Skin-based identification from multispectral image data using CNNs

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

Abstract

User identification from hand images only is still a challenging task. In this paper, we propose a new biometric identification system based solely on a skin patch from a multispectral image. The system is utilizing a novel modified 3D CNN architecture which is taking advantage of multispectral data. We demonstrate the application of our system for the example of human identification from multispectral images of hands. To the best of our knowledge, this paper is the first to describe a pose-invariant and robust to overlapping real-time human identification system using hands. Additionally, we provide a framework to optimize the required spectral bands for the given spatial resolution limitations.

1. Introduction

Personal identification using unique physiological features such as a face, an iris, fingerprints or vein, is required in a wide variety of systems and applications. Especially in the past couple of years, there has been significant increase of practical logon applications for different types of devices such as cellular phones [3, 28], laptops [24], and video game consoles [33]. In the case of tabletop devices, which fall into the category of a so-called tangible user interface [18, 34], identifying a user from his hands only is desired for individual access controls and natural user interfaces. The system is required to work without any constraints on hand pose, in contrast to fingerprint or vein which usually forces users to pause at the ideal pose. In recent years, thanks to progress in deep learning, there has been outstanding progress in a variety of computer vision tasks. The standard way to perform image-based recognition is to use geometric information. While this approach is suitable for relatively rigid objects such as faces, hand recognition often requires a particular hand pose [29, 5, 43, 30, 13]. Our method is pose-invariant and can deal with occlusions.

Recently, multispectral image acquisition systems which capture data at several specific wavelength ranges (usually more narrowly than RGB image acquisition systems) have become easier available. Due to the complexity of traditional multispectral acquisition systems, proposed applications have been limited to very specific fields [16, 22]. However, various types of systems [7, 11] are being developed recently and spreading to wider commercial applications [27, 6] including biometrics [45, 1]. Especially the advent of multispectral mosaic-array sensors [42, 21, 15] enables the acquisition of video as a sequence of single-snapshots. However, the disadvantage of these sensors is the trade-off between spatial resolution and the number of spectral bands, which means the spatial resolution is sacrificed if the number of spectral bands is increased or vice versa.

Several skin spectra models have been proposed in early works [2, 40]. According to their optical considerations, perceived color is mainly composed of dermis scattering, melanin and vascular absorption which are different among individuals due to the differences of skin chromophore concentrations [35, 40]. Our approach has been motivated by knowledge from these studies.

In this paper, we propose a framework of hand identification using spatial and spectral distributions of skin, without using any geometric information, as shown in Figure 1. From a small patch (i.e. 16x16 pixels) of a hand, our CNN model can distinguish among registered users, without using any additional information about a hand's shape. One advantage of our patch-based identification is that it works even when a hand of a user to be identified is overlapped by a hand of a different user, or a part of a hand is out of the view. Additionally, our model can distinguish between the left and right hand of a person because our CNN model learns different spectral and spatial features of skin for each hand. Finally our approach works in a frame-by-frame fashion, making real time processing more feasible. We demonstrate this user identification framework in the scenario of a tabletop projection system as shown in Figure 2. A multispectral camera, including a projection system, is mounted over the tabletop and acquires images of hands which are moving without any constraints on the table. These capabilities may provide a novel natural user interaction for many

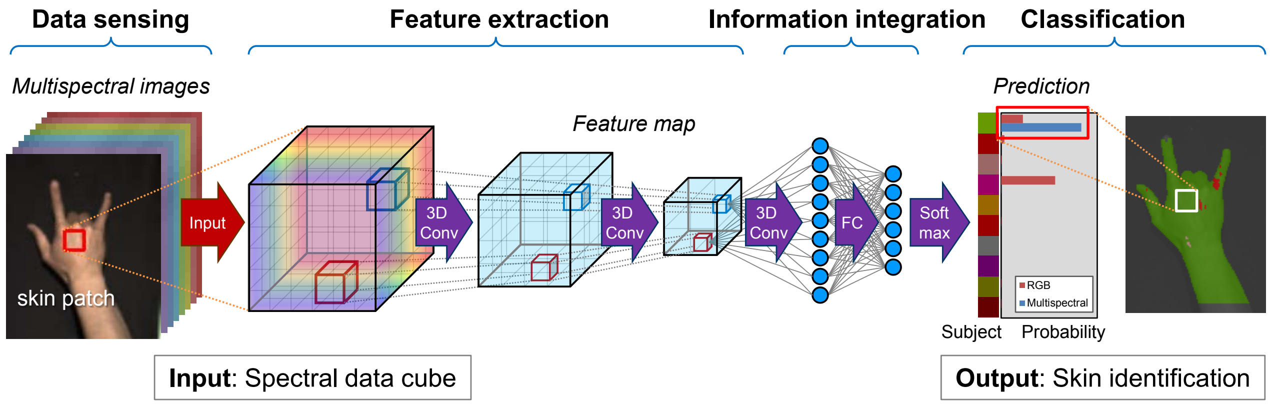

Figure 1. Skin-based user identification with local spatial-spectral features: We propose a novel framework for pose-invariant user identification by the combination of multispectral image data and an algorithm based on CNN.

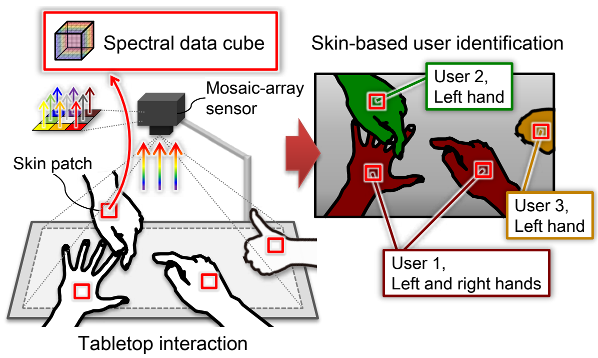

Figure 2. Use case in the scenario of a tabletop interface: In this case, users' hands are moving without any shape constraints and without any restriction in partial overlapping each other.

applications.

