discriminative. There are also fewer parameters (27 times the number of channels instead of AlexNet's 49 times the number of channels). VGG incorporates 1×1 convolutional layers to make the decision function more non-linear without changing the receptive fields.

ResNets

A residual neural network (ResNet) is a family of networks proposed as a solution for vanishing gradients. These CNNs use the so-called identity shortcut connection that skips connections, or shortcuts to jump over some layers. Typical ResNet models are implemented with double- or triple- layer skips that contain nonlinearities (ReLU) and batch normalization in between. The network constructs pyramidal cells in the cerebral cortex during data processing. ResNet networks are proposed in many variants. During this study ResNet18, ResNet50 and ResNet101 were investigated.

Inception networks

An inception network also referred to as GoogLeNet was used. The Inception network introduces Inception layers, sparsely connected network architectures which have replaced fully connected network architectures. Each Inception Layer is a combination of a 1×1 Convolutional layer, a 3×3 Convolutional layer, and a 5×5 Convolutional layer with their output filter banks concatenated into a single output vector forming the input of the next stage. The Inception network aims to be more effective than traditional CNNs reducing redundant or unnecessary activations. Moreover, the Inception network replaced the fully-connected layers at the end with a global average pooling which averages out the channel values across the 2D feature map, after the last convolutional layer. This reduces the total number of parameters. In effect, the network is less prone to overfitting. The second version of the Inception network introduced several changes including smart factorization and expansion of filter banks.

4. Measurement methodology

4.1. Introduction

The thermal image of a human face presents its unique heat-signature which can be used for facial recognition. Analysis of relative temperature distribution on the surface of a face can reveal individual patterns of intensity variations. Thermal infrared imaging does not need illumination since it relies on heat emitted by all bodies and objects in the field of view of the camera. The radiation registered is proportional to relative distribution of the apparent temperature of objects. The projection of objects depends on *noise equivalent temperature difference* (NETD) and optics as well as temperature difference between objects and their emissivity. The acquired image strongly depends on environmental conditions during the acquisition process.

In order to capture the shape and structure of the face, the imager needs to distinguish very small amounts of energy. The ability to capture and quantify the thermal energy depends on the camera's parameters, in particular the *noise equivalent temperature difference*. Thermal face imaging is sensitive to changes of emotional, physical and health condition of the subject. Moreover, thermal properties of the face depend on the temperature of the body, environmental conditions and occlusions present on the face such as scarfs, hairs, facial hairs, glasses or any disguise accessories that alter the emitted heat pattern. Below we present the methodology of the experiment.

4.2. Dataset

Several face datasets collected in thermal infrared are available, however none is annotated. In this study, three thermal face datasets were used to train and test the algorithms. Two of the datasets (in-house and PROTECT) were used for training, while part of all three was used for testing. All the datasets were manually annotated. The datasets are briefly described below:

a) The first database is an in-house dataset reported in [28]. The thermal images of subjects' faces were collected using a thermal imager operating in the range of 7.5–13 μm. The pixel resolution of the camera was 640×512 pixels with *minimum resolvable temperature difference* (MRTD) below 50 mK for 300 K. We collected over 1100 thermal images in frontal and various head positions for over 100 subjects. All images have been recorded at a distance of 1.2 meters. The dataset is composed of subsets containing images acquired during respective measurement sessions. The dataset is composed of intensity images with 14-bit grayscale depth. The imager was not calibrated to acquire accurate temperature values.

b) PROTECT Multimodal database [29] presents a combination of 9 biometric traits, namely 2D face, 3D face (RGB and Depth Field), thermal face, iris, voice, finger veins, hand veins and anthropometrics. For our studies we used a thermal face dataset. The pixel resolution of the thermal images was 640×512 pixels with minimum resolvable temperature difference of 50 mK for 300 K. All the data were collected indoors, simultaneously for 47 subjects.

c) The Carl database [30, 31] contains images collected in three spectra, visible, near infrared and thermal infrared. Thermal infrared images are of resolution of 160×120 pixels with MRTD of 100 mK. All images were collected indoors in stand-alone positions of 41 subjects.

