Cascaded Refinement Network: Thermal and Cross-spectral Palm Image Matching in the Visual Domain by Robust Image Transformation

Ewelina Bartuzi*, Naser Damer

Biometrics and Machine Learning Groups
Institute of Control and Computation Engineering
Faculty of Electronics and Information Technology, WUT

Seminarium naukowe 4

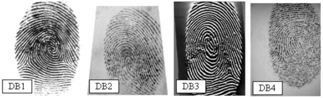
Contents

- synthesis visual-like palm images from thermal images
- utilizing Cascaded Refinement Network with contextual loss function
- assessing the images quality for original, targeted, and generated samples

Personal features of the hand

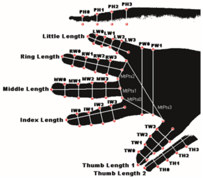
Hand biometric has been used since the XIX century...

fingerprint



Source: FVC2004

geometric features

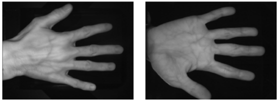


Source: Use of hand in biometrics, A. Czajka

palmprint

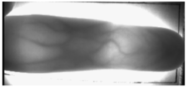
Source: IIT Delhi

hand vein pattern



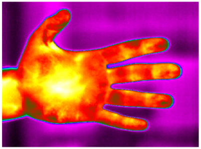
Source: A Novel Biometric System Based on Hand Vein, X. Wu et al.

finger vein pattern



Source: PolyU

thermal features



Source: BioBase-Hand-Thermal

Motivation

- Systems based on hand characteristics have many advantages, mainly related to the simple acquisition process and usability.
- Acquisition can be performed in a contactless manner, resulting in a hygienic and convenient process that can enable continuous verification.
- The palm side of the hand is also rarely exposed in whole (less imitation attack-prone), and is socially acceptable.
- Recent works confirmed the discriminatory information content in the thermal hand images.
- More importantly, these thermal features are independent of external illumination variations and it is hard to reconstruct heat maps to perform presentation attacks on such a system.
- On the other hand, thermal representations are highly dependent on temperature changes caused by variations in the body metabolism and other physiological processes.

A challenge: developing a comparison pipeline that is robust to such variations, e.g. a thermal-to-visual image conversion approach that considers such variations.

Image Transformation Methods

Image Transformation Methods

- The recent image spectral transformation methods relaid on the recent advancements around the idea of **generative adversarial networks** (GAN).

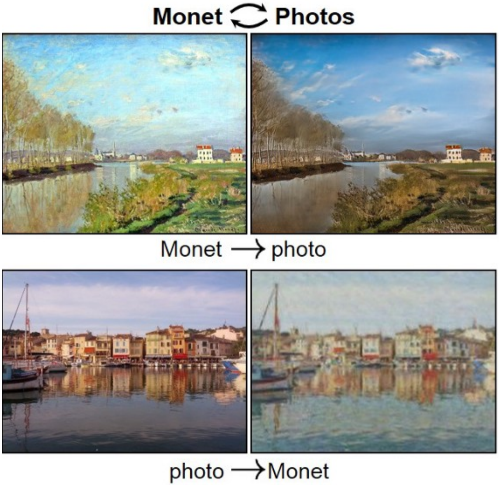
- The first GAN solutions **lacked stability**,

- More advanced designs, such as Boundary Equilibrium Generative Adversarial Networks (BEGAN) and Deep Convolutional Generative Adversarial network (DCGAN), **offered more stability** but failed to produce images of high **resolution**.

- This issue was addressed by the Cycle-Consistent Adversarial Networks (CycleGAN) and Image-to-Image Translation with Conditional Adversarial Nets (Pix2Pix), which **increased the complexity** of the network and thus its training data needs.

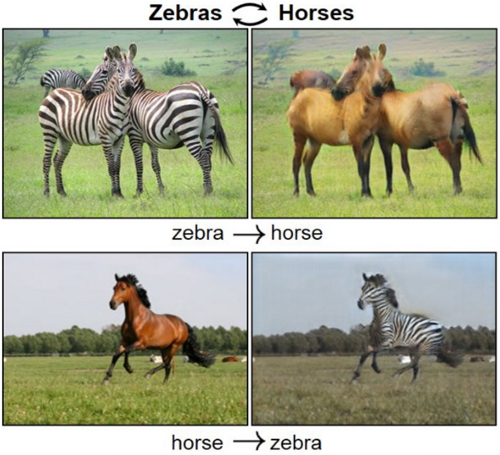
GAN: Image-to-Image Translation using GANs

"The coolest idea in deep learning in the last 20 years." - Yann LeCun on GANs.