In this work, we are dealing with a mosaic-array based multispectral sensor where its spatial resolution is divided into spectral bands. Thus, by increasing the number of spectral bands we reduce the resolution of each band. This paper provides a framework for finding the best trade-off between the number of spectral bands and the spectral spatial resolution in order to maximize classification accuracy. In other words, given the fixed 3D spectral data cube volume (specified by sensor's basic resolution), our framework optimizes the shape of the cube with the same volume which provides the best classification accuracy. This optimization framework requires input from various configurations with different numbers of spectral bands. This would require data capture with multiple different multispectral cameras with the same basic sensor resolution, which is not practical. Instead, we propose to use a simulation approach, which we describe in detail in Section 4.

Contributions The key technical contributions of this paper are summarized below.

- Showing superiority of our approach with respect to conventional RGB image skin-based identification when using the same amount of data.
- Demonstrating feasibility of hand identification based on spatial-spectral features of skin using CNNs with synthetic and real datasets.
- Proposing a novel 3D CNN which enhances relevant spectral bands for skin-based identification.
- Providing multispectral image dataset generating pipelines for finding an optimal shape of spatial-spectral data cube.

Outline This paper is organized as follows. We begin with reviewing prior work in Section 2. In Section 3, we introduce our network architecture utilizing a multispectral data cube as input. We explain our strategy of generating synthetic datasets in Section 4 and show superiority of multispectral image input via multiple experiments with our synthetic datasets in Section 5. The feasibility with real multispectral data is shown in Section 6. In Section 7, we discuss supportive evidence of our proposal with an explanation tool for deep networks. Conclusions, limitations and future works are provided in Section 8.

2. Prior Work

Identification approach using hands: There are already many commercial biometric user authentication systems which require an image of hands. Most of them can be categorized into fingerprint, vein and geometric identification. As an example of fingerprints identification, in [32], the authors extracted ridge ending and ridge bifurcation of fingerprint as feature values. In [26], the authors claimed multispectral fingerprint image acquisition improved robustness against environmental and physiological conditions like bright ambient lighting, wetness, poor contact between the finger and sensor. In the case of vein authentication, vascular patterns are recognized by analyzing deoxidized hemoglobin absorption of near-infrared light in [39]. The other category is the physical dimensions of a human hand. In [5], a user identification approach using 25 geometric features of finger and palm was proposed. In [43], features of hand silhouette extracted by using independent component analysis showed satisfactory performance for groups of about 500 users. A user authentication system with RGB camera on multi-touch tables was proposed in [30]. The authors used a support vector machine classifier with features of palm width, finger length and breadth. In [13], a non-contact identification method with CNN was proposed. Here, users hold their hands in front of a ToF camera and they are classified by shape features from their palm.

Multispectral image capturing system: A traditional multispectral imaging system operates in a sweeping manner and utilizes a prism or grating to disperse light [25]. The next category of multispectral imaging system employs either liquid-crystal tunable filters or acousto-optic tunable filters to modulate the input spectrum over time [14]. Recently, as the newest category, single-snapshot multispectral imaging systems are being developed to rapidly acquire a 3D spectral data cube which allow to avoid motion artifacts and thus enabling video acquisition [41, 42, 21, 15]. However, this category of mosaic-array multispectral imaging system usually sacrifices its spatial resolution for spectral resolution. To overcome this problem, some papers recently proposed the framework of combining image sensor architecture and image signal reconstruction [38, 12].

Potential of skin spectra for user identification: Early works have shown that skin has much personal information. The history began with the famous skin model in [2]. In [37], the authors proposed a novel skin model with two-region chromophore fitting and estimated consistency of pigments such as melanin, oxy- and deoxy-hemoglobin, by measuring the spectra of skin optical properties of 18 subjects of different skin phototypes I–VI [35] in the range from 500 to 1000 nm. In another work [40], the authors showed that absorption spectra and scattering spectra properties of skin sub-surface scattering are very different among 149 subjects.

3. Network architecture for hand identification

Recently, CNNs using 3D convolutions have been successful in various applications [10, 8, 23] with high dimensional data. In our skin identification task, with a spectral data cube, 3D convolution is expected to extract spectral-spatial features more efficiently than 2D convolution. The proposed network architecture is shown in Figure 3. Although any architecture can be used as the base network for extending to 3D, we selected the wide residual networks (Wide-ResNet) [44] because the ResNet architecture and its variants are commonly used in image classification field. The difference from a normal 3D Wide-ResNet is that

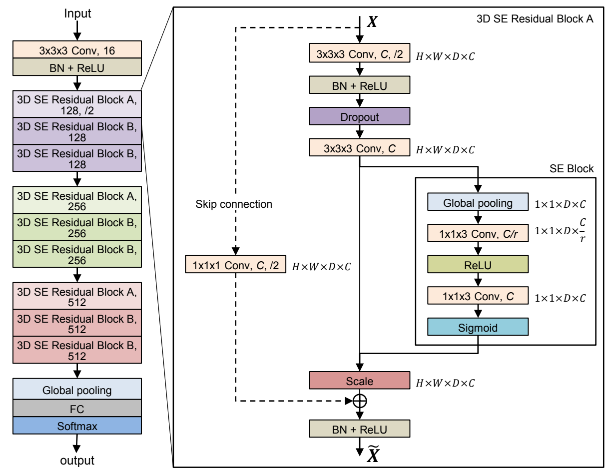

Figure 3. Network architecture: Our architecture is based on the Wide-ResNet [44] which has two types of residual blocks with different skip connections (left). These skip connections of block A and B are the projection shortcut and the identity shortcut respectively. We used a 3D convolution kernel and extracted the characteristics of a spectral band. In addition, to enhance relevant spectral bands, we added squeeze-and-excitation (SE) blocks (right).

