Postprint of: Kwaśniewska A., Rumiński J., Szankin M., Kaczmarek M.: Super-resolved Thermal Imagery for High-accuracy Facial Areas Detection and Analysis ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE Vol. 87, (2020), pp. 1-11 DOI: 10.1016/j.engappai.2019.103263

Super-resolved Thermal Imagery for High-accuracy Facial Areas Detection and Analysis

Alicja Kwasniewska[a,b,*], Jacek Ruminski[a], Maciej Szankin[b], Mariusz Kaczmarek[a]

[a]*Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Department of Biomedical Engineering, Narutowicza 11-12, Gdansk, 80-233 Poland*
[b]*Intel Corporation, Artificial Intelligence Products Group, AI Applications 12220 Scripps Summit Dr, San Diego, CA 92131, USA*

Abstract

In this study, we evaluate various Convolutional Neural Networks based Super-Resolution (SR) models to improve facial areas detection in thermal images. In particular, we analyze the influence of selected spatiotemporal properties of thermal image sequences on detection accuracy. For this purpose, a thermal face database was acquired for 40 volunteers. Contrary to most of existing thermal databases of faces, we publish our dataset in a raw, original format (14-bit depth) to preserve all important details. In our experiments, we utilize two metrics usually used for image enhancement evaluation: Peak-Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM). In addition, we present how to design a SR network with a widened receptive field to mitigate the problem of contextual information being spread over larger image regions due to the heat flow in thermal images. Finally, we determine whether there is a relation between achieved PSNR and accuracy of facial areas detection that can be analyzed for vital signs extraction (e.g. nostril region). The performed evaluation showed that PSNR can be improved even by 60% if full bit depth resolution data is used instead of 8 bits. Also, we showed that the application of image enhancement solution is necessary for low resolution images to achieve a satisfactory accuracy of object detection.

Keywords: super resolution, deep learning, thermal imagery, object detection

*Corresponding author
Email addresses: alicja.kwasniewska@pg.edu.pl, alicja.kwasniewska@intel.com (Alicja Kwasniewska), jacek.ruminski@pg.edu.pl (Jacek Ruminski), maciej.szankin@intel.com (Maciej Szankin), mariusz.kaczmarek@pg.edu.pl (Mariusz Kaczmarek)

1. Introduction

High resolution (HR) image restoration from corresponding low resolution (LR) data is known as image super resolution (SR). Specifically, if a single image is used for the enhancement, the approach is called single image super resolution (SISR). Inherently, this problem is ill-posed since it's possible to recover various HR outputs for a single LR input. Such inverse problem is usually solved by utilizing the prior knowledge. The prior knowledge can be learned by predicting a pixel value with interpolation methods, e.g. bicubic interpolation [1], edge-guided interpolation [2], or adaptive non local sparsity-based modeling [3]. Another ways to acquire the knowledge is to exploit the internal structure of pixels within the same LR image [4][5][6], or learn it from corresponding pairs of LR and HR examples, i.e. example-based algorithms [7][8][9][10][11][12][13][14][15]. The first group, known as interpolation-based SISR [16], often intend to mitigate a down-sampling process only. Also, the interpolation techniques are based on generic smoothness priors and therefore are indiscriminate, as they smooth both edges and object parts, what leads to the blurring effect [7].

Hence, due to limited applicability of interpolation approaches, the learning-based methods are becoming more popular and are being further investigated, e.g. by combining learning based gradient regularization with reconstruction approach that aims at preserving consistency between HR and LR images, while satisfying the prior knowledge [17]. In particular, deep learning based SR algorithms, which allow to establish the mapping between LR and HR patches of the image using a stack of convolutional operations have recently become state-of-the-art solutions. Majority of the conducted work considered visible light images only. The pioneer research of applying deep learning to SISR problem for RGB data, conducted by Jain and Seung [18], aimed at image denoising. In later studies, stacked collaborative local auto-encoders were proved to be successful for a low resolution RGB images up-scaling [6]. At each stacked layer, high frequency components are enhanced using similarity search applied to the input LR image, that are then fed to auto-encoders in order to suppress the noise and take into account the correspondence of the overlapping reconstructed patches. To overcome the disadvantage of independent optimization of the similarity search and the auto-encoder

for each layer, as well as the absence of steps other than the learning part in the framework pipeline, Dong et al. proposed to formulate both mapping and feature extraction as convolutional operations [11]. As a result, proposed SR pipeline, called SR Convolutional Neural Network (SRCNN), can be fully obtained through end to end mapping. Additional modifications introduced to the SRCNN allowed to further improve the Peak Signal-to-Noise Ratio (PSNR) index, a metric usually used for quantitative evaluation of super-resolution algorithms. Kim et al. at first introduced Very Deep Super Resolution (VDSR) model with residual connections that correlate LR input with HR output, outperforming SRCNN by 0.87dB PSNR [12] on RGB dataset called Set5 [9] downscaled by a factor of 2. Later same authors proposed Deeply Recursive Convolutional Network (DRCN) [13], which incorporates recursive supervision to 1) increase the depth while keeping number of parameters constant; 2) eliminate the vanishing gradient problem. Afterwards, Tai et al. [14] further enhanced the efficiency with Deep Recursive Residual Network (DRRN) that adopts weight-shared residual connections both in global and local manner to increase the network depth, achieving PSNR 1.08dB higher than SRCNN on RGB data. Generative Adversarial Network (GAN) based solutions have been also already applied to visible light image enhancement, allowing for successful restoration of high frequency details (e.g. SRGAN [19] or EnhanceNet [20]).

HR data is especially desired for medical applications in order to make proper diagnosis, as super-resolved image can usually offer more diagnostically important details [21] [22]. Health care sectors that use imaging techniques can utilize resolution enhancement, e.g. for computer-aided diagnosis (CAD) of breast tumors to improve accuracy of malignancy classification [23] or reconstruct computed tomography images to provide clinicians with important details to make correct decisions [24]. Undoubtedly, providing more detailed HR samples can ease the analysis of medical imaging. The interesting research question is whether it can also help with improving accuracy of remote medical diagnostics. Due to global aging, the medicine is expected to deliver novel solutions that allow for performing basic diagnostic and monitoring tests at home [25]. Some studies have already proved that basic vital signs can be estimated from a single camera stream in a non-contact way, e.g. heart rate from visible light sequences

[26] or a breathing rate from thermal data [27]. The starting point of image processing-based vital signs estimation is accurate detection of proper facial regions. Despite the majority of object detection research focuses on visible light spectrum, some attempts to localize facial features from low resolution thermal imagery have been done [28][29], recently also using Deep Learning (DL) techniques [30]. The achieved results, though, were not satisfactory, giving accuracy of 0.53 ± 0.15 for eyes area and 0.60 ± 0.18 for nostrils, expressed as Intersection over Union (IoU) metric. Localizing a proper region is crucial for the robustness of signal processing algorithms aimed at estimating vital signs [27]. It has already been shown that motion magnification can improve performance of heart rate estimation at distances above 6 meters [31]. However, to the best of our knowledge, it hasn't been evaluated yet whether image enhancement with DL-based SR has a positive effect on the accuracy of object detectors, especially in thermal imaging. Simultaneously, analysis of other than visible light image domains is crucial, as representation of features may differ across them, e.g. thermal images are characterized by blurring and lower contrast between adjacent regions due to the heat flow, hence networks designed for extracting high frequency components (e.g. edges, lines) from visible light spectrum images may not be sufficient in the thermal domain.

The contribution of our work is threefold: 1) First, we collect and publish a dataset of raw thermal sequences of a face in the original full 14-bit precision. 2) Second, we applied deep learning based super resolution algorithms to thermal image sequences. We experimentally compared the state of the art algorithms and our algorithm analyzing selected spatiotemporal properties of thermal sequences including temporal frames averaging and the influence of various bit depths. 3) Finally, we evaluate the effect of enhancing image resolution (and related PSNR/SSIM measures) on the robustness of areas detection in the auditing thermal domain.