It should be stressed here that when comparing algorithms using different datasets we compare thermal images recorded in different conditions. That fact implies that the differences in accuracy may not necessarily result from the specificity of algorithms, but also from differences in the

measurement process of image acquisition. However, the aim of this study is also to provide insight on how the algorithms perform across various conditions.

In order to train the Viola–Jones algorithm, we prepared a database of non-facial thermal images (a negative dataset). The database contains over 3000 images of various objects. A gallery of the thermal face images is presented in Fig. 1. The datasets were divided into train, and test subsets with a split ratio set to 75% to 25%, respectively.

4.3. Training

The training process of an object detector requires a database with annotated images containing properly selected regions of interest. For training, 1148 thermal images, each containing a single face were manually annotated. The training dataset is a composition of images brought from two databases ("in-house" and PROTECT). In order to prepare the algorithm to provide optimal RoI for face recognition, only facial regions without hair and ears were selected as presented in Fig. 2. The desired facial region was defined starting from the below the chin and through cheeks and eyes and finishing above the eyebrows. The selection of region was aimed to minimize the impact of changing hairstyle which may strongly affect the facial thermogram.

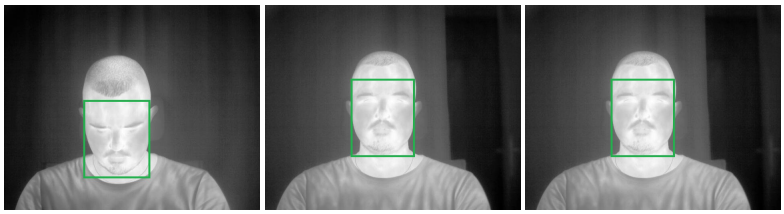


Fig. 2. Examples of manually annotated bounding boxes for training the neural network model.

All three algorithms were trained and tested using the same set of images. In order to provide an in-depth study, all algorithms were trained with various parameters including different features, various number of epochs and other parameters.

5. Results and discussion

Each face detector was tested to assess its performance. During the study, we used the Intersection over Union metric which is the most commonly used metric to measure the accuracy of an object detector [19]. This metric determines the percentage of overlap between a predicted bounding box and the ground-truth bounding boxes as presented in Fig. 3. The green rectangle refers to the ground-truth bounding box, while the red one refers to the predicted bounding box. The ratio of the common part of the surface areas of two rectangles to the sum of two surface areas defines accuracy of detection. The three algorithms were tested using testing datasets. Each dataset was composed of over 250 images containing one face per image.

The number of thermal face databases is limited and very often collected with low thermal and spatial resolution equipment. This situation can be due to thermal imagers being relatively expensive. In our study, however, for the evaluation purposes the authors' dataset of thermal faces, the Carl dataset as well as the PROTECT thermal face dataset which were collected using state-of-the-art thermal infrared imagers. Subsequently, in order to evaluate the trained algorithms, all the datasets were manually annotated with definitions of ground-truth bounding boxes.

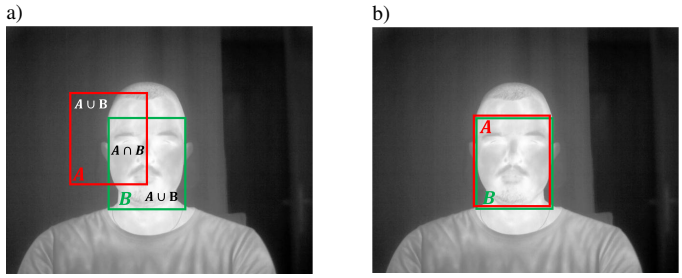


Fig. 3. a) Visualization of bounding box regions for Intersection over Union metric; b) Example of well-fitted bounding boxes for Intersection over Union metric.