spectral attention is involved to enhance the relevant spectral bands. It was inspired by the squeeze-and-excitation block (SE-block) [17]. SE block enhances the performance with small computational effort, and pays attention to a single weight for each channel of the feature maps. Position-SE block for facial attribute analysis was proposed in [46], which focuses on highlighting the relevant spatial position. Unlike [46], our SE block pays attention to spectra as well as channels of feature maps.

In our implementation we replaced all 2D convolutions of Wide-ResNet by 3D convolutions, while keeping the dimension of spectral band constant until the last global pooling layer. SE blocks were applied to each residual module. Concerning the global pooling layer in the SE blocks, only spatial axes were averaged. At the end of the SE blocks, weights for each spectral band, as well as each channel of feature maps, were obtained.

4. Generating synthetic datasets

We need datasets for evaluating the identification performance in various types of input data cubes. Therefore, we provide two types of synthetic multispectral dataset generating pipelines. One of them is for creating datasets which include actual hand skin textures. The other is for evaluating performance with a large number of subjects. In this framework, 1D spectral profiles are converted to 3D data cubes with measured distributions of skin textures.

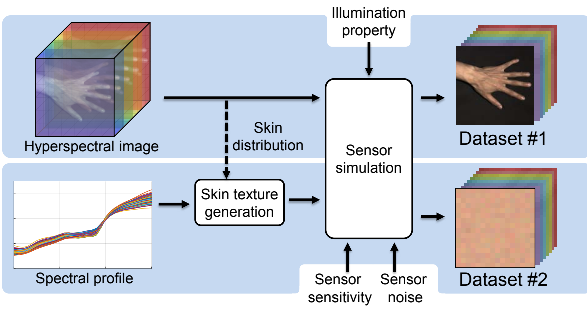

Figure 4. Pipelines of synthetic multispectral skin dataset generation: We constructed two types of framework. One of them is for creating datasets which include real hand skin textures. Another is a framework that converts 1D spectral profiles into 3D data cubes by incorporating measured distributions of skin.

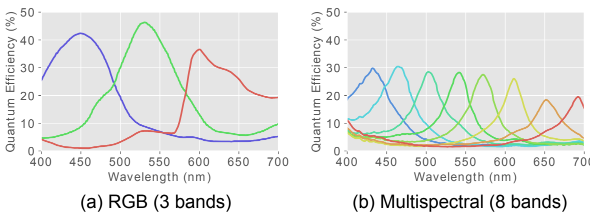

Figure 5. Sensor sensitivity characteristics: (a) represents the spectral profile of the acA2500-14gc of Basler AG. (b) shows the spectral profile of the multispectral camera CMS-C of SILIOS Technologies. Both cameras were used in the experiment described in Section 6.

4.1. Dataset #1 based on 2D spectral measurement

In order to create this dataset #1, we utilized a 2D spectrometer which has been prototyped internally in our affiliation. It acquires hyperspectral images in steps of 1 nm with 168x128 pixel resolution in the visible. 20 hands of skin from 12 subjects are in the source data and each of the hands is represented by 18 to 30 hyperspectral images with various poses. All subjects were Asian males. Here, we can emulate images according to a sensor specification. The upper line of Figure 4 shows the pipeline of generating our skin multispectral dataset #1. In general, an image capturing system with multiple wavelength channels is represented by equation 1:

$$I_c = \int_{400}^{700} R(\lambda)L(\lambda)S_c(\lambda)d\lambda + n \tag{1}$$

Where I_c means the intensity of spectral band c, λ is the wavelength over which is integrated and n is the noise. R is the spectral reflectance of a target in the scene and L is the spectral distribution of the illumination. Finally, S_c represents the sensor's spectral sensitivity of spectral band c. To acquire the correct reflectance R for each pixel, we had to normalize the illumination under which we collected the data. We captured the spectral response of a gray uniform board whose reflectance is known in advance. Then, we normalized the skin spectra by using equation 2:

$$R(\lambda) = M_s(\lambda) \oslash M_g(\lambda) \tag{2}$$

Where M_s and M_g denote the measured data of skin and the gray board. The symbol $\oslash$ represents the Kronecker division. We mainly assumed a white illumination which has a flat spectral distribution over all wavelengths as L. With the sensor spectral sensitivities S_c, we mainly assumed profiles as shown in Figure 5 (a) and (b) respectively for a multispectral and RGB sensor which were used in actual camera experiments in Section 6. The final stage of the pipeline consists of adding sensor noise. We adopted a noise model described in equation 3 and 4:

$$\hat{I}_c(x) = G(I_c(x), s_n(x)) \tag{3}$$

$$s_n(x) = \alpha \times \sqrt{I_c(x)} \tag{4}$$

$G(m, s)$ denotes a Gaussian distribution function with a mean value m and a standard deviation s. $\hat{I}_c(x)$ indicates the intensity in a patch image at position x and $s_n(x)$ is the standard deviation of noise. α is a noise scaling coefficient. We usually set $\alpha = 0.25$ as a base. In this case, the standard deviation s becomes 0.78% in a bright region of a 10-bit image. According to the procedure described above, we could generate a skin dataset #1 which was assumed to be captured by a multispectral sensor and an RGB sensor. Some examples of this dataset #1 are shown in our Supplementary Material.