Since in our research we focus on facial features detection, we analyse PSNR changes for both extracted facial areas and images as a whole. In this way, we aim at determining whether PSNR is sensitive to the change of pixel values caused be the presence of breathing patterns in the extracted areas (e.g. nostrils). In addition, we evaluate how PSNR changes by a) using averaging operation of subsequent frames in a sequence, what allows for reducing random noise; b) utilizing 16-bit resolution data in

order to preserve important image components that may be invisible after conversion to lower bit resolution (e.g. 8-bit). To the best of our knowledge, the majority of existing publicly available face thermal datasets contain images that were converted to lower precision image format with loss of some of the information available in the higher precision.

Thus, the data collected by us possesses advantages over them, as it is shared in the raw format that contains unprocessed pixels, what is potentially useful for extracting intensity changes, caused e.g. by breathing. Collected dataset [1] and supplementary materials [2] are publicly available.

The rest of the paper is organized as follows: Section II introduces the problem statement, including acquired thermal data characteristics. In Section III we describe the experimental methodology used to evaluate the influence of super-resolution methods on facial features detection accuracy. Section IV overviews achieved preliminary results, further discussed in Section V. Finally, we conclude our work in Section VI.

2. Problem statement and preliminaries

2.1. Data characteristics, collection and processing

In this study, we are focusing on evaluating face hallucination algorithms on images acquired in thermal spectrum, i.e. intensities of electromagnetic radiation in the range of 8-12um (LWIR, Long-Wave Infrared). The intensity is represented as a sequence of arrays. Each array contains digital values of a bit resolution higher than 8, typically 14 bits. Radiation values may be converted to temperature data in order to form a final thermal image using data pre-processing algorithms (e.g., radiation to temperature mapping, data range selection) and a Color Look Up Table (CLUT) to assign colors or shades of gray to the digital values.

When 8-bit color models are used for intensities with higher bit resolution, the conversion from electromagnetic radiation to color palette values is lossy, so the contrast

[1] Database of thermal image sequences can be obtained by sending an e-mail request to {alicja.kwasniewska/jacek.ruminski}@pg.edu.pl

[2] Supplementary materials available at https://github.com/akwasnie/Super-Resolved-Thermal-Imagery

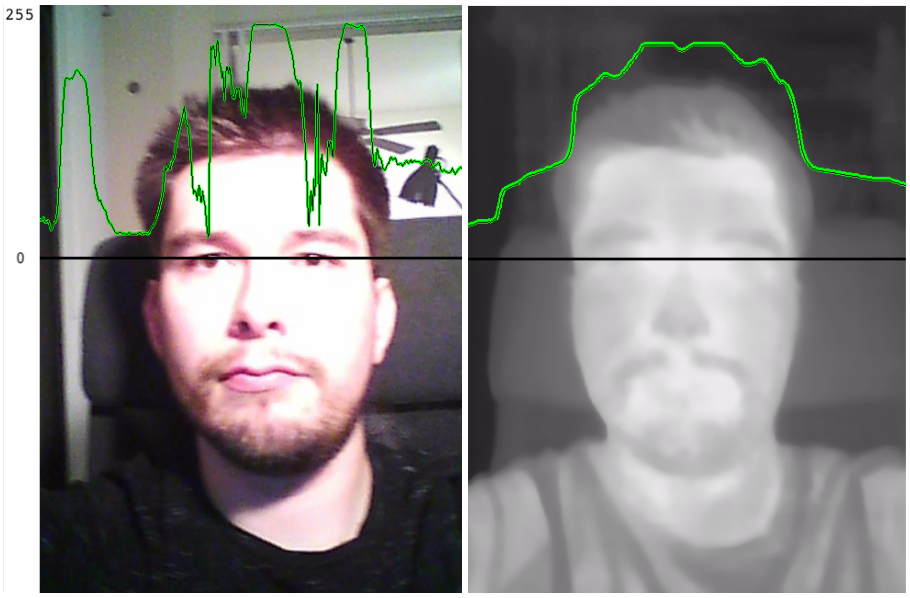

Figure 1: Color values (green plot) of pixels at the eyes level marked with a black line in RGB (left) and thermal (right) image collected with a single device FLIR One equipped with the VGA Visible Camera and the 120x80 IR Sensor

between regions may be reduced, eliminating some important details. Another important factor represented in thermal images is related to the heat transfer in objects of different temperature. When heat is transferred from one object to another, the temperatures equalize, resulting in lower gradient and smoother color change between pixels within adjacent areas (see Fig. 1). For the thermal image we can observe smoother color change between pixels within adjacent area. These characteristics of thermal data should be carefully considered while applying image processing algorithms. Typically, CNNs are based on high frequency features (e.g. edges, corners, lines), and trained on visible light images to extract them. In visible spectrum edges between different areas are clearly distinguishable, therefore filters based on high frequency patterns lead to very good recognition results [32]. Smooth representation of areas in thermography lead to worse detection accuracy while using high frequency components [30]. This problem can be partially mitigated by using higher resolution thermal cameras but then the cost of the solution rapidly grows, what makes it unsuitable for home based

health care devices. That's why, in most cases low resolution data is more common. Although the cost of thermal imaging hardware is still significantly higher than the price of comparable visual light cameras, the price of thermal cameras is continuously decreasing [33]. Recently, thermal cameras have become more affordable and thus commercially available (e.g. FLIR Lepton camera modules <$175 [34]), what enabled various practical applications in different industries, e.g. in remote medical diagnostics (non-contact vital signs estimation [35]), or in studies on autonomous vehicles (detection of people on roads [36] and classification of challenging road conditions [37]). We believe that utilization of hallucinated images can further improve the performance of proposed solutions by improving distinguishability of the object parts, especially in the case of thermal images with small spatial resolution (e.g. 80x60, 160x120) and smaller temperature resolutions.

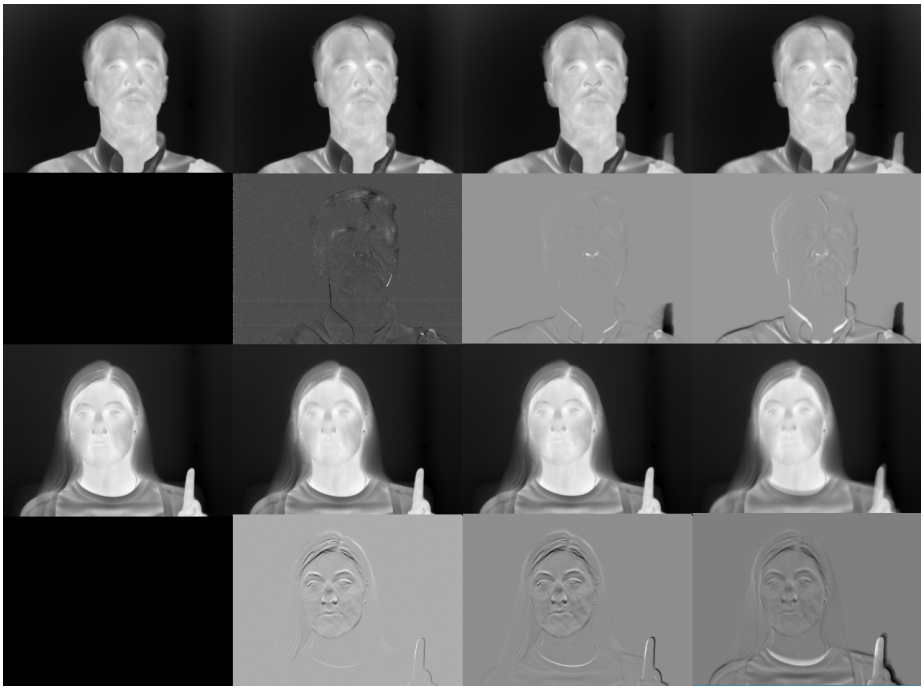

Figure 2: Examples of thermal images for volunteer 1 and 2 breathing through nose)

Experiments were carried out on a thermal face dataset collected by us from 40 users (19 male, 21 female, age: 34.11±12). The thermal camera FLIR SC3000,

used for data acquisition, (320x240 spatial resolution, 30 frames per second (FPS), temperature measurement range from -20°C to +80°C, 20° lens, measurements using noise reduction mode) was placed on a tripod, 112 centimeters from the ground and 120 centimeters from the volunteer's head. During data collection, volunteers were asked to perform 5 tasks:

1. breathing trough a nose for 2 minutes
2. breathing through a mouth for 1 minute
3. turning a head very slowly from a left to a right side during 1 minute, so that after 30 seconds the volunteer was looking straight
4. turning a head very slowly from a top to a bottom during 1 minute, so that after 30 seconds the volunteer was looking straight
5. mimicking 3 emotions, each for 20 seconds (happiness, anger, surprise)

Additionally, in scenario 1 and 2, participants were asked to point their finger upward during inhaling and down during exhaling, so that we were able to calculate the reference breathing per minute (BPM) value, by counting the number of finger movements. At the same time, we also collected the ambient temperature inside the laboratory room, where experiments were conducted. As a result, we created a relatively big thermal face dataset (240 minutes of thermal sequences, 78.14 GB), that can be used for training deep neural networks, since the amount of data that we feed to models is crucial for them to succeed in producing accurate predictions on new samples [25]. It is very important to note that the created dataset consists of original raw 14-bit image sequences. We haven't found any other dataset that contains images saved in original resolution before lossy conversion to other image formats. The entire set consists of 40 2-minute sequences and 160 1-minute sequences. All sequences were recorded with 30 FPS frequency rate, so after extracting all frames from them, we got 7200 images in PNG file format with 16-bit depth per channel and 7200 images in PNG file format with 8-bit depth per channel. Each image is 320 pixels wide and 240 pixels high.