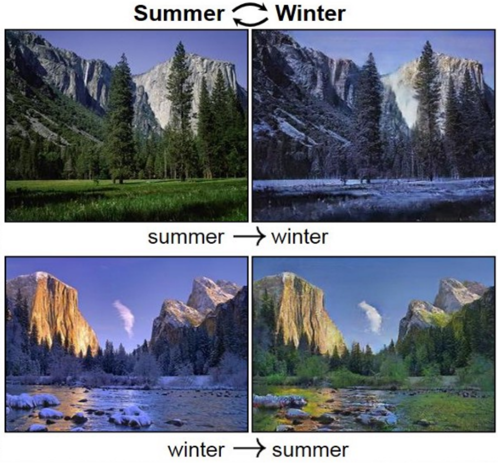
GAN: Image-to-Image Translation using GANs

"The coolest idea in deep learning in the last 20 years." - Yann LeCun on GANs.



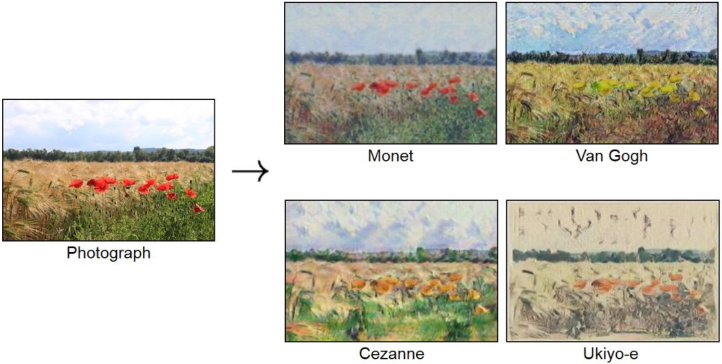
GAN: Image-to-Image Translation using GANs

"The coolest idea in deep learning in the last 20 years." - Yann LeCun on GANs.

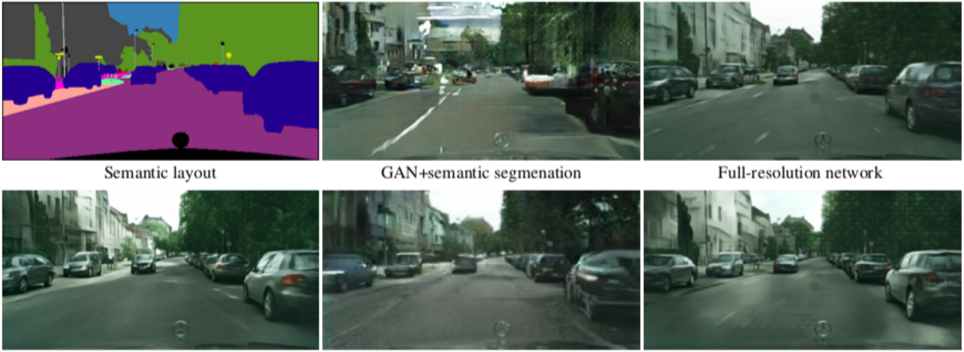


GAN: Image-to-Image Translation using GANs

"The coolest idea in deep learning in the last 20 years." - Yann LeCun on GANs.



CRN: Cascaded Refinement Network



Semantic layout

GAN+semantic segmenation

Full-resolution network

Our result

Isola et al. [16]

Encoder-decoder

Figure 5. Qualitative comparison on the Cityscapes dataset.