4.2. Dataset #2 based on large scale spectral profiles

For generating a large scale skin multispectral dataset, using the pipeline of dataset #1 is a hard task. Especially collecting the source data with the 2D spectrometer is taking a lot of time. Therefore, we used the standard object color spectra database (SOCS) [19] as the source of our dataset. SOCS contains only 1D skin spectral profiles which were acquired at a point using a spectrometer. We picked up bare skin spectral profiles from the forehead of 123 Japanese females. These profiles are shown in Figure 6. We intended to generate a skin dataset #2 which was assumed to be captured by any formats of multispectral sensors. The pipeline is shown at the bottom of Figure 4. The biggest difference with the pipeline of dataset #1 is that R in equation1 has only spectral distributions, but does not have spatial distributions. Hence, we synthesized skin textures based on measurement of real skin. We measured the standard deviation of real skin s_r from the source of dataset #1. This standard deviation s_r is shown in Figure 6 as the yellow band. Then, we calculated the desired texture pixel number using equation 5:

$$\hat{R}(\lambda) = G(R(\lambda), s_r(\lambda)) \tag{5}$$

Here, we could acquire a 3D data cube $\hat{R}(\lambda)$ as input of the pipeline. The subsequent procedure is the same as with

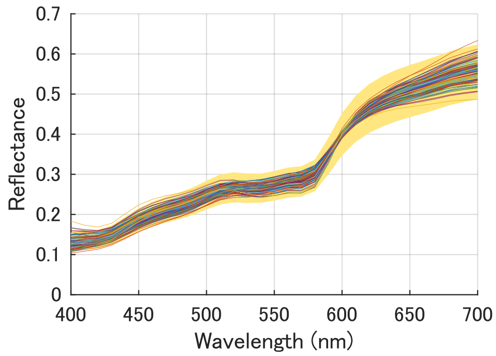

Figure 6. Skin spectral profiles in the SOCS dataset: 123 subjects' spectral skin profiles were picked up for generating dataset #2. The wide yellow band means a spatial standard deviations of real skin. Because distribution among individuals are less than a spatial distribution, it looks difficult for a human to distinguish the person from them.

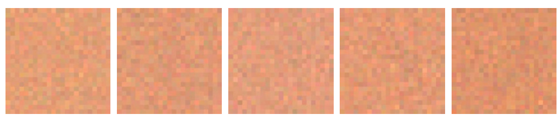

Figure 7. Samples of synthetic dataset #2: Multispectral data patches were generated from 1D spectral profiles from the SOCS dataset. Here, 16×16×3 (RGB) samples are shown. They look very similar and it seems to be difficult for a human to distinguish a subject.

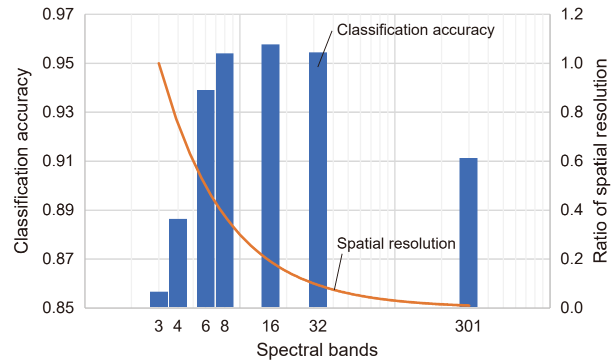

Figure 8. Performance comparison under same amount of data: The amount of data is defined by the multiplication of the number of spectral bands × the spatial resolution ratio. We evaluated performances under this trade-off conditions and found that at 16 spectral bands the combination of spectral and spatial information is optimal.

Section 4.1. We mainly generated patches with skin texture having a size of 16x16 pixels, but the size is flexible depending on a requirement of each experiment. Some samples of this dataset #2 are shown in Figure 7.

5. Evaluation on synthetic datasets

In this section, we analyzed identification performances with synthetic datasets which were generated as described in the previous section. To validate superiority of a multispectral image as input in various aspects, we evaluated on data trade-off conditions (Section 5.1), different numbers of classes (Section 5.2) and different noise conditions (Section 5.3). Finally, we compared the performance between 2D and 3D CNNs including our proposed network architecture (Section 5.4).

5.1. Data cube trade-off comparison

In the case of a mosaic-array sensor, the number of spectral bands and the spatial resolution are in a trade-off relationship, if keeping the amount of data (image width × image height × number of bands) constant. In this experiment, we evaluated performances among this trade-off conditions.

Experimental setup: We conducted this experiment with the generation pipeline of dataset #1. At first, 7 types of multi-band sensors which have respectively 3, 4, 6, 8, 16, 32 and 301 bands were generated. Their sensor sensitivities were defined by dividing the range $400 - 700\ nm$ into their band numbers equally (whose spectral profiles shaped into squares). An ideal white illumination was assumed. From the obtained multispectral images, we cropped 16x16xD (D=3, 4, 6, 8, 16, 32, 301) data cubes of hand skin regions which have been detected by thresholding in HSV color space in advance. In order to align the amount of data among cubes, they were reduced by skip down-scaling with the ratio $\sqrt{3/D}$ from all of their original data amount of 16x16xD. Finally, sensor noise was added with coefficient $\alpha = 0.25$ in equation 4. Then, data cubes were upscaled again by filtering for evaluation with the same network. We used the Wide-ResNet [44], which was implemented by [36]. Implementation details with network parameters are explained in our Supplementary Material. We split the created dataset into training and evaluation data in the ratio of 7 to 3. We randomly cropped 500 patches for each hand from the training data, and in the same manner, we prepared 215 patches per hand for the evaluation.