All the algorithms were evaluated with regard to detection rate, false detection rate, processing speed and mean IoU. The detection rate corresponds to overall ratio between the number of correctly detected faces and the number of faces in the testing dataset. It is defined by the following formula:

$$DetRate = \frac{CD}{NI}, \tag{4}$$

where CD is the number of correctly detected faces in the testing dataset, NI is the overall number of faces in the testing dataset.

The false detection rate corresponds to overall ratio between the number of non-detected faces and incorrectly detected faces to the number of faces in the testing dataset. All the cases when non-facial object is detected are assumed to be incorrectly detected faces. The false detection rate takes values from 0 and is defined as follows:

$$FalseDetRate = \frac{(ND + IN)}{NI}, \tag{5}$$

where ND is the number of non-detected faces and IN is the number of incorrectly detected faces. It should be noted that all the images in the testing split contain a single face per image. In the case of the detection rate equal to 1, the non-zero false detection rate means that all the false detections result from incorrect detections.

The statistical values of detection performance were presented in Table 1. The presented values correspond to the best performing algorithms.

All three compared algorithms achieved high detection rates (the number of correctly detected faces over the testing dataset), very close or equal to 1.00. The difference between algorithms is visible when comparing false detection rates. False detection rate defines for how many images the algorithm detected additional objects over the entire training set. Considering both performance indicators, Faster R-CNN provided the highest performance, achieving an almost perfect detection rate with a very low false detection rate. YOLO also provides a very high detection rate, however, some variants of this algorithm provided a very high false detection rate up to 29%. In the case of both deep-learning algorithms, ResNet101 achieved the highest performance scores.

The Viola–Jones was outperformed by both deep-learning algorithms because this algorithm generated an extremely high number of false detections. Sample images of thermal face images after detection by the Viola–Jones are presented in Fig. 4. Both images present additional, false bounding boxes generated by the algorithm. The high false detection rate eliminates this algorithm from real-life applications.

Table 1. Detection performance of investigated algorithms.

Viola–Jones						
Feature extraction method	**Det. rate**	**False Det. Rate**	**Det. rate**	**False Det. Rate**	**Det. rate**	**False Det. Rate**
	In-house		**PROTECT**		**CARL**	
Haar	1.00	0.21	1.00	0.34	1.00	0.03
LBP	1.00	0.54	1.00	0.57	1.00	0.05
HOG	1.00	0.23	0.99	0.25	0.97	0.02
YOLOv2						
ResNet18	0.98	0.03	1.00	0.00	1.00	0.00
ResNet50	1.00	0.01	1.00	0.00	1.00	0.03
ResNet101	1.00	0.00	1.00	0.00	1.00	0.00
Alexnet	0.99	0.00	1.00	0.07	1.00	0.02
GoogLeNet	1.00	0.01	1.00	0.02	1.00	0.01
VGG16	0.99	0.03	1.00	0.09	1.00	0.29
VGG19	0.98	0.00	0.98	0.00	1.00	0.00
Darknet-53	0.99	0.01	0.99	0.02	1.00	0.01
Faster R-CNN						
ResNet18	0.98	0.01	1.00	0.00	1.00	0.00
ResNet50	0.99	0.00	1.00	0.00	1.00	0.00
ResNet101	1.00	0.00	1.00	0.00	1.00	0.00
Alexnet	0.99	0.01	1.00	0.00	0.98	0.00
GoogLeNet	0.98	0.00	1.00	0.00	1.00	0.00
Darknet-53	0.97	0.00	1.00	0.00	1.00	0.00

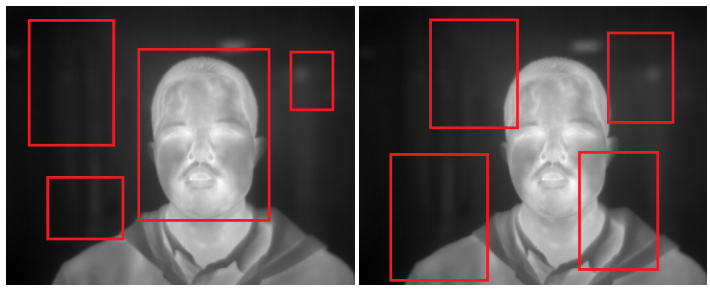


Fig. 4. Sample thermal images showing results of the Viola–Jones face detector.