Q. Chen and V. Koltun, "Photographic image synthesiswith cascaded refinement networks", IEEE International Conference on Computer Vision, ICCV 2017

CRN: Cascaded Refinement Network

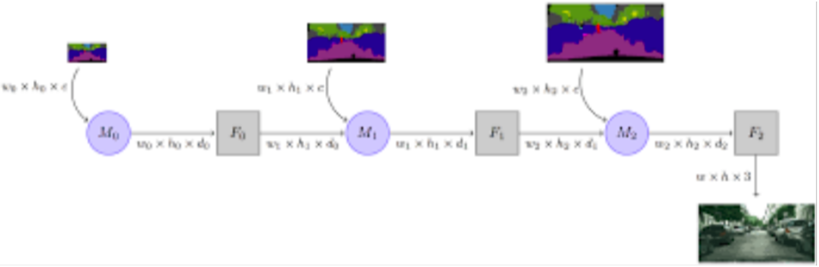
- A new approach to synthesis photographic images conditioned on semantic layouts.
- Using a semantic label map, this approach produces an image with photographic appearance that conforms to the input layout.
- The approach thus functions as a rendering engine that takes a two-dimensional semantic specification of the scene and produces a corresponding photographic image.
- This network has a specific, modular architecture using convolutional layers.
- Solution in the first work was trained with direct regression loss = perceptual loos + trics to synthesize a diverse collection of images (collection of images with difference appearance - white/black/silver cars, day/nighttime iamges)

Q. Chen and V. Koltun, "Photographic image synthesiswith cascaded refinement networks", IEEE International Conference on Computer Vision, ICCV 2017

CRN: Cascaded Refinement Network (1/2)

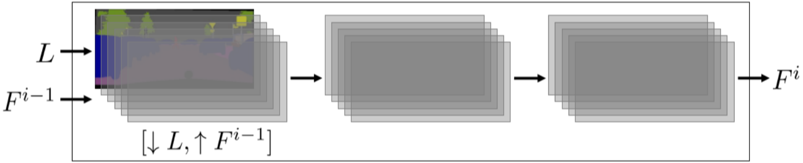
The CRN is a convolutional neural network that consists of inter-connected refinement modules.

- each module consists of **three layers**: input, intermediate, and output layer,
- the first module considers the lowest resolution space (4×4 in this case)
- this resolution is duplicated in the successor modules until the last module, matching the target image resolution,
- from the second to the last module expects two inputs, the output representation of the previous module and the input image at the specific resolution of the module,


$w_0 \times h_0 \times c$
$w_1 \times h_1 \times c$
$w_2 \times h_2 \times c$
M_0
$w_0 \times h_0 \times d_0$
F_0
$w_1 \times h_1 \times d_0$
M_1
$w_1 \times h_1 \times d_1$
F_1
$w_2 \times h_2 \times d_1$
M_2
$w_2 \times h_2 \times d_2$
F_2
$w \times h \times 3$

- not colors, but activation in a pretrained perceiver network are matched

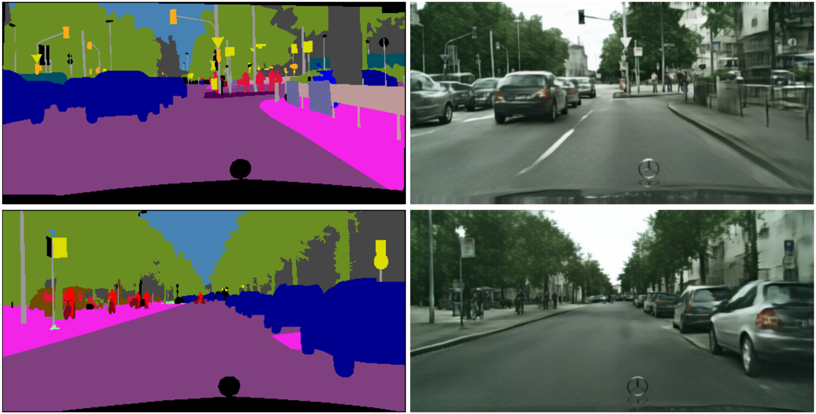
CRN: Cascaded Refinement Network



Rysunek: A single module of a CRN, which is at every octave/level of resolution.