Analysis: Figure 8 shows the classification accuracy for each input. The classification accuracy is increasing with the number of spectral bands up to 95.7% at 16 bands. Basically, at 16 spectral bands the combination of spectral and spatial information is optimal. Even though the maximum classification accuracy is achievable at 16 spectral bands, 8 bands provide sufficiently good results. Thus, taking into account the current mutispectral cameras market, we decided to use SILIOS Technologies' CMS-C multispectral camera with 8 narrow spectral bands.

5.2. Performance scalability in large scale datasets

We show the scalability of performance related to the number of classes in the identification.

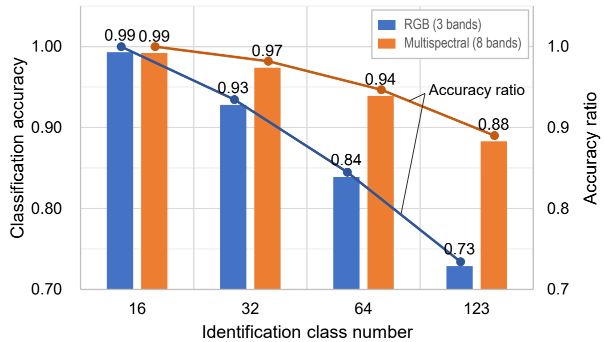

Figure 9. Classification accuracy with different number of subjects: The performance gap between RGB and multispectral input becomes larger with an increasing class number. Overall the identification performance with multispectral input remains more stable with varying number of classes.

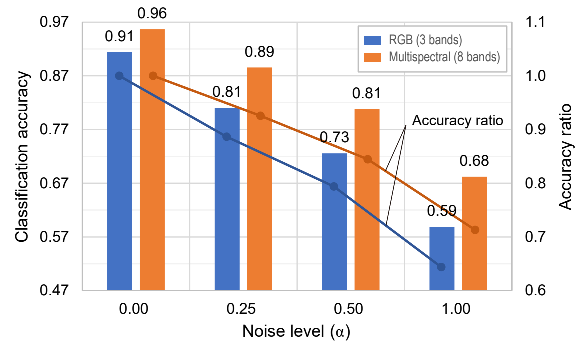

Figure 10. Classification accuracy at different noise levels: As noise level becomes bigger, the performance gap between with RGB and with multispectral input becomes larger.

Experimental setup: To facilitate this experiment, we prepared skin data cubes by utilizing the procedure as described in Section 4.2 which was used previously for generation dataset #2. In this experiment, we intended to emulate and compare existing sensors, therefore we adopted sensor sensitivities of the 3-band sensor (Figure 5 (a)) and the 8-band sensor (Figure 5 (b)). An ideal white illumination was assumed and the noise level was set to $\alpha = 0.5$ in equation 4. Then we got 16x16x8 data cubes for the 8-band sensor and 23x23x3 data cubes for the RGB sensor. Their data volumes are almost the same when considering the Bayer pattern [4] in RGB. We prepared sub-datasets with different numbers of subjects (16, 32, 64 and 123) based on the original dataset #2 containing 123 subjects. The number of patches for each subject were the same in the training and evaluation. The network architecture and its parameters were as in the previous experiment explained in Section 5.1.

Analysis: The results of classification accuracy are shown in Figure 9. In the case of 16 classes, high accuracy over 99.0% was achieved with both RGB and multispectral inputs. This means that this dataset with 16 classes might be simpler than the experiment in Section 5.1. However, the performance gap between them became larger as the number of classes increased. The accuracy with multispectral input was still kept 88.3% even in the case of 123 classes, though the accuracy with RGB was declined to 72.9%.

5.3. Robustness for noise

We also compared the robustness for sensor noise between RGB and multispectral input.

Experimental setup: Here, we regenerated dataset #1 with various noise levels. We set the noise scaling coefficient in equation 4 as $\alpha = 0, 0.25, 0.5, 1.0$, for creating different levels of noise. We used the same sensor sensitivities as in Section 5.2. Except for noise and sensor sensitivity, all other conditions were the same as in the experiment in Section 5.1. The network, its parameters and training procedures were also according to Section 5.1.

Analysis: The comparison with different noise levels is shown in Figure 10. The performance with multispectral input was superior to the one with RGB input at every noise level. Although both decreased, the performance gap increased with the noise level. This result implies that spectral information keep contributing to the performance, even when spatial information is degraded by noise.

Table 1. Classification accuracy with 2D and 3D based networks

Approach	RGB (3 bands) (%)	Multispectral (8 bands) (%)
2D CNN [44]	81.0	88.6
3D CNN	80.0	88.0
3D CNN with SE (Ours)	83.2	91.1

5.4. Comparison between 2D and 3D CNNs

In this experiment, we show the performance improvement by our network architecture described in Section 3.

Experimental setup: We used the same dataset with noise level $\alpha = 0.25$ which was generated in Section 5.3.

We compared the performances of 2D and 3D convolutions in Wide-ResNet architecture. Then, we evaluated the effectiveness of the SE-blocks which pay attention to relevant spectral bands.

Analysis: Table 1 shows the results obtained from RGB and multispectral input with different networks. In the comparison between different inputs in 2D CNN, multispectral input led to 7.6% improvement. However, from the comparison between 2D and 3D convolutions, 3D convolutions did not enhance the performance for each input. As mentioned in [20], this might due to the fact that 3D convolutions do not work well with insufficient number of training data. On the other hand, our proposal with multispectral input had 2.5% improvement from the 2D CNN. From this result, SE-blocks enhanced the performance of 3D CNN even with the insufficient amount of training data. Visualized results are shown in our Supplementary Material.