Other works reported the results presented in Table 2. The works in [16] rely on thermal infrared images acquired with a very sensitive imager of higher resolution (1024 x 768 pixels). The results reported in [15] were achieved on a low resolution Carl database.

Another performance indicator used for validation of face detection algorithms is *Intersection over Union* (IoU). This metric was used to verify how well the bounding box generated by fthe ace detector algorithm fits the ground-truth bounding box.

Table 2. Detection performance reported in related works.

Method	True Positive Rate[1]	False Positive Rate[1]
Deformable Parts Model [16]	0.98	0.00
Eye Corner Detection [16]	0.52	0.48
Projection Profile Analysis [16]	0.85	0.15
HoG [16]	0.93	0.00
LBP [16]	0.90	0.03
V-J [16]	0.91	0.07
PICO [16]	0.90	0.10
V-J [15]	0.93	–

[1] All numbers are given in absolute values.

Below we present mean values of IoU for each of the algorithms (Table 3). The presented values correspond to the best performing algorithms. As presented in Table 3, both deep-learning algorithms outperformed the Viola–Jones and provided comparable results. YOLO did not achieve the expected mean IoU pre-defined during training (0.887).

Table 3. The best values of mean IoU.

Feature extraction methods	Mean IoU[1]		
	In-house	PROTECT	CARL
Viola–Jones			
Haar	0.47	0.46	0.52
LBP	0.48	0.48	0.53
HOG	0.46	0.48	0.48
YOLOv2			
ResNet18	0.64	0.68	0.73
ResNet50	0.77	0.78	0.77
ResNet101	0.76	0.78	0.79
Alexnet	0.57	0.58	0.49
GoogLeNet	0.74	0.76	0.73
VGG16	0.64	0.67	0.65
VGG19	0.70	0.69	0.60
Darknet-53	0.74	0.76	0.73
Faster R-CNN			
ResNet18	0.78	0.81	0.80
ResNet50	0.78	0.81	0.77
ResNet101	0.68	0.71	0.76
Alexnet	0.76	0.80	0.63
GoogLeNet	0.72	0.74	0.73
Darknet-53	0.68	0.71	0.73

[1] All numbers are given in absolute values.

Values of Intersection over Union provide very important information for multi-stage algorithms including face detectors. Face detection is only a part of a longer process thus the correct generation of the bounding box significantly influences the working of the face recognition system. Results of our studies indicate that the sufficient value of mean IoU is around 80%. Such values can be achieved by YOLO and Faster R-CNN based on ResNet101.

It has to be noted that values of mean IoU vary for YOLO over the training process while for Faster R-CNN they do not vary significantly. The exception for this rule is ResNet101 which for selected epochs provides the highest scores but also achieves zero mean IoU for some epochs.

The presented methods were implemented on an NVIDIA based GPU (GTX 1080-Ti) in MATLAB 2019a environment and are reported to process a single thermal image (640 x 512 pixels) at various speeds. Average processing times for single images for each of the algorithms are provided in Table 4.

Table 4. The best values of mean IoU.

Method	Feature CNN	Processing time [s]
Viola–Jones	LBP	0.0105
	Haar	0.0116
	HoG	0.0564
Faster R-CNN	Alexnet	0.0553
	GoogLeNet	0.3558
	ResNet18	0.1405
	ResNet50	0.3996
	ResNet101	0.5225
	Darknet-53	0.1405
YOLOv2	Alexnet	0.0085
	GoogLeNet	0.0130
	ResNet18	0.0101
	ResNet50	0.0204
	ResNet101	0.0218
	VGG16	0.0109
	VGG19	0.0117
	Darknet-53	0.0109

The slowest and the most greedy algorithm is Faster R-CNN which needs a high power GPU to process a single frame in less than one second. Average processing times of Viola–Jones are much shorter than those for Faster R-CNN. The biggest processing speed achieved by the Viola–Jones was up to 95 frames per second. YOLO showed the highest processing speed. The fastest version of YOLO based on AlexNet is able to process over 110 frames per second.