L - sementic layouts
F - feature maps

CRN: Cascaded Refinement Network



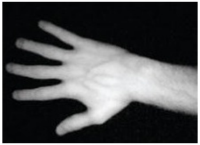
(a) Input semantic layouts

(b) Synthesized images

Experimental data

Experimental data

- *Tecnocampus Hand Image Database* (THID),
- multispectral database, which contains hand images obtained in three wavelength ranges: visible light, near-infrared, far-infrared,
- two subsets of this database were used in experiments:
  - **thermal images – oT**, acquired by thermal camera Testo 882-3 with sensor of range from 3 to $14\mu m$ $(320 \times 240$ px),
  - **visible light images – oV**, collected by digital camera $(380 - 750nm, 640 \times 480$ px)
- 111 subjects × 5 session × 2 images,
- constrained test-bench: it is possible to estimate hand translation



Experimental data

- *Tecnocampus Hand Image Database* (THID),
- multispectral database, which contains hand images obtained in three wavelength ranges: visible light, near-infrared, far-infrared,
- two subsets of this database were used in experiments:
    - **thermal images – oT**, acquired by thermal camera Testo 882-3 with sensor of range from 3 to $14\mu m$ $(320 \times 240$ px),
    - **visible light images – oV**, collected by digital camera $(380 - 750nm$, 640×480 px)
- 111 subjects $\times$ 5 session $\times$ 2 images,
- constrained test-bench: it is possible to estimate hand translation

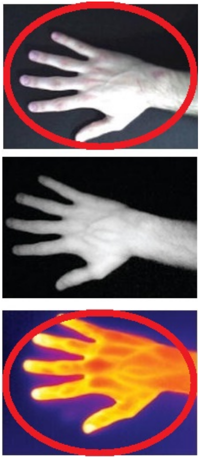
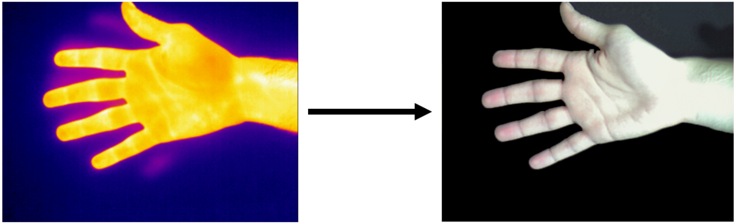
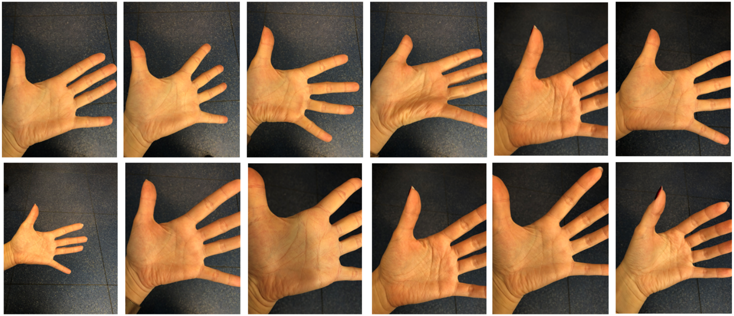


Image transformation



Generation methodology: challenges

a. Palms have a high dynamic nature ⇒ palm images transformation should focus on the entire image containing the object
b. The palm scale within the image is variable and dependent on the capture setup ⇒ automatic scale correction or scale-invariant method
c. Unconstrained measurement (palm aligning)
d. The generated images have to be of realistic resolution



Generation methodology: Cascaded Refinement Network

Why CRN?

- The CRN considers **multi-scale information** and is based on training a relatively **small number of parameters** (leading to lower need of training data)
- output images are characterized by:
  - high resolution
  - does not require data alignment or scaling (if we use a suitable loss function)
- capture photographic appearance
- image synthesis is fast

CRN details:

- the target (generated) image resolution in the experiments was set to 512×512 pixels
- 40 epochs, learning rate = 0.0001
- training: 6h

Generation methodology: contextual loss function

- the CRN training is controlled by a loss function, which should be invariant to exact **scale**, **alignment**, and shape of the hand/fingers position,
- this also inherently covers the training with not perfectly aligned pairs of visual and thermal images

Challange: a loss function that neglects outliers on the pixel level (in comparison to pixel-level loss).
Answer: the contextual loss function (CL)
The CL function is calculated between the **source (S)** and the **generated (G)** images, and between the target (ground-truth) and the generated images.