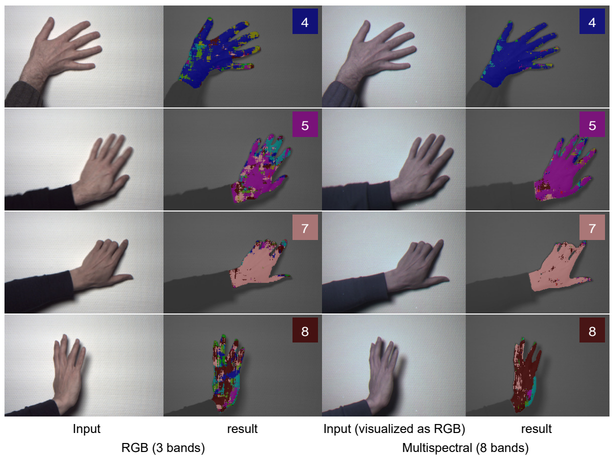

Figure 11. Result on the actual camera dataset of 10 hands: The first and second columns show the RGB inputs and the prediction results with 2D CNN. The third and fourth columns show the multispectral inputs and the prediction results with Ours. For each prediction result, each color means the predicted class and the label at the upper-right corner shows the ground truth color.

Table 2. Classification accuracy with the actual camera dataset

Approach	RGB (3 bands) (%)	Multispectral (8 bands) (%)
2D CNN [44]	88.3	91.5
3D CNN	87.9	92.5
3D CNN with SE (Ours)	89.1	93.1

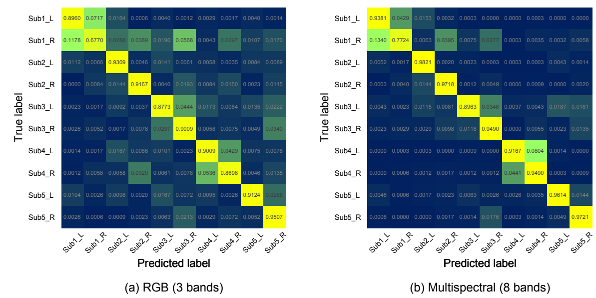

Figure 12. Confusion matrices comparison between (a) 2D CNN with RGB input and (b) Ours with multispectral input: More confusion among subjects as well as between hands of a subject are found in (a). Ours with multispectral input distinguishes among hands more accurately.

6. Evaluation on an actual camera dataset

To validate experiments with synthetic data in the previous section, we did an experiment with an actual camera setup. We made a real hand skin dataset which was acquired by a multispectral and RGB camera, and compared the identification performance achieved with each camera dataset against each other.

6.1. Experimental setup

We used a commercially available multispectral camera, the CMS-C of SILIOS Technologies, for the image acquisition. This camera can acquire 8 narrow spectral bands and a broad monochrome band. In our experiment, we used the 8 color spectral bands only. For the RGB image acquisition, we used the acA2500-14gc of Basler AG. Both cameras were synchronized by software with a framerate of about 5 frames per second. Their spectral sensitivities have been already shown in Figure 5. Other camera specifications and the experimental setup are explained in our Supplementary Material. We acquired images of both hands from 5 males, 1 Asian and 4 Caucasians. During the acquisition, the subjects moved their hands freely on the wall which was 0.8 meters away from the cameras. A conventional 100W bulb was used as a light source. As pre-processing before inputting to CNN, illumination normalization represented by equation 2 were applied to both datasets. In addition to that, we adjusted the field of view of a pixel by nearest neighbor resizing and the bit depth between both cameras by rounding RGB images. Measured noise levels were corresponding to $\alpha = 0.36$ in an RGB image and $\alpha = 0.43$ in a multispectral image on average of spectral bands.

Then, we split images of each hand into training data and evaluation data in the ratio of 7 to 3. In order to detect skin pixels, we preprocessed both RGB and multispectral images by transforming them into the HSV color space and identified skin pixels by a reference sub-color space. From the training data, we randomly extracted 8100 patches of size 16x16 pixels, centered around the detected pixels. In the same manner, we prepared 3471 patches per hand for the evaluation. We evaluated the performance with each input using the same three networks described in Section 5.4.

6.2. Result

Figure 11 shows selected results for comparing between the 2D CNN with RGB input and our proposed 3D CNN added SE with multispectral input. Since our goal is to identify a person from just a small skin area, this performance gap is considerable. Table 2 shows the quantitative results obtained from actual RGB and multispectral inputs with different networks. When comparing the performances between 2D CNN and 3D CNN without SE, multispectral based performance was improved with more than eight times the amount of training data to Section 5.4. On the other hand, RGB based performance was decreased. We believe that this result comes from 3D convolution, which benefit more from the relevant spectral information of multispectral data, and not only from the model capacity increase. We also confirmed that our proposed network led to more improvement by involving SE-blocks. The final performance gap was 4.8%. When considering the experimental setup of the class number and noise levels, this result is reasonable. Figure 12 shows another comparison with confusion matrices. This result also supports our conclusion