6. Conclusions

Face detection algorithms play a crucial role in facial recognition systems. Due to a relatively low number of works on thermal infrared facial recognition, studies in this field are desirable. We have presented a comparative study of three object detection algorithms presenting three different approaches to object detection. The Viola–Jones, YOLO and Faster R-CNN were trained and

fine-tuned for thermal detecting of faces. Both methods were used in various configurations. Both algorithms were evaluated using three datasets using four different performance indicators. All the presented results show that in order to assess the overall capability of detection algorithm, several performance indicators have to be considered together.

Faster R-CNN provided the highest performance, achieving almost perfect detection rate with a very low false detection rate. However, this algorithm requires the biggest computational resources and is the slowest among the three investigated. YOLO is the fastest tested algorithm achieving a high detection rate together with a low false detection rate.

In the case of deep-learning algorithms, both based on ResNet101 achieved the highest scores of detection rate, lowest false detection rates and high mean IoU.

The Viola–Jones algorithm was outperformed by Faster-RCNN and YOLO in most configurations. The study also shows that selection of a proper feature extraction method is of great importance.

Acknowledgements

This research was funded by the Military University of Technology; grant number ZBW/08-894/2020/WAT.

References

[1] Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active Contour Models. *International Journal of Computer Vision*, *1*, 321–331. https://doi.org/10.1007/BF00133570

[2] Crowley, J. L., & Coutaz, J, (1997). Vision for Man Machine Interaction. *Robotics and Autonomous Systems*, *19*, (3-4), 347–358. https://doi.org/10.1016/S0921-8890(96)00061-9

[3] Sakai, T., Nagao, M., & Kanade, T. (1972). Computer Analysis and Classification of Photographs of Human Faces. *Proceedings of First USA-JAPAN Computer Conference*, Japan, 55–62.

[4] Viola, P. & Jones, M.J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision, 57*, 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[5] Lin, S., Cai, L., Lin, X., & Ji, R. (2016). Masked face detection via a modified LeNet. *Neurocomputing*, *218*, 197–202. https://doi.org/10.1016/j.neucom.2016.08.056

[6] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C (2016). SSD: Single Shot MultiBox Detector. *Proceedings of European Conference on Computer Vision,* The Netherlands, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), USA, 779–788. https://doi.org/10.1109/CVPR.2016.91

[9] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards Real-Time Detection with Region Proporsal Networks. *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence*, USA, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[10] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Spain, 397–387. https://arxiv.org/abs/1605.06409

[11] Kim, K.H., Hing, S., Roh, B., Cheon, Y., & Park, M. (2016). PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. *Proceeding of Conference on Computer Vision and Pattern Recognition*, USA. https://arxiv.org/abs/1608.08021

[12] Vu, T. H., Osokin, A., & Laptev, I. (2015). Context-Aware CNNs for Person Head Detection. *Proceedings of IEEE International Conference on Computer Vision*, Chile, 2893–2901. https://doi.org/10.1109/ICCV.2015.331

[13] Qin, H., Yan, J., Li, X., & Hu, X. (2016). Joint Training of Cascaded CNN for Face Detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, USA, 3456–3465. https://doi.org/10.1109/CVPR.2016.376

[14] Jiang, H., & Miller, E. L. (2017). Face Detection with the Faster R-CNN. *Proceedings of 12th IEEE International Conference on Automatic Face & Gesture Recognition, USA, 650–657*. https://doi.org/10.1109/FG.2017.82

[15] Mekyska, J., Duro, V. E., & Zanuy, M. F. (2010). Face Segmentation: A comparison between visible and thermal images. *Proceedings of 44th Annual 2010 IEEE International Carnahan Conference on Security Technology*, USA, 185–189. https://doi.org/10.1109/CCST.2010.5678709