In this study, we utilize data that allow for calculating reference BPM (scenario 1 and 2), as we focus on improving facial feature detection, used for remote estimation of breathing rate. From all these sequences, we extracted every 300^{th} data frame to

ensure proper variability of facial data in the compact training dataset. Radiation data acquired by us in this study is represented as a sequence of arrays, containing digital values of a 14-bit resolution. In order to feed data to convolutional-based deep learning model, it is necessary to convert collected raw data to image formats. Therefore, the minimum and maximum values (other than 0) were identified in the selected data frames. These values were used in linear scaling of original, 14-bit radiometric data to 8 and 16-bit grayscale images in the PNG file format. Facial regions detection is typically based on features of a single frame. Since collected dataset contains thermal sequences instead of single images, utilization of temporal information can be potentially beneficial to improve desired detection results. For example, temporal average can potentially remove random noise (e.g. related to image sensor noise, influence of local environment, etc.) keeping image features important for better detection of facial regions. On the other hand, it is important to note that averaging operation is possibly favorable on the assumption that the subject stands still in front of the camera. In case of the movement occurrence, the edges of object areas will become blurred after applying the averaging operation. Our data collection process assumes that the volunteer stands still, yet some small movements may still be present. Thus, in our research we evaluate how averaging operator influence image quality enhancement (PSNR) and the object detection accuracy (IoU). For this, the average of W subsequent frames is calculated, where W denotes the size of a window, expressed as a number of neighbouring frames *(-W/2;W/2)*, used for calculating the average frame. We chose window sizes as 7 (relatively small window insensitive to body movements and respiratory events), 30 (1 image for every 1-second intervals), 90 (1 image for every respiratory event, assuming the respiration rate for an adult is ~12 while resting, every 2-3 second we can observe the inhalation/exhalation event). The used set consisted of 1296 single 8-bit frames, 1296 single 16-bit frames, and 1296 images for each of the window sizes: *W=7, W=30, W=90* both in 8 and 16-bit resolution, named *single-{8/16}*, *avg{7/30/90}-{8/16}* respectively. In total, 10368 thermal images are utilized. For all SR networks applied in this study, we used LR images generated by dowscaling and upscaling original HR images with a scale of 2. LR images were used to train and evaluate SR topologies by comparing enhanced results against original HR data. Examples of thermal images

and differences between calculated average frame and the middle frame in each window for two different volunteers breathing through nose are presented in Fig. 2. Odd rows present images generated using various window sizes for average operation, even rows difference between calculated average frame and the middle frame in the window (from the left: 1, 7, 30, 90 window size).

2.2. Formulation of the research statement

The goal of SISR is to estimate the HR output $\widehat{Y}$ from the corresponding LR image X, created by downscaling the ground truth (GT) sample Y with a scale s. SR network S defined by parameters θ has to find the HR output $\widehat{Y}$ as close to Y, as possible:

$$min(L_\theta(Y, \widehat{Y})), \quad \text{where} \quad \widehat{Y} = S_\theta(X) \tag{1}$$

L_θ is the cost function used to optimize the model.

In general, network S_θ realizes 3 tasks: feature extraction F_{fe}, non-linear mapping F_{nlm} and reconstruction F_{rec}. All operation can be formed by convolution operations ($*$), and defined as:

$$\widehat{Y} = F_{rec}(F_{nlm}(F_{fe}(X))) = W_{rec} * (\sigma(W_{nlm} * (\sigma(W_{fe} * X + B_{fe})) + B_{nlm})) + B_{rec} \tag{2}$$

where $\{W_{rec}, W_{nlm}, W_{fe}, B_{fe}, B_{nlm}, B_{rec} = \theta\}$ are weights (W) are biases (B) matrices, respectively for each of the network task:fe, nlm, rec and σ is the activation function. The goal is to optimize network parameters θ, so that end-to-end mapping S_θ accurately predicts $\widehat{Y} = S_\theta(X)$. In the supervised setting, that we utilize in this study, the relation between reconstructed HR image $\widehat{Y}$ and the ground truth image Y formulates the cost function, that is used for parameters optimization. In SISR, the commonly used cost function is Mean Squared Error (MSE), averaged across N training samples:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \|Y_i - S_\theta(X_i)\|^2 \tag{3}$$

Although lower MSE favors higher Peak Signal to Noise Ratio (PSNR), it has been observed that satisfactory performance can be also obtained by using other evaluation metrics, e.g. Structural Similarity Index (SSIM) [11].

Please note, that equation (2) is the general form of CNN-based SISR, that contains only one convolutional layer at each network step, i.e. W_{fe}, W_{nlm}, W_{rec} correspond to $n_{(fe/nlm/rec)}$ filters of a size $w_{(fe/nlm/rec)} \times h_{(fe/nlm/rec)} \times c_{(fe/nlm/rec)}$ (*w*-width, *h*-height, *c*-input channels). Yet, as previous studies showed, the deeper the network, the better performance can be achieved [13][14]. Thus, in this work, we utilize the novel version of CNN based SR network, designed specifically for thermal data. The network architecture is explained in details in Section 3.

3. Methods

3.1. Object detection

Previous attempts to facial features detection from low resolution thermal images using Artificial Neural Networks (ANNs) have shown that achieved accuracy is limited (0.32 0.38, 0.55 0.42 for eyes and nostril areas respectively, expressed as Intersection over Union (IoU)) [30]. Thus, in our study, we evaluate the effect of applying SR algorithms on object detection accuracy using a relatively large dataset of thermal images. Specifically, the proposed SR network and other state-of-the-art SR models are used to generate hallucinated images of a face, that are then used for facial features detection training. Accuracy is measured using IoU metric and compared across various SR algorithms, as well as original HR data, and LR samples generated via bicubic interpolation.

To evaluate the influence of applying SR on the detection task, we utilize Inception based Single Shot Detector (SSD) model [38], the same network as in [30]. SSD is a simple relative to other deep neural networks, as it does not require generation of object proposals during the run-time. Instead, it creates a set of default boxes over different aspect ratios and scales during training. Then, predictions are performed by assigning scores for the presence of each object category in default boxes. This makes SSD a real time solution that can be easily trained for a new task.

Although SSD has more lightweight feature extractor than faster R-CNN, the fast processing time is not indispensable in our case, as the person is supposed to sit still for 10-15 seconds in order to evaluate vital signs at the distance [31].