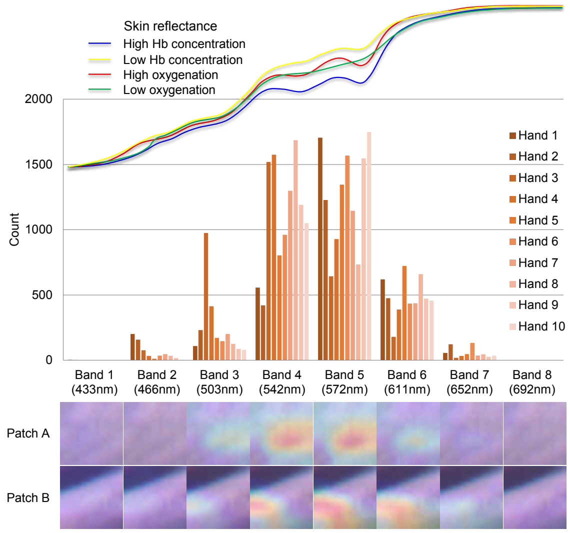

Figure 13. Analysis with Grad-CAM: The center bar graph shows a histogram of the most contributing spectral band which had the largest weight in 8 bands. The top plots show skin profiles with different underlying blood conditions. Bottom the series of images show some heat map examples of visualization.

that the experimental results with synthetic dataset was validated with the actual dataset.

7. Discussion with network explanation tool

In this section, we reveal the contributing factors to identification performances with Grad-CAM [31] which is a technique to produce visual explanations of decisions from a large class of CNN-based models.

Contribution degree of each spectral band: We can analyze contributing spectral bands of input. To facilitate this evaluation, we applied Grad-CAM to the 3D Wide-ResNet with SE model which was trained in Section 6. Then, we observed feature maps output from the last residual block. They had (No. of channels, height, width, No. of spectral bands) = (512, 4, 4, 8) dimensions and got weights for each spectral band. Figure 13 shows the histogram of the most contributing spectral band. From the aggregated result, we found that band #5 (center wavelength $\lambda_c = 572nm$) and #4 ($\lambda_c = 541nm$) contributed more than other bands. Some Grad-CAM visualizations as heat maps are shown in the bottom of Figure 13. We can see that the peak of contribution is located mostly within band #4 and #5.

Discussion: As shown above, there were significant bias among contributions of each input spectral bands. The plots in the top of Figure 13 are quoted from [9]. They are typical cases of skin reflectance with different conditions of the underlying blood. They indicate cases of high and low hemoglobin concentration as well as high and low oxygenation. [9] claims that a skin spectrum shapes into a "W" curve around $550nm$ due to underlying blood conditions which are depending on individuals. This is consistent with our Grad-CAM analysis, showing that spectral bands of this wavelength region were the most relevant for skin-based identification. Our 3D CNN with SE architecture enhanced the relevant spectral characteristics. Also, it could learn from the underlying hemoglobin more than other approaches. We consider that these are the main reasons for higher performance.

8. Conclusion

In this paper we have presented a novel framework for skin-based user identification, by combining multispectral imaging and CNNs. We showed superiority of our approach, with respect to conventional RGB imaging, when using the same amount of data. Feasibility of the approach was demonstrated with synthetic and actual image datasets. We proposed a novel 3D CNN model which is enhancing the influence of spectral image data. This paper is the first to involve SE-blocks for boosting relevant wavelengths. Additionally, we developed multispectral image dataset generation pipelines for finding an optimal shape of spatial-spectral data cubes.

Limitations and future work: One limitation is that we do not consider variations in illumination. A solution would be to add information from a spectrometer which senses the illumination spectra of a light source in a scene. Spectrometers recently became affordable devices, making this approach feasible. Another limitation is that we do not account for a sudden change of skin color due to sunburn or coloration by hand cream. Finally, the current work is restricted to the back-side of the hand. This is mainly due to the targeted use case scenario of a tabletop device. In principle, our approach can be also applied to the palm or other parts of the human body.

Optimizing spectral combinations of a mosaic-array sensor is the most interesting future work. Recently, there are several proposals which enable to customize capturing spectral bands such as [42]. We consider that our synthetic data generation framework and the CNN visualization technique described in Section 7 are valid for this work. Additionally, there should be many spatial clues of personal identification in infrared spectral bands, although our current camera configuration is limited to visible wavelengths only. We are looking forward to finding the best combination of bands and extending wavelengths for improving the performance of our identification framework.

Acknowledgement

We are grateful to our colleagues from Sony Europe B.V. and Sony Corporation for their fruitful discussions and support.

References

[1] Faisal AlGashaam, Kien Nguyen, Mohamed Alkanhal, Vinod Chandran, Wageeh W. Boles, and Jasmine Banks. Multispectral periocular classification with multimodal compact multi-linear pooling. *IEEE Access*, 5:14572–14578, 2017.

[2] R. R. Anderson and J. A. Parrish. The optics of human skin. *Journal of Investigative Dermatology*, 77:13–19, 1981.

[3] Apple Inc. Face ID, 2017.

[4] B.E. Bayer. Color imaging array. 1976. US Patent, US05685824.

[5] G Boreki and A Zimmer. Hand geometry: A new approach for feature extraction. *IEEE Workshop on Automatic Identification Advanced Technologies*, pages 149–154, October 2005.

[6] A. Burns and W. U. Bajwa. Multispectral imaging for improved liquid classification in security sensor systems. *Proceedings of SPIE Conference on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIV*, pages 1–7, Apr. 2018.

[7] Zach D. Caratao, Kelsey F. Gabel, Abijit Arun, Brett Myers, David L. Swartzendruber, and Christopher W. Lum. Micasense aerial pointing and stabilization system: Dampening in-flight vibrations for improved agricultural imaging. *2018 AIAA Information Systems-AIAA Infotech @ Aerospace AIAA SciTech Forum*, 2018.

[8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2017.

[9] Mark Changizi. The vision revolution: How the latest research overturns everything we thought we knew about human vision. *BenBella Books*, 2009.