[16] Kopaczka, M., Nestler, J., & Merhof, D. (2017). Face Detection in Thermal Infrared Images: A Comparison of Algorithm- and Machine Learning-Based Approaches. *Proceedings of International Conference on Advanced Concepts for Intelligent Vision Systems*, Belgium, 518–529. https://doi.org/10.1007/978-3-319-70353-4_44

[17] Ma, C., Thanh, T. N., Uchiyama, H., Nagahara, H., Shimada, A., & Taniguchi, R. I. (2017). Adapting Local Features for Face Detection in Thermal Image, Sensors, *17*(12). https://doi.org/10.3390/s17122741

[18] Panasiuk, J., Prusaczyk, P., Grudzień, A., & Kowalski, M., (2020). High-resolution thermal face dataset for face and expression recognition. *Metrology and Measurement Systems*, *27*(3), 399–415. https://doi.org/10.24425/mms.2020.134591

[19] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition*, USA, 886–893. https://doi.org/10.1109/CVPR.2005.177

[20] Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623

[21] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, USA, 658–666. https://doi.org/10.1109/CVPR.2019.00075

[22] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, USA, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[23] Szegedy, Ch., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, USA, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[25] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*(2), 84–90. https://doi.org/10.1145/3065386

[26] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of 3rd International Conference on Learning Representations*, USA. https://arxiv.org/abs/1409.1556

[27] Redmon, J. & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. https://arxiv.org/abs/1804.02767

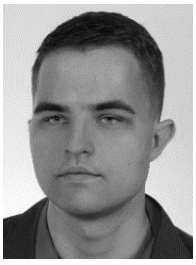
[28] Kowalski, M., Grudzień A. (2018). High-resolution thermal face dataset for face and expression recognition. *Metrology and Measurement Systems*, *25*(2), 403–415. https://doi.org/10.24425/119566

[29] Sequeira, A. F., Chen, L., Ferryman, J., Galdi, C., Chiesa, V., Dugelay, J. L., Maik, P., Gmitrowicz, P., Szklarski, Ł., Prommegger, B., Kauba, C., Kirchgasser, S., Uhl, A., Grudzień, A., & Kowalski, M. (2018). PROTECT Multimodal DB: a multimodal biometrics dataset envisaging Border Control. *Proceedings of International Conference of the Biometrics Special Interest Group*, Germany, 1–5. https://doi.org/10.23919/BIOSIG.2018.8552926

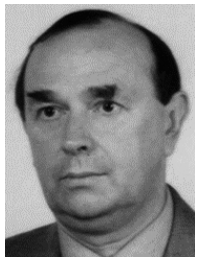
[30] Espinosa-Duró, V., Faundez-Zanuy, M., & Mekyska, J. (2013). A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums. *Cognitive Computation*, *5*, 119–135. https://doi.org/10.1007/s12559-012-9163-2

[31] Espinosa-Duró, V., Faundez-Zanuy, M., Mekyska J., Monte-Moreno, E. (2010). A Criterion for Analysis of Different Sensor Combinations with an Application to Face Biometrics. *Cognitive Computation*, *2*, 135–141. https://doi.org/10.1007/s12559-010-9060-5

Marcin Ł. Kowalski received the PhD degree in optoelectronics from the Military University of Technology, Warsaw, Poland in 2014. Currently he is working as an assistant professor in the Institute of Optoelectronics at the Military University of Technology where he leads the biometric research group. His research efforts focus on computer vision, in particular biometrics and multispectral imaging.



Artur Grudzień received the M.Sc. (2016) degree from the Military University of Technology (MUT) in Warsaw, Poland. His research activities focus on biometrics, face recognition and machine learning.



Wiesław M. Ciurapiński received the PhD degree in optoelectronics from the Military University of Technology, Warsaw, Poland in 1985. Currently he is working as Assistant Professor in the Institute of Optoelectronics at the Military University of Technology where is Head of the Optoelectronics Systems Division. His research efforts focus on fiber optics and biometrics.