

Supplementary Material I: Skin-based identification from multispectral image data using CNNs

Takeshi Uemori¹ Atsushi Ito² Yusuke Moriuchi² Alexander Gatto¹ Jun Murayama²
¹Sony Europe B.V., Stuttgart, Germany ²Sony Corporation, Tokyo, Japan
{Takeshi.Uemori, Atsushi.C.Ito, Yusuke.Moriuchi, Alexander.Gatto, Jun.Murayama}@sony.com

Abstract

In this supplementary material, we provide additional details on two fronts. Firstly, we provide the experimental details including source data specifications of generated datasets, camera configurations in the actual image experiment, and implementation parameters of CNNs. Secondly, we present additional results to facilitate a deeper understanding of our method.

1. Experimental details

Source data specifications: Table 1 in this material shows the specifications of source data for generating synthetic datasets in Section 4. Samples from the dataset #1 are shown in the columns (a) of Figure 2 in this material. The 123 spectral profiles of dataset #2 are shown in Figure 6 in the main paper.

Camera configurations: Table 2 in this material shows the configurations of RGB and multispectral cameras which we used in the actual image experiment described in Section 6. The RGB camera was operated with binning mode, and the bit depth was converted from 12 to 10 bit by rounding in order to adjust to bit depth of the multispectral camera. We used only 8 color bands, and discarded black-and-white pixels of the multispectral camera. The sensor sensitivities of each camera are shown in Figure 5 in the main paper. To adjust the total data amount in the patch for identification, we resized RGB images by nearest neighbor with the ratio 0.784. Figure 1 in this material shows our experimental setup.

Implementation parameters of CNNs: We implemented three types of CNNs using [36]. They are based on the Wide-ResNet architecture [44]. The first is 2D CNN which is the original Wide-ResNet. The second is 3D CNN in which 2D convolutions in 2D CNN are extended to 3D convolutions. The third is our proposed network in which SE-blocks are added to 3D CNN. Our proposed network is shown in Figure 3 in the main paper. The dropout with the

Table 1. Source specifications for synthetic dataset generation

Pipeline	Dataset #1	Dataset #2
How to collect	By ourselves	Purchased SOCS dataset [19]
Number of subjects	12 people / 20 hands (incl. 8 pairs of each hand)	123 people
Skin type	Japanese males	Japanese females
Skin condition	Bare skin at back of hand	Bare skin at forehead
Spatial resolution	168×128 pixels	1 pixel
Wavelength range	400 - 700 nm	400 - 700 nm
Wavelength step	1 nm	10 nm
Light source	Halogen lamp	Tungsten lamp

Table 2. Camera configurations

Camera type	RGB	Multispectral
Name	acA2500-14gc	CMS-C
Manufacture	Basler AG	SILIOS Technologies
Pixel number in all mosaic-array	1294×970 pixels	1280×1024 pixels
Pixel number in each band	647×485 pixels	426×339 pixels
Number of band	3	8 colors and 1 black-and-white
Bit depth	12	10

ratio 0.4 was used for all CNNs. The reduction ratio r in SE-block was set to 8. We used the momentum stochastic gradient descent as the optimizer for all CNNs with the parameters: $momentum = 0.9$, $weight\ decay = 0.0001$, and $learning\ rate = 0.1$. The learning rate was decreased by a factor of 10 at a half and at three-quarters of total epochs which was 200. The batch size for 2D CNN was set to 128, and the batch size for 3D CNN and our proposed network were set to 64.

2. Additional results

In this section we present additional results to facilitate a deeper understanding of our method. Figure 2 in this material shows the experimental results with synthetic datasets in Section 5.4. It demonstrates the identification performance dependent on the number of input spectral bands and CNN algorithms. The additional results on the actual image ex-

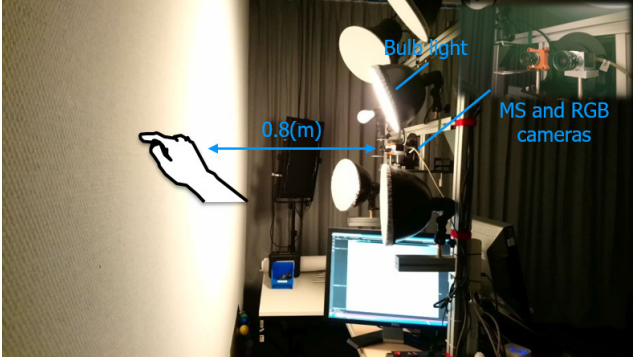


Figure 1. Experimental setup with RGB and multispectral cameras: both cameras were installed on the system which was about 0.8 meters away from the wall. They were synchronized by software with a framerate of about 5 frame per second. Subjects moved their hands on the wall lit by a bulb light.

periment in Section 6 are shown in a video named Supplementary Material II. It includes frame-by-frame prediction results with the actual multispectral camera and our proposed network.