[10] Lele Chen, Yue Wu, Adora M. DSouza, Anas Z. Abidin, Axel Wismüller, and Chenliang Xu. MRI tumor segmentation with densely connected 3d CNN. *Proceedings of SPIE Image Processing*, 2018.

[11] Valerie C. Coffey. Multispectral imaging moves into the mainstream. *Optics and Photonics News*, 23, 2012.

[12] K. Degraux, V. Cambareri, B. Geelen, L. Jacques, and G. Lafruit. Multispectral compressive imaging strategies using fabry-perot filtered sensors. *IEEE Transactions on Computational Imaging*, Pre-print, 2018.

[13] DArmin Dietz, Joachim Hienzsch, and Eduard Reithmeier. Contactless hand identification using machine learning. *CMBBE 2018, 15th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering*, 2018.

[14] N. Gat. Imaging spectroscopy using tunable filters: a review. *Proceedings of the SPIE*, 4056:50–64, 2000.

[15] Bert Geelen, Nicolaas Tack, and Andy Lambrechts. A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. *SPIE The International Society for Optical Engineering*, (1), 2014.

[16] E. Keith Hege, Dan O'Connell, William Johnson, Shridhar Basty, and Eustace L. Dereniak. Hyperspectral imaging for astronomy and space surviellance. *Proceeding of Imaging Spectrometry IX*, 5159, 2004.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Hiroshi Ishii. Tangible bits: beyond pixels. *Invited paper in Proceedings of the 2nd International Conference on Tangible and Embedded Interaction*, 2008.

[19] ISO/TR16066:2003. Graphic technology - standard object colour spectra database for colour reproduction evaluation (socs). *Technical report (International Organization for Standardization)*, 2003.

[20] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[21] Pierre-Jean Lapray, Xingbo Wang, Jean-Baptiste Thomas, and Pierre Gouton. Multispectral filter arrays: Recent advances and practical implementation. *Sensors*, 14(11):21626, 2014.

[22] G. Lu and B. Fei. Medical hyperspectral imaging: a review. *Journal of Biomedical Optics*, 19, 2014.

[23] Yanan Luo, Jie Zou, Chengfei Yao, Tao Li, and Gang Bai. HSI-CNN: A novel convolution neural network for hyperspectral image. *Proceedings of International Conference on Audio, Language and Image Processing*, 2018.

[24] Microsoft Corporation. Windows Hello face authentication, 2017.

[25] P. Mouroulis, R. O. Green, and T. G. Chrien. Design of pushbroom imaging spectrometers for optimum recovery of spectroscopic and spatial information. *OSA Applied Optics*, 39:2210–2220, 2000.

[26] R. Rowe, K. Nixon, and P. Butler. Multispectral fingerprint image acquisition. *Advances in biometrics*, pages 3–23, 2008.

[27] Inkyu Sa, Marija Popovic, Raghav Khanna, Zetao Chen, Philipp Lottes, Frank Liebisch, Juan I. Nieto, Cyrill Stachniss, Achim Walter, and Roland Siegwart. Weedmap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing*, 10(9):1423, 2018.

[28] Samsung Electronics Co., Ltd. Security - Iris Scanner, 2017.

[29] R Sanchez-Reillo, C Sanchez-Avila, and A Gonzalez-Marcos. Biometric identification through hand geometry measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1168–1171, 2000.

[30] Dominik Schmidt, Ming Ki Chong, and Hans Gellersen. Handsdown: hand-contour-based user identification for interactive surfaces. *NordiCHI*, pages 432–441, 2010.

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[32] Tatsuya Shimahara. Technologies for improving the speed and accuracy of fingerprint identification systems in support

of public bodies. *NEC Technical Journal*, 9(1):128–131, 2015.

[33] Sony Interactive Entertainment Inc. PlayStation4 User's Guide, 2013.

[34] Jim Spadaccini and Hugh McDonald. The evolution of tangible User interfaces on touch tables: New frontiers in UI & UX design. Technical report, Ideum, 2017.

[35] B. Thomas and MD Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Arch Dermatol*, 124(6):869–871, April 1988.

[36] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[37] Sheng-Hao Tseng, Paulo Bargo, Anthony Durkin, and Nikiforos Kollias. Chromophore concentrations, absorption and scattering properties of human skin in-vivo. *Optic Express*, (17):14599–14617, 2009.

[38] P. Wang and R. Menon. Computational multispectral video imaging. *Journal of the Optical Society of America A*, 35(1):189–199, 2018.

[39] Masaki Watanabe. Palm vein. *Encyclopedia of Biometrics*, pages 1027–1033, 2009.

[40] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (Proceedings on SIGGRAPH 2006)*, 25(3):1013–1024, 2006.

[41] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. Generalized assorted pixel camera: Post-capture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

[42] Sozo Yokogawa, Stanley P. Burgos, and Harry A. Atwater. Plasmonic color filtersfor cmos image sensor applications. *Nano Letter*, 12:4349–4354, 2012.

[43] Erdem Yoruk, Ender Konukoglu, Jerome Darbon, and Bulent Sankur. Shape-based hand recognition. *IEEE Transaction on Image Processing*, 15:1803–1815, 2006.

[44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *Proceedings of the British Machine Vision Conference*, 2016.

[45] David Zhang, Zhenhua Guo, and Yazhuo Gong. Multispectral Biometrics: Systems and Applications. *Springer*, 2015.

[46] Yan Zhang, Wanxia Shen, Li Sun, and Qingli Li. Position-squeeze and excitation block for facial attribute analysis. *The British Machine Vision Conference*, page 279, 2018.