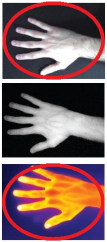
- The source-generated loss aims at saving the details of the source image such as detailed boundaries.
- The target-generated loss maintains the properties of the target image in the generated image, e.g. target image style and content.

In our case, the source (thermal) and target (original visible-light) training image pairs are of identical, but not aligned or have the same scale.

Generation methodology: contextual loss function

$$
\begin{aligned}
L_{CX}(s,t,g,l_s,l_t) = & \lambda_1(-log(CX(\Phi_1^{l_s}(g),\Phi^{l_s}(s))))+ \\
& \lambda_2(-log(CX(\Phi_2^{l_t}(g),\Phi^{l_t}(t)))),
\end{aligned}
\qquad (1)
$$

where:



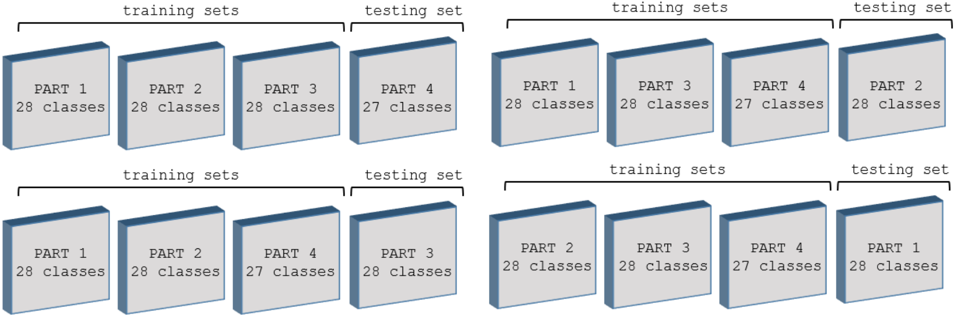
s, t, g – the source, target, and generated images, respectively,
CX – the rotation and scale invariant contextual similarity*

Φ – a perceptual network, VGG19 in our work,
$\Phi^{l_s}(x)$, $\Phi^{l_t}(x)$ – the embeddings vectors extracted from the image at layer l_s (conv4_2) and l_t (conv3_2 and conv4_2) of the perceptual network, respectively,
$\lambda_1 = 0.01$ and $\lambda_2 = 0.99$ by checking the resulting generated image visually. That was motivated to capture the style of the target image (visual), but maintain the main properties of the source image, shape and location.

**R. Mechrez, et.al, "The contextual loss for image transformation with non-aligned data", ECCV, Germany, 2018, proceedings, vol. 11218, Springer, pp. 800–815,*

Generated dataset: process

- **Cascaded Refinement Network** architecture,
- minimizing **contextual loss function** ,
- data of 111 identities was split into 4 parts,
- 4 trained transformation model,
- batch size = 1, learning rate = 0.0001, image size = 512×512 px



training sets | testing set

PART 1 28 classes | PART 2 28 classes | PART 3 28 classes | PART 4 27 classes

training sets | testing set

PART 1 28 classes | PART 3 28 classes | PART 4 27 classes | PART 2 28 classes

training sets | testing set

PART 1 28 classes | PART 2 28 classes | PART 4 27 classes | PART 3 28 classes

training sets | testing set

PART 2 28 classes | PART 3 28 classes | PART 4 27 classes | PART 1 28 classes

Generated dataset: examples (1/4)

- **'Warm' palm**

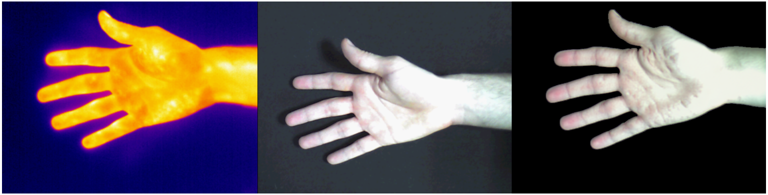


Figure: From left to right: **original visible light images (oV)**, **original thermal images (oT)**, and **generated visible-like images from thermal spectrum (gV)**.