Additionally, we prepared an extra video as a trailer. We recommend to watch this video at first so that you can understand the overall content of this paper.

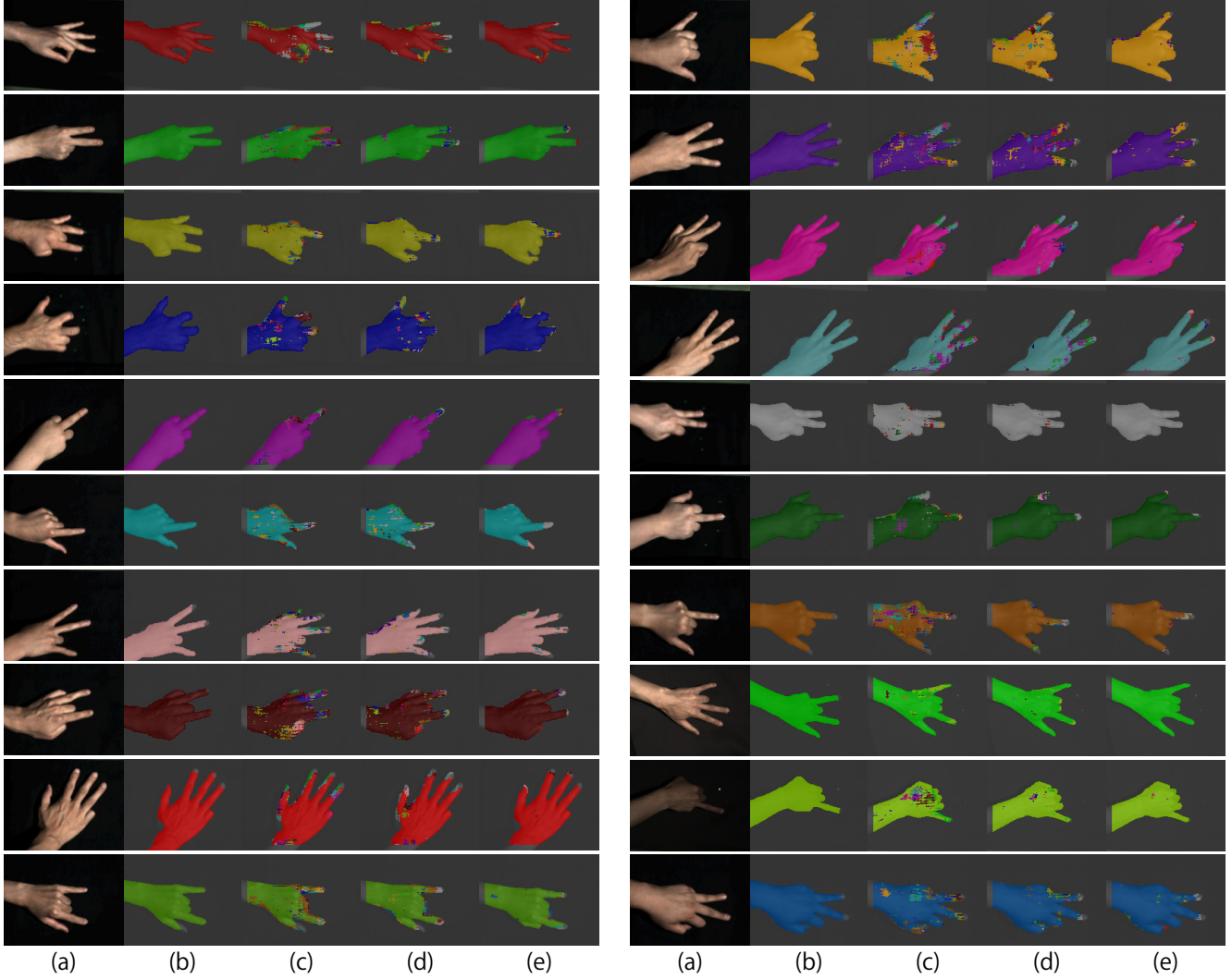


Figure 2. Selected results on the synthetic dataset of all 20 hands: (a) input data visualized as RGB images, (b) the labels with ground truth color, (c) the prediction results on RGB (3 bands) input with 2D CNN, (d) the prediction results on multispectral (8 bands) input with our CNN, and (e) the prediction results on multispectral (16 bands) input with our CNN. (c) and (d) were simulated to be captured with the acA2500-14gc and the CMS-C respectively, and their performance were evaluated in Section 5.4 in the main paper. Input of (e) was generated in Section 5.1, and was the optimal combination of spectral and spatial information. We can see potential ability of our approach from (e). Basically results on multispectral inputs with our CNN (d) and (e) are better than results on RGB with 2D CNN (c), but failures are still found around boundaries and on nails.