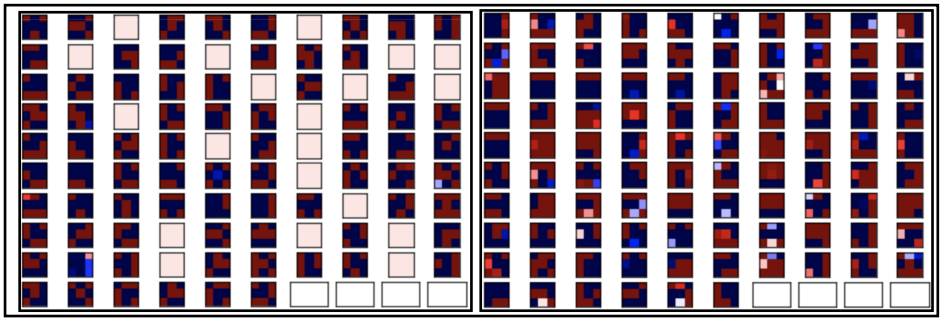

Figure 3: Examples of filters learnt by the proposed SR CNN network. On the left 96 filters learnt if a single convolution is applied for feature extraction. On the right 96 filters of a convolution after applying 3 residual blocks

3.2. Super resolution

3.2.1. Proposed network architecture

The wider receptive field in the feature extraction step is potentially beneficial for mitigating the problem of lower contrast of adjacent regions in thermal imagery. As presented in Fig. 3, we can observe that after applying a set of convolutions filters learnt by a network represent more complex features that filters in the a single convolution, what is crucial for reconstructing details in the SR task. The drawback of deeper networks, though, is increased number of parameters, what leads to huge models sizes and more difficult optimization process [14]. Therefore, we propose to apply residual mappings [39] constructed from following operations: $conv1_{fe/nlm}$, $batch\,norm$, activation function σ (in our case Rectified Linear Unit (ReLU) [40]), $conv2_{fe/nlm}$, $batch\ norm$ to both feature extraction and non-linear mapping steps. Weights of $conv1_{fe/nlm}$ and $conv2_{fe/nlm}$ are shared across all residual blocks within each network step, i.e. fe and nlm respectively, to avoid the increased number of parameters. After each residual block, the addition operation $\oplus$ sums up the $shortcut\,connection$ $conv0_{fe/nlm}$ (i.e. input of the residual block that is skipping the convolutional and batch norm layers) with the output from this block. Operation $\oplus$ is followed by activation σ. Output from the feature extraction sub-network F_{fe} after *e-th* residual block can be defined as:

$$F_{fe}^{(e)} = \begin{cases} g_{fe}(I_{fe}, W_{fe}) + I_{fe},\ e = 1 \\ g_{fe}(F_{fe}^{(e-1)}, W_{fe}) + I_{fe},\ e \in (1, E\rangle \end{cases} \tag{4}$$

where g_{fe} represents residual mapping $g_{fe}(x, W_{fe}) = W_{conv2_{fe}}\sigma(W_{conv1_{fe}} * x)$ to be learnt, and $I_{fe} = W0_{fe} * X$ is the LR input convoluted with the first weights matrix $W0_{fe}$. *E* is the total number of residual blocks used for *feature extraction*. To simplify the mathematical formulation biases were skipped.

Following [13], we also introduce recursive supervision to the non-linear mapping sub-network in order to eliminate vanishing gradient problem. Let's define an input to the recursive block d, as $I_{rec}^{(d)}$:

$$I_{rec}^{(d)} = \begin{cases} W0_{nlm}^{(d)} * F_{fe}^{(e=E)}, d = 1 \\ W0_{nlm}^{(d)} * F_{nlm}^{(d-1)}, d \in (1, D\rangle \end{cases} \tag{5}$$

where D denotes total number of recursions, $W0_{nlm}^{(d)}$ is the weight matrix of the first convolution in the *d-th* recursive block, and $F_{nlm}^{(d-1)}$ is the output of the previous (i.e. *d-1*) recursive block in the non-linear mapping sub-network. Since residual blocks are also introduced to recursive blocks, we can define them as:

$$F_{nlm}^{(d)} = B_{res}^{(U)} = \begin{cases} g_{nlm}(I_{rec}^{(d)}, W_{nlm}) + I_{rec}^{(d)}, u = 1 \\ g_{nlm}(B_{res}^{(u-1)}, W_{nlm}) + I_{rec}^{(d)}, u \in (1, U\rangle \end{cases} \tag{6}$$

where *U* is the number of residual blocks in the recursion *d*, and g_{nlm} represents residual mapping $g_{nlm}(x, W_{nlm}) = W_{conv2_{nlm}}\sigma(W_{conv1_{nlm}} * x)$ to be learnt. The output $F_{nlm}^{(d)}$ from *d-th* recursion is simultaneously the output from the last (*U-th*) residual block $B_{res}^{(U)}$ within this recursion. All *D* outputs from the *nlm* sub-network are weighted in the reconstruction sub-network F_{rec} to produce the final output. Also, additional *identity mapping* is added in order to correlate LR input (X) with restored HR data and, in this way, preserve detailed image components.

$$F_{rec} = \sum_{d=1}^{D} w^{(d)}(F_{nlm}^{(d)} + X) \tag{7}$$

Then, taking into account eq. 3, similarly to [13], the cost function can be defined as:

$$L(\theta) = \frac{1}{N}\frac{1}{D}\sum_{i=1}^{N}\sum_{d=1}^{D}\left\| Y_i - \sum_{d=1}^{D} w^{(d)}(F_{nlm}^{(d)} + X)\right\|^2 \tag{8}$$

3.2.2. Reference SR solutions

In this subsection, we specify the differences between the proposed network architecture and state of the art algorithms. The pioneer research of applying CNN to SR

task (model knows as SRCNN) was conducted by Dong et. al [11]. Yet, it had been soon confirmed that deeper representation can lead to better results, e.g. in DRCN [13] and DRRN [14] models. Motivated by these findings, we base our idea on these solutions, but introduce some additional steps to better fit thermal data. Similarly to DRCN, we use recursive supervision to reduce the risk of overfitting, while increasing the depth of the network. Besides this operation, we propose to further deepen the model and, in this way, gain more accuracy by applying residual blocks, proved to ease the optimization process [39]. This approach has been previously utilized by DRRN, yet, there are some important differences to be noted. First of all, best DRRN results were achieved for the configuration *B1U25*, where *B* denotes number of recursions and *U* number of residual blocks within each recursion. It can be easily observed, that in fact recursions were not applied in this setup, as *B=1*. In our network we utilize recursive approach. Secondly, we propose to widen the receptive field in the feature extraction sub-network to mitigate the problem of blurring and bigger distances between interesting components in thermography. Specifically, we utilize residual blocks in both feature extraction and non linear mapping steps, contrary to DRRN, which uses them only in the mapping part. In this way, we design the model that fits other image domain, apart from visible light, which most of the networks are designed and tested for. Last but not least, we use shared weights at each network step, i.e. for feature extraction we use only two weights matrices $W_{conv2_{fe}}, W_{conv1_{fe}}$ shared across all *E* residual blocks, similarly, for the non linear mapping, we use $W_{conv2_{nlm}}, W_{conv1_{nlm}}$ shared across all *U* residual and all *D* recursive blocks, opposed to DRRN, where weights are shared across residual blocks but are unique across recursions. Thus, comparing to DRRN, the number of parameters is reduced by *2D* times in the non-linear sub-network. Since we utilize residual blocks in both feature extraction (embedding) and non-linear mapping sub-networks, our model is called DRESNet - Deep Residual Embedding and Supervised-recursion. The proposed model architecture and recent state-of-the-art models are presented in Fig. 4. Weights used in each convolution or a block are labeled with a capital letter W. Weights that are shared across operations/blocks are marked with the same background color. Thus, we may easily note that weights in recursions in DRRN are not shared, while in DRESNet they are, except

one convolutional block applied before all following residual blocks.