Generated dataset: examples (2/4)

- **Palm with cold fingertips**

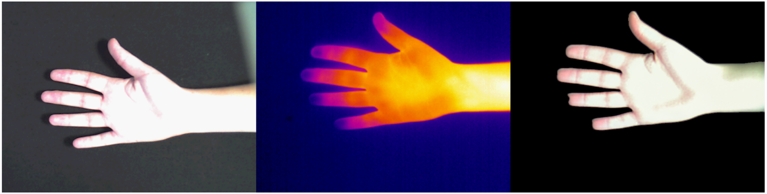


Figure: From left to right: **original visible light images (oV)**, **original thermal images (oT)**, and **generated visible-like images from thermal spectrum (gV)**.

Generated dataset: examples (3/4)

- **Palm with complete cold fingers**

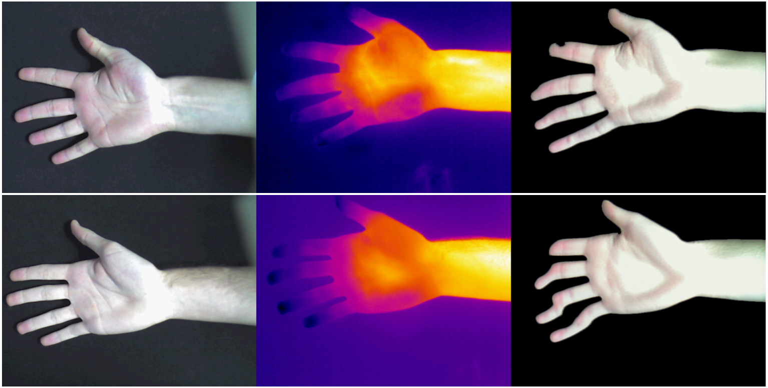
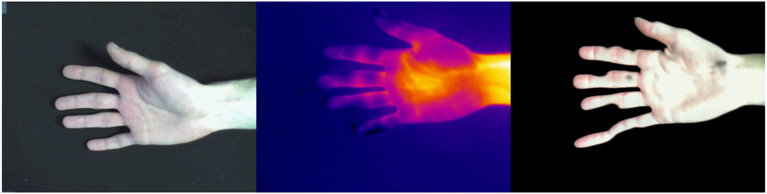


Figure: From left to right: **original visible light images (oV)**, **original thermal images (oT)**, and **generated visible-like images from thermal spectrum (gV)**.

Generated dataset: examples (4/4)

- **Palm with large cold regions**



- **Palm with heat-shade effect**

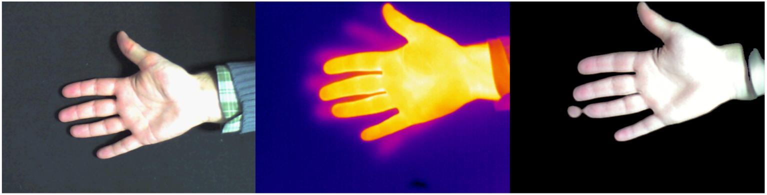


Figure: From left to right: **original visible light images (oV)**, **original thermal images (oT)**, and **generated visible-like images from thermal spectrum (gV)**.

Image quality: no-reference image quality metrics

Unlearned metrics:

- **Global Contrast Factor (GCF)** closer to the human perception of contrast, measures richness of detail as perceived by a human observer, and is used in various application areas (generating images, rendering, tone mapping, volume visualization, and lighting design etc.)(K. Matković, 2005)
- **Sharpness** calculated via sum of image gradient (*X. Gao, 2007*)

Learned metrics:

- **Natural Image Quality Evaluator (NIQE)** measures the distance between the NSS-based features calculated from image A to the features obtained from an image database used to train the model. The features are modeled as multidimensional Gaussian distributions (*A. Mittal, 2013*)
- **Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)** does not compute distortion specific features such as ringing, blur or blocking, but instead uses scene statistics of locally normalized luminance coefficients to quantify possible losses of 'naturalness' in the image due to the presence of distortions, thereby leading to a holistic measure of quality

Image quality: no-reference image quality metrics

Unlearned metrics:

- **Global Contrast Factor (GCF)** weighted local contrast at various resolution level,
- **Sharpness** calculated via sum of image gradient,

Learned metrics:

- **Natural Image Quality Evaluator (NIQE)** measures the distance between the NSS-based features,
- **Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)** – uses scene statistics of locally normalized luminance coefficients to quantify possible losses of 'naturalness'

Tabela: Mean values (± standard deviation) of quality metrics.