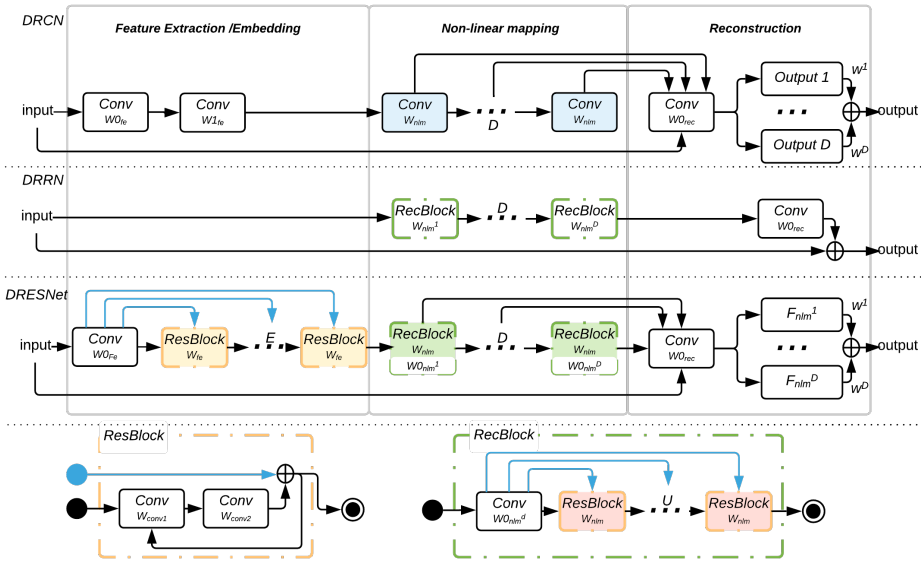

Figure 4: Architecture of the proposed DRESNet model and comparison with state-of-the-art networks

3.3. Training

Training is carried on all created sets separately (i.e. *single*-{*8/16*}, *window*{*7/30/90*}-{*8/16*}). At first each of image sets is divided into training, test and validation parts using a 70:15:15 split. Training and validation sets are used to optimize SR models.

Next, test sub-sets are fed into the trained SR models in order to generate hallucinated face images. After this step, we have 24 HR sub-sets, 8 for each of the SR models. Within each model 4 for each of the bit resolution (8/16 bits): *avg*{*7/30/90*} and *single*. Generated subsets are then again split into training, test and validation parts in a 70:15:15 proportion. Object detection network *(SSD)* is trained using each of these subsets. In addition, we also use bicubic data and original HR data to train the models and compare achieved results with results computed for super-resolved object detectors. As a result, 40 object detection models are created. The nomenclature is as follows: {*object detection*}-{*data source*}-{*data pre-processing*}-{*resolution*}: {*SSD*}-{*bicubic/orig/DRCN/DRRN/DRESNet*}-{*avg*{*7/30/90*}/*single*}-{*8/16*}.

In order to find the most optimal DRESNet architecture, various configurations of the proposed model are tested. Number of residual blocks in feature extraction sub-network (E), number of recursions (D), and number of residual blocks within each recursion (U) are randomly chosen from the range (1-10). For each configuration, training is performed using the same set of hyperparameters. Each convolutional layer contains of 96 3x3 filters with weights initialized using He algorithm [41]. Following [12], training data is cropped to 41x41 patches with a stride of 21. The model is optimized using back-propagation with the cost function defined by eq. 8, minimized using Adam optimizer [42], momentum 0.9, and weight decay 0.0001. Initial learning rate is set to 10^{-2} and then we reduce it by an order of magnitude after each 5 subsequent epochs, for which the decrease of the validation error is not observed. After evaluating all configurations, we found that the best performing network in terms of PSNR had 3 residual blocks in the feature extraction sub-network (*E=3*) and 9 recursions (*D=9*) in the non-linear mapping sub-network. Both residual and recursive blocks are described in details in Section 3.2.1. Contrary to DRRN, it turned out that residual blocks in recursions don't produce better results. Instead it's better to place them before recursions. The final architecture of the introduced SR CNN is presented in Fig. 5. Selected model has 3 residual blocks with shared weights in the feature extraction part and 9 recursions in the non-linear mapping part. The final output is constructed as the weighted sum of all recursions. This configuration is used in all further experiments and we thereafter refer to it as DRESNet.

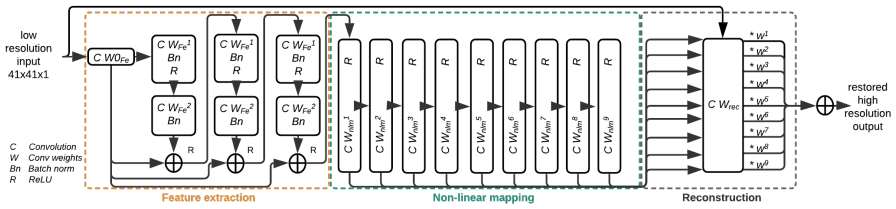

Figure 5: The best configuration of the proposed DRESNet in terms of PSNR

DRCN and DRRN are trained with hyperparameters suggested by their authors [13][14] using TensorFlow implementation [43] mentioned as the alternative code in the original DRRN repository [44]. This implementation was DRRN configuration

uses 9 residual blocks and 1 recursion, while for DRCN 16 recursions are applied. For a fair comparison with DRESNet, number of filters in both models was set to 96. Our motivation is based on results achieved by SRCNN [11], which proved that better performance is achieved by increasing the number of filters in convolutional layers. Taking it into account, using different number of filters may affect results, leading to false conclusions about the architecture itself, i.e. placement of recursions, residuals etc. To avoid results being biased by different number of filters, we decided to use the same filter width for all SR networks.

Since the number of super-resolved data used for training object detection model is limited, we utilize transfer learning technique [45] to tune publicly available checkpoint [46] on our thermal dataset. The random search approach [47] was used to find the best training configuration. After that, the same hyperparameters were applied to SSD object detection, i.e. training steps 40000, batch size 32, initial learning rate 0.004, learning rate decay steps 5000, decay factor 0.95.

4. Results

Calculated Peak Signal-to-Noise Ratio and Structural Similarity Index (for each marked region and the frame as a whole) for 8 and 16-bit images are collected in Table 1 and 2, respectively. We compare results achieved for images enhanced with various SR algorithms and resized with bicubic interpolation both for extracted single images and images calculated as the average of 7, 30, and 90 subsequent frames. Table 3 presents Intersection over Union between regions marked manually (ground-truth) and regions detected by object detection models (SSD) trained on images with improved and decreased resolution and evaluated on test sets corresponding to the applied enhancement/degradation algorithm, i.e. model trained on bicubic data was evaluated on bicubic data, etc. Relation between IoU metric (average for all detected regions) and PSNR (average for all marked facial areas) is presented in Fig. 6.

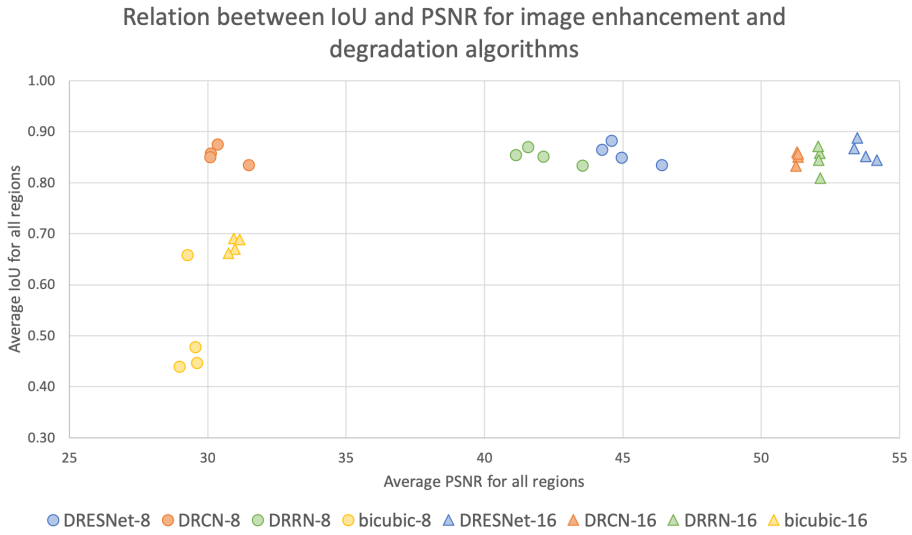

Figure 6: Relation between average IoU calculated for all detected regions and average PSNR calculated for annotated ground-truth facial areas (resolution enhanced with SR models or decreased with bicubic interpolation)

Table 1: Peak Signal-to-Noise Ratio [dB] (for each manually marked region and the frame as a whole) for 8 and 16-bit LR images, generated by extracting frames from raw thermal sequences and then downscaling and upscaling them with a scale of 2 using bicubic interpolation. Generated LR images were then enhanced with DRCN, DRRN or DRESNet (our) SR models. blue - first best for each region and within each input category (single; window 7, 30, 90) separately, **blue** - first best across all input categories

	single-8				single-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.90 ±0.10	29.86 ±1.84	41.01 ±1.81	43.87 ±1.58	27.90 ±0.11	51.29 ±0.11	52.12 ±0.34	53.06 ±0.39
face	27.92 ±0.10	30.28 ±1.86	40.73 ±1.65	44.20 ±1.78	27.90 ±0.05	51.31 ±0.09	52.11 ±0.53	53.88 ±0.56
nose	27.93 ±0.21	30.36 ±1.52	41.72 ±1.55	44.98 ±1.86	27.89 ±0.14	51.28 ±0.13	52.13 ±0.30	53.38 ±0.63
frame	27.91 ±0.16	31.49 ±2.37	43.07 ±1.06	47.49 ±1.28	27.90 ±0.11	51.29 ±0.11	52.12 ±0.39	53.36 ±0.61
	avg7-8				avg7-16			

region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.88	30.14	41.94	44.72	27.89	51.33	52.12	53.39
	±0.10	±2.73	±1.97	±1.70	±0.10	±0.26	±0.37	±0.46
face	27.90	30.66	41.22	44.33	27.90	51.34	51.95	53.98
	±0.03	±2.30	±1.50	±1.64	±0.02	±0.16	±0.47	±0.48
nose	27.89	30.36	41.28	44.57	27.91	51.32	52.03	53.11
	±0.14	±2.54	±1.71	±1.62	±0.15	±0.24	±0.26	±0.51
frame	27.89	31.15	43.18	47.45	27.90	51.33	52.05	53.47
	±0.10	±2.67	±1.01	±1.21	±0.11	±0.23	±0.38	±0.58
	avg30-8				avg30-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.91	29.44	42.11	44.85	27.90	51.33	52.12	53.44
	±0.11	±2.47	±2.16	±2.00	±0.02	±0.14	±0.32	±0.63
face	27.90	30.59	41.68	44.49	27.90	51.35	51.95	54.28
	±0.02	±2.36	±1.72	±1.73	±0.02	±0.12	±0.39	±0.60
nose	27.89	30.78	42.59	45.54	27.91	51.31	52.09	53.90
	±0.13	±2.93	±1.65	±1.93	±0.11	±0.14	±0.30	±0.77
frame	27.90	31.51	43.76	47.69	27.90	51.33	52.07	53.78
	±0.10	±2.92	±1.18	±1.28	±0.09	±0.14	±0.34	±0.75
	avg90-8				avg90-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.89	30.90	43.77	**46.55**	27.91	51.24	52.22	**53.95**
	±0.10	±3.32	±2.29	**±2.08**	±0.15	±0.21	±0.46	**±0.82**
face	27.90	31.64	42.84	**46.00**	27.90	51.26	52.02	**54.64**
	±0.03	±2.81	±1.07	**±1.77**	±0.03	±0.18	±0.55	**±0.77**
nose	27.90	32.34	43.82	**46.54**	27.90	51.26	52.13	**54.13**
	±0.11	±3.14	±1.86	**±1.80**	±0.13	±0.19	±0.44	**±0.81**
frame	27.89	31.81	44.62	**49.02**	27.90	51.25	52.14	**54.18**
	±0.09	±3.09	±1.19	**±1.25**	±0.10	±0.20	±0.49	**±0.85**

Table 2: Structural Similarity Index (for each manually marked region and the frame as a whole) for 8 and 16-bit LR images, generated by extracting frames from raw thermal sequences and then downscaling and upscaling them with a scale of 2 using bicubic interpolation. Generated LR images were then enhanced with DRCN, DRRN or DRESNet (our) SR models. blue - first best for each region and within each input category (single; window 7, 30, 90) separately, **blue** - first best across all input categories

	single-8				single-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.71 ±0.28	0.88 ±0.04	0.98 ±0.01	**0.99 ±0.00**	0.71 ±0.28	0.85 ±0.04	0.97 ±0.01	**0.99 ±0.01**
face	0.64 ±0.27	0.93 ±0.01	0.98 ±0.00	**0.99 ±0.00**	0.64 ±0.27	0.91 ±0.01	0.98 ±0.01	**0.99 ±0.00**
nose	0.53 ±0.39	0.92 ±0.04	0.99 ±0.01	0.99 ±0.00	0.53 ±0.39	0.90 ±0.03	0.99 ±0.01	0.99 ±0.00
frame	0.64 ±0.32	0.89 ±0.06	0.98 ±0.01	**0.99 ±0.01**	0.64 ±0.32	0.92 ±0.01	0.98 ±0.01	**0.99 ±0.01**
	avg7-8				avg7-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.73 ±0.25	0.96 ±0.01	0.98 ±0.01	**0.99 ±0.01**	0.73 ±0.25	0.94 ±0.0.05	0.98 ±0.01	**0.99 ±0.01**
face	0.63 ±0.26	0.96 ±0.01	0.98 ±0.00	**0.99 ±0.00**	0.63 ±0.26	0.96 ±0.01	0.98 ±0.01	**0.99 ±0.00**
nose	0.50 ±0.41	0.96 ±0.01	0.99 ±0.01	0.99 ±0.01	0.50 ±0.41	0.96 ±0.01	0.99 ±0.01	0.99 ±0.01
frame	0.64 ±0.32	0.89 ±0.06	0.98 ±0.01	**0.99 ±0.01**	0.64 ±0.32	0.94 ±0.01	0.98 ±0.01	**0.99 ±0.01**
	avg30-8				avg30-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.70 ±0.25	0.96 ±0.01	0.98 ±0.01	**0.99 ±0.01**	0.70 ±0.25	0.87 ±0.06	0.98 ±0.02	**0.99 ±0.01**
face	0.61 ±0.26	0.96 ±0.01	0.98 ±0.00	**0.99 ±0.00**	0.61 ±0.26	0.93 ±0.01	0.99 ±0.01	**0.99 ±0.00**

nose	0.50 ±0.36	0.98 ±0.01	0.99 ±0.01	0.99 ±0.00	0.50 ±0.36	0.91 ±0.03	0.99 ±0.01	**1.00 ±0.00**
frame	0.62 ±0.30	0.90 ±0.06	0.98 ±0.01	**0.99 ±0.01**	0.62 ±0.30	0.93 ±0.01	0.98 ±0.01	**0.99 ±0.01**
	avg90-8				avg90-16			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.75 ±0.21	0.97 ±0.01	0.99 ±0.01	**0.99 ±0.00**	0.75 ±0.21	0.96 ±0.03	**0.99 ±0.01**	**0.99 ±0.01**
face	0.64 ±0.24	0.98 ±0.00	**0.99 ±0.00**	**0.99 ±0.00**	0.64 ±0.24	0.97 ±0.01	**0.99 ±0.01**	**0.99 ±0.00**
nose	0.48 ±0.37	0.98 ±0.01	0.99 ±0.01	0.99 ±0.00	0.48 ±0.37	0.96 ±0.02	0.99 ±0.00	**1.00 ±0.00**
frame	0.65 ±0.29	0.90 ±0.06	0.98 ±0.01	**0.99 ±0.01**	0.65 ±0.29	0.93 ±0.01	0.98 ±0.01	**0.99 ±0.01**

Table 3: Intersection over Union for detected facial regions for 8 and 16-bit images extracted from raw thermal sequences, original, enhanced with Super Resolution algorithms or resized with bicubic interpolation, evaluated on test sets corresponding to the applied enhancement/degradation algorithm ; blue - first best for each region within each input category (single; window 7, 30, 90) separately, **blue** - first best across all input categories

	single-8					single-16				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	0.90 ±0.03	0.79 ±0.12	0.91 ±0.02	0.90 ±0.04	0.91 ±0.03	0.91 ±0.03	0.85 ±0.08	0.91 ±0.03	0.90 ±0.04	0.91 ±0.04
face	**0.84 ±0.06**	0.33 ±0.38	0.83 ±0.05	0.83 ±0.05	**0.84 ±0.06**	0.80 ±0.20	0.62 ±0.29	0.83 ±0.07	0.83 ±0.61	0.84 ±0.06
nose	0.83 ±0.06	0.31 ±0.38	0.84 ±0.06	0.83 ±0.07	0.85 ±0.08	0.85 ±0.08	0.59 ±0.35	0.84 ±0.08	0.85 ±0.07	0.86 ±0.07
avg.	0.86	0.48	0.86	0.85	0.87	0.85	0.69	0.86	0.86	0.87
	avg7-8					avg7-16				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	**0.95 ±0.03**	0.57 ±0.34	**0.95 ±0.02**	**0.95 ±0.03**	**0.95 ±0.02**	**0.95 ±0.02**	0.88 ±0.09	0.94 ±0.03	**0.95 ±0.02**	**0.95 ±0.02**