	GCF	Sharpness	NIQE	BRISQUE
oV	8.13(±0.49)	0.060(±0.007)	7.03(±1.69)	47.44(±1.49)
gV	8.19(±0.54)	0.062(±0.006)	6.35(±0.49)	48.03(±0.30)
oT	7.32(±1.43)	0.091(±0.018)	4.49 (±0.45)	33.25(±4.85)

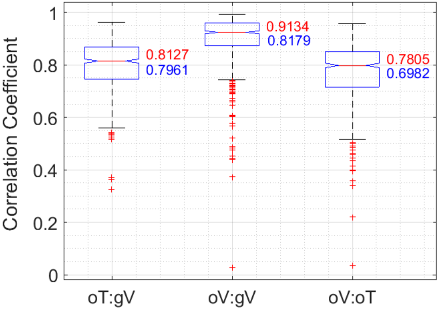
Conclusions: One can notice the high similarity in the quality measures between the generated visual-like images and the original visual one. This is an indication of the success of the generated images in imitating the quality of the targeted original images. The original thermal images, as expected, have a different quality nature compared to the visual ones.

Correlation to original images

The inter-spectral correlation coefficient (r) was used to assess the similarity between original and generated images:

$$r_{o:g} = \frac{\sum_{i=1}^{n}(O_i - \overline{O})(G_i - \overline{G})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2 \sum_{i=1}^{n}(G_i - \overline{G})^2}},$$

where O_i denotes the i-th pixel value of the original image, and G_i corresponds to the value of the i-th pixel in the generated image. $\overline{O}$, $\overline{G}$ are mean pixel values.



Comment: The determination of the correlation coefficient was preceded by the alignment of the hand position on the images using the SURF features and applying geometric transformation.

Palm recognition

Texture-based approach: Local Binary Patterns

- LBP is one of the most popular texture descriptors used in image analysis, and represents the contrast differences in the neighbourhood of each pixel in the image.

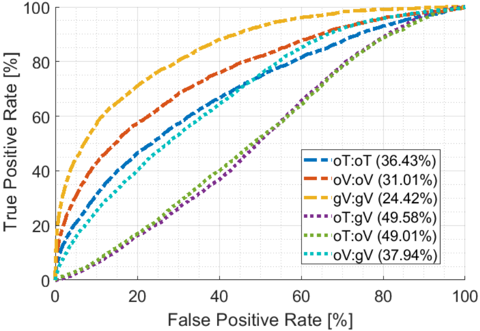
$$LBP_{N,R}(I_C) = \sum_{n=1}^{N} s(I_n - I_C)2^{n-1}, \quad (2)$$

where:
N, R are the number of surrounding neighbors and the radius, respectively,
I_C – central pixel, I_N – N-th pixel from neighborhood of central pixel,
$s(I_n - I_C) = 1$, while $I_n - I_C > 0$ and $s(I_n - I_C) = 0$ otherwise.

- The obtained values are formed as a histogram for **the whole image** (no-patches), as well as concatenated vector from vectors representing histograms from image patches in size 28×28 or 56×56 pixels.
- Feature vectors were compared using χ^2 metrics.

Texture-based approach: Experiments Results

- verification scenario
- session-disjoint splits
- the proposed transformation approach allowed to extract static features of thermal images that are more resistant temperature changes over time and les sensitive to external and physiological factors (EER = 24.42% for gV:gV)



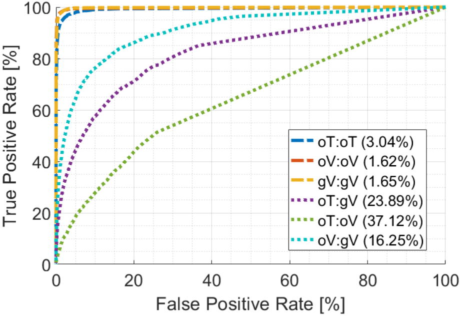
CNN-based approach

- pre-trained AlexNet model: 5 convolutional layers and 3 fully connected layers (11×11×96, 5×5×256, 3×3×384, 3×3×384, 3×3×256)
- modified classifier
- fine-tuned with palm images obtained in visible light, thermal images, and generated images, respectively