face	0.82 ±0.08	0.32 ±0.37	0.81 ±0.08	0.81 ±0.08	0.83 ±0.06	0.82 ±0.06	0.48 ±0.39	0.75 ±0.27	0.81 ±0.8	**0.85 ±0.07**
nose	0.86 ±0.05	0.45 ±0.39	0.87 ±0.06	0.86 ±0.06	**0.88 ±0.06**	0.86 ±0.05	0.65 ±0.25	0.86 ±0.05	0.86 ±0.6	**0.87 ±0.05**
avg.	**0.88**	0.45	**0.88**	0.87	**0.88**	0.88	0.67	0.86	0.85	0.89
	avg30-8					avg30-16				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	**0.95 ±0.02**	0.83 ±0.10	0.94 ±0.02	**0.95 ±0.03**	0.94 ±0.03	0.94 ±0.03	0.89 ±0.05	0.94 ±0.03	0.94 ±0.03	0.94 ±0.04
face	0.80 ±0.09	0.45 ±0.33	0.80 ±0.10	0.80 ±0.08	0.81 ±0.09	0.81 ±0.09	0.48 ±0.34	0.82 ±0.07	0.80 ±0.80	0.81 ±0.08
nose	0.80 ±0.08	0.70 ±0.21	0.81 ±0.07	0.81 ±0.10	0.80 ±0.09	0.82 ±0.07	0.62 ±0.32	0.82 ±0.09	0.80 ±0.10	0.81 ±0.08
avg.	0.66	0.85	0.85	0.85	0.85	0.66	0.86	0.86	0.84	0.85
	avg90-8					avg90-16				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	0.88 ±0.04	0.76 ±0.11	0.89 ±0.05	0.89 ±0.04	0.89 ±0.04	0.89 ±0.04	0.84 ±0.09	0.90 ±0.05	0.88 ±0.05	0.89 ±0.04
face	0.78 ±0.20	0.22 ±0.36	0.78 ±0.19	0.77 ±0.20	0.78 ±0.20	0.78 ±0.20	0.44 ±0.39	0.77 ±0.19	0.72 ±0.26	0.83 ±0.07
nose	0.84 ±0.10	0.34 ±0.39	0.81 ±0.07	0.84 ±0.07	0.83 ±0.09	0.82 ±0.09	0.79 ±0.06	0.83 ±0.07	0.83 ±0.08	0.81 ±0.08
avg.	0.44	0.83	0.84	0.83	0.84	0.69	0.83	0.83	0.81	0.84

5. Discussion

The extensive benchmark evaluation performed for the collected thermal images showed that PSNR can be significantly improved if residual blocks are used in the *Feature Extraction* part of the network. As shown in Table 1, the presented SR CNN model called DRESNet outperformed other state-of-the-art solutions in terms of PSNR by a large margin (in the best case 21.13dB comparing the to bicubic interpolation, 17.21dB comparing to DRCN and 4.4dB comparing to DRRN). Analysis of the SSIM

Figure 7: Thermal images after decreasing and increasing image resolution. From the left: original image with a marked facial area, the enlarged region of the marked facial area and enlarged chosen facial features in images after applying bicubic interpolation, DRCN, DRRN and the proposed DRESNet model.

index also proves the robustness of the DRESNet model, but DRRN achieves similar results. This may be cause by the fact that PSNR is more sensitive to the additive noise [48], that even if very small in low resolution sequences may become exceedingly prevalent. The results confirm that the widened receptive field helps with mitigating the problem of smooth representation of thermal features by analysing more distant

dependencies between interesting components.

In addition, we also proved that utilization of the average of subsequent frames instead of single images help to further increase performance. For all tested SR solutions, the best results were achieved for images calculated as the average of the window covering 90 adjacent frames. Taking it into account, we believe that analysis of differential images is potentially a very interesting research area that can lead to new conclusions. Since it is very hard to avoid uncontrolled movements of volunteers during data collection, differential images may contain some important information about the object of interest (e.g. person) while the rest of the image (e.g background) is removed. Thus, in the future study we would like to explore whether facial features can be detected from the image calculated as a difference between a given frame and the average of subsequent frames (similarly to images presented in even rows in Fig. 2.

Also, the significant finding of this study is that accuracy can be greatly improved by preserving the original bit resolution of images. The dataset collected by us contained raw images with 14-bit resolution what allowed for generating 8 and 16-bit images (16-bit format, but up to 14-bit useful information) that were then used in our experiments. The performed analysis showed that 16-bit resolution produces PSNR of values higher by at least 10% comparing to results achieved for 8-bit images for the presented DRESNet model. In the best case (single frame, eye area) the difference was even higher - 25%. For other SR models utilization of higher resolution data was also helpful. Results achieved by DRCN were improved by ~66%, reducing the PSNR difference between DRCN and DRESNet from ~15.36dB to ~2.97dB for frames creates using 90 subsequent images. For DRRN, the improvement of PSNR was around 10% if 16-bit images were used. This confirms the need of creating thermal face databases in raw formats that contain unprocessed data. We believe that the database published by us may become a very useful reference for further studies on thermal image analysis and processing for e.g. non-contact vital signs estimation [27] from automatically detected facial areas [30].

Another important aim of this research was to evaluate whether increased image quality metrics (PSNR, SSIM) lead to the better accuracy of facial features detection. As presented in Fig. 6, we observe that to a certain level of PSNR (~30dB), the higher

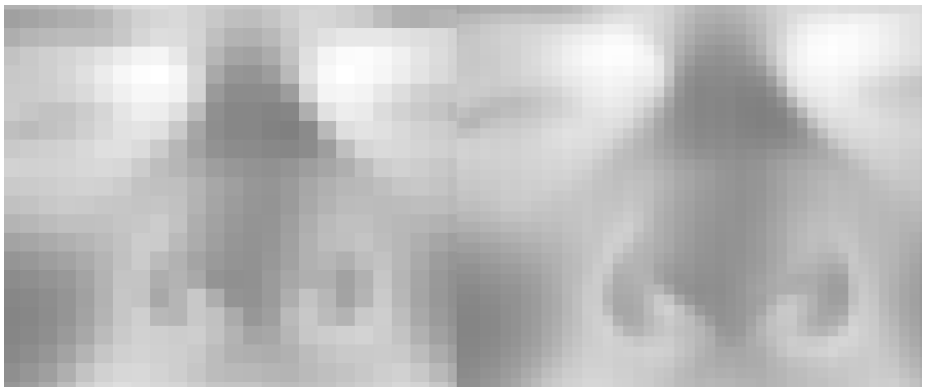

Figure 8: Nostril area extracted from image after scaling it down using bicubic interpolation (on the left) and after applying the proposed DRESNet model (on the right)

PSNR, the better IoU. Yet, once PSNR exceeds this level, the detection model is able to learn correct predictions regardless of PSNR values. Also, it saturates and can't further outperform its state-of-the-art accuracy (for SSD model ~0.85 [38]).

Theoretically, the universal deep learning model should be able to learn a single function *D(x)* (where *D* denotes the detection network applied to the image *x*) equally well as two functions *S(D(x))* (S denotes the SR model). Thermal images though are characterized by smoothed representation of features. Downscaling leads to even more blurred version of the image, where edges of facial features are not clear and their shapes may be distorted (see Fig. 8). Application of CNN-based models, that frequently utilize high frequency component may be not sufficient, as it is hard for the detector to correctly adjust bounding boxes. Thus, the IoU value decrease. Our experiments confirmed this assumption. The use of low resolution data (images downscaled with bicubic interpolation) for facial features detection lead to very poor results (IoU values around 0.5). This proves the need of enhancing images before feeding them to object detection models in order to create an accurate solutions that use thermal image processing.

To further improve the accuracy of facial areas detection, we plan to train our SR network with the augmented data to create solution that is able to better generalize to various databases. Also, we will use gradient clipping that helps to mitigate the problem of exploding gradients [14].