Options	modified AlexNet
training/validation/test	60:20:20
batch size	8
data augmentation	random rotation: $\pm 60^0$, random translation: ±30 pixels, random reflection
optimization method	SGD (with momentum=0.9)
learning rate	0.0001
other	data shuffling before each training epoch L2 regularization validation-stop

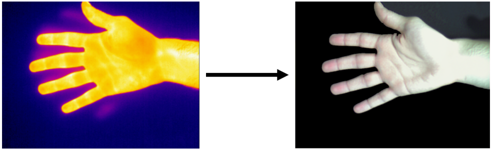
CNN-based approach: Experiments Results

- verification scenario
- session-disjoint splits
- data was divided into 3 parts: trainig/validation/test in ratio of 3:1:1
- for comparisons between images of the same type, the EER was about 1.6% for oV:oV and gV:gV, and for thermal images over 3%.
- the verification error between the thermal images was reduced from 3.04% to 1.65% after the proposed transformation into the visual space (oV:oV). This proves our goals in being more robust to thermal variations and taking advantage of visual-trained networks after our transformation. For the cross-spectral verification, our proposed transformation improved the accuracy from 37.12% EER (oT:oV) down to 16.25% (oV:gV).



Conclusions

- **The first purpose of this work:** a novel approach to transfer thermal images into the visual domain by utilizing the **CRN** trained with a loss function based on the **contextual similarity**.
    - the use of high performing deep networks pre-trained on images from that domain (visual)
    - a solution that targets the **alignment-free** and **scale-invariant** nature of palm images, as well as the **limited size** of the training data
    - the proposed transformation produced images of similar quality as the original visual images, characterized by high correlation to them



Conclusions

- **The second purpose of this work:** comparison of thermal images in the visual domain, and cross-spectral comparison of thermal and visual palm images.
    - we used these generated images to perform thermal-to-thermal and thermal-to-visual verification.
    - two approaches: using hand-crafted features, and embedding generated by pre-trained CNN
    - the comparison of generated images from thermal-to-thermal comparisons lead to higher verification accuracy in comparison to using the original thermal images.
    - the EER of cross-spectrum verification using CNN-extracted features was reduced from 37.12% to 16.25% after transferring the thermal image into the visual domain.

Comparison type	LBP	CNN-based	Comparison type	LBP	CNN-based
oT:oT	36.43	3.04	**oT:gV**	49.58	23.89
oV:oV	31.01	1.62	**oT:oV**	49.01	37.12
gV:gV	24.42	1.65	**oV:gV**	37.94	16.25

It must be noted that using thermal images for verification helps avoiding presentation attacks. However, an attacker may transfer a thermal image to the visual-spectrum and attack a visual verification system.

Bibliography

1. K. Matković et.al., '*Global Contrast Factor - a New Approach to Image Contrast*', Computational Aesthetics in Graphics, Visualization and Imaging (2005)
2. X. Gao et.al., '*Standardization of Face Image Sample Quality*', ICB 2007, proceedings Springer
3. A. Mittal, R. Soundararajan, and A. Bovik, 'Making a Completely Blind Image Quality Analyzer', IEEE Signal Processing Letters. Vol. 22, Number 3, 2013, pp. 209–212
4. A. Mittal, A. Moorthy, and A. C. Bovik. 'No-Reference Image Quality Assessment in the Spatial Domain', IEEE Transactions on Image Processing. Vol. 21, Number 12, December 2012, pp. 4695–4708
5. Q. Chen and V. Koltun, "Photographic image synthesiswith cascaded refinement networks," ICCV 2017,
6. L. Xu, J. S. J. Ren, C. Liu, and J. Jia, "Deep convolutionalneural network for image deconvolution," Advances in Neural Information Processing Systems, 2014
7. L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," CVPR 2016

Cascaded Refinement Network: Thermal and Cross-spectral Palm Image Matching in the Visual Domain by Robust Image Transformation

Ewelina Bartuzi*, Naser Damer

Biometrics and Machine Learning Groups
Institute of Control and Computation Engineering
Faculty of Electronics and Information Technology, WUT

Seminarium naukowe 4