Finally, we believe it would be useful to evaluate other image quality metrics as well (e.g. Information Fidelity Criterion IFC, Noise Quality Measure NQM, Signal to Noise Ratio SNR, Universal Quality Index UQI, Visual Information Fidelity VIF, Visual Signal to Noise Ratio VSNR, etc.). We would like to perform in-depth comparison of these metrics in future work and determine which one would be the best for thermal images. This evaluation is very important because thermal data have different characteristic than RGB images, so maybe different metrics would reflect thermal image enhancement better.

6. Conclusion

The aim of this study was to evaluate various image enhancement methods in low resolution thermal imagery of original bit-resolution and after lossy compresion. For this, we collected and published a database of raw facial images. Extensive benchmark evaluation proved that Peak-Signal-to-Noise Ratio can be improved by 60% (in the best case) if 16-bit resolution data is used instead of 8 bits. Additionally, we presented how DL-based SR model should be designed to address the issue of contextual information being spread over larger image regions due to the heat flow that is visible in thermography. The DRESNEt model presented by us outperformed other SR networks on low resolution thermal images by a margin of ~15dB and ~3dB comparing to DRCN and DRRN, respectively. Also, we showed that it is important to enhance images to improve facial features detection, as LR inputs produce IoU of 0.5 at most for 8-bit images. In the future work we will focus on the detection of facial features from other images that are included in the collected dataset, specifically how the proposed model deals with sequences, where volunteers perform small head movements.

Acknowledgement

This work has been partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology, Intel Corporation, USA and NCBiR, FWF, SNSF, ANR and FNR in the framework of the ERA-NET

CHIST-ERA II, project eGLASSES The interactive eyeglasses for mobile, perceptual computing. We thank all our colleagues, who provided insight and expertise that greatly assisted the research.

References

[1] R. Keys, Cubic convolution interpolation for digital image processing, IEEE transactions on acoustics, speech, and signal processing 29 (6) (1981) 1153–1160.

[2] L. Zhang, X. Wu, An edge-guided image interpolation algorithm via directional filtering and data fusion, IEEE transactions on Image Processing 15 (8) (2006) 2226–2238.

[3] Y. Romano, M. Protter, M. Elad, Single image interpolation via adaptive nonlocal sparsity-based modeling, IEEE Transactions on Image Processing 23 (7) (2014) 3085–3098.

[4] D. Glasner, S. Bagon, M. Irani, Super-resolution from a single image, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 349–356.

[5] G. Freedman, R. Fattal, Image and video upscaling from local self-examples, ACM Transactions on Graphics (TOG) 30 (2) (2011) 12.

[6] Z. Cui, H. Chang, S. Shan, B. Zhong, X. Chen, Deep network cascade for image super-resolution, in: European Conference on Computer Vision, Springer, 2014, pp. 49–64.

[7] H. Chang, D.-Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, Vol. 1, IEEE, 2004, pp. I–I.

[8] K. I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE transactions on pattern analysis & machine intelligence (6) (2010) 1127–1133.

[9] M. Bevilacqua, A. Roumy, C. Guillemot, M. L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

[10] K. Jia, X. Wang, X. Tang, Image transformation based on learning dictionaries across image spaces, IEEE transactions on pattern analysis and machine intelligence 35 (2) (2013) 367–380.

[11] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on pattern analysis and machine intelligence 38 (2) (2016) 295–307.

[12] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[13] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1637–1645.

[14] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017, p. 5.

[15] J. Liu, W. Yang, X. Zhang, Z. Guo, Retrieval compensated group structured sparsity for image super-resolution, IEEE Transactions on Multimedia 19 (2) (2017) 302–316.

[16] X. Li, M. T. Orchard, New edge-directed interpolation, IEEE transactions on image processing 10 (10) (2001) 1521–1527.

[17] H. Chen, X. He, L. Qing, Q. Teng, Single image super-resolution via adaptive transform-based nonlocal self-similarity modeling and learning-based gradient regularization, IEEE Transactions on Multimedia 19 (8) (2017) 1702–1717.

[18] V. Jain, S. Seung, Natural image denoising with convolutional networks, in: Advances in Neural Information Processing Systems, 2009, pp. 769–776.

[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 105–114.

[20] M. S. Sajjadi, B. Schölkopf, M. Hirsch, Enhancenet: Single image super-resolution through automated texture synthesis, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 4501–4510.

[21] S. C. Park, M. K. Park, M. G. Kang, Super-resolution image reconstruction: a technical overview, IEEE signal processing magazine 20 (3) (2003) 21–36.

[22] L. G. Villanueva, G. M. Callicó, F. Tobajas, S. López, V. De Armas, J. F. López, R. Sarmiento, Medical diagnosis improvement through image quality enhancement based on super-resolution, in: Digital System Design: Architectures, Methods and Tools (DSD), 2010 13th Euromicro Conference on, IEEE, 2010, pp. 259–262.

[23] M. Abdel-Nasser, J. Melendez, A. Moreno, O. A. Omer, D. Puig, Breast tumor classification in ultrasound images using texture analysis and super-resolution methods, Engineering Applications of Artificial Intelligence 59 (2017) 84–92.

[24] Y. Gao, H. Li, J. Dong, G. Feng, A deep convolutional network for medical image super-resolution, in: Chinese Automation Congress (CAC), 2017, IEEE, 2017, pp. 5310–5315.

[25] A. Kwaśniewska, A. Giczewska, J. Rumiński, Big data significance in remote medical diagnostics based on deep learning techniques, Task Quarterly 21 (2017) 309–319.

[26] M. Lewandowska, J. Rumiński, T. Kocejko, J. Nowak, Measuring pulse rate with a webcama non-contact method for evaluating cardiac activity, in: Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on, IEEE, 2011, pp. 405–410.

[27] J. Ruminski, A. Kwasniewska, Evaluation of respiration rate using thermal imaging in mobile conditions, in: Application of Infrared to Biomedical Sciences, Springer, 2017, pp. 311–346.

[28] M. Hanmandlu, et al., A new entropy function and a classifier for thermal face recognition, Engineering Applications of Artificial Intelligence 36 (2014) 269–286.

[29] A. Kwaśniewska, J. Rumiński, Face detection in image sequences using a portable thermal camera, in: Proceedings of the 13th Quantitative Infrared Thermography Conference, 2016.

[30] A. Kwaśniewska, J. Rumiński, K. Czuszyński, M. Szankin, Real-time facial features detection from low resolution thermal images with deep classification models, Journal of Medical Imaging and Health Informatics 8 (5) (2018) 979–987.

[31] M. Szankin, A. Kwasniewska, T. Sirlapu, M. Wang, J. Ruminski, R. Nicolas, M. Bartscherer, Long distance vital signs monitoring with person identification for smart home solutions, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 1558–1561.

[32] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, Vol. 1, MIT press Cambridge, 2016.

[33] R. Gade, T. B. Moeslund, Thermal cameras and applications: a survey, Machine vision and applications 25 (1) (2014) 245–262.

[34] Flir lepton camera modules, `https://www.flir.com/products/lepton/`, accessed: 2018-11-10.

[35] J. Rumiński, Analysis of the parameters of respiration patterns extracted from thermal image sequences, in: Biocybernetics and Biomedical Engineering, Vol. 36, 2016, pp. 732–741.

[36] B. Qi, V. John, Z. Liu, S. Mita, Pedestrian detection from thermal images: A sparse representation based approach, Infrared Physics & Technology 76 (2016) 157–167.

[37] K. A. R. J. N. R. Szankin, Maciej, Road condition evaluation using fusion of multiple deep models on always-on vision processor, in: Proc. Of the 44th Annual Conference of the IEEE Industrial Electronics Society, in print, 2018.

[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[40] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[41] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[42] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[43] Tensorflow implementation of sr models, `https://github.com/LoSealL/VideoSuperResolution`, accessed: 2018-10-10.

[44] Drrn repository, `https://github.com/tyshiwo/DRRN_CVPR17`, accessed: 2018-10-10.

[45] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, IGI Global, 2010, pp. 242–264.

[46] Tensorflow detection model zoo, https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md, accessed: 2018-11-10.

[47] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (Feb) (2012) 281–305.

[48] A. Hore, D. Ziou, Image quality metrics: Psnr vs. ssim, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 2366–2369.