# CONVOLUTIONAL NEURAL NETWORKS AND TRAINING STRATEGIES FOR SKIN DETECTION

*Yoonsik Kim, Insung Hwang, Nam Ik Cho*

INMC, Dept. of Electrical and Computer Engineering
Seoul National University, Seoul, Korea
nicho@snu.ac.kr

## ABSTRACT

This paper presents two convolutional neural networks (CNN) and their training strategies for skin detection. The first CNN, consisting of 20 convolution layers with $3 \times 3$ filters, is a kind of VGG network. The second is composed of 20 network-in-network (NiN) layers which can be considered a modification of Inception structure. When training these networks for human skin detection, we consider patch-based and whole-image-based training. The first method focuses on local features such as skin color and texture, and the second on the human-related shape features as well as color and texture. Experiments show that the proposed CNNs yield better performance than the conventional methods and also than the existing deep-learning based method. Also, it is found that the NiN structure generally shows higher accuracy than the VGG-based structure. The experiments also show that the whole-image-based training that learns the shape features yields better accuracy than the patch-based learning that focuses on local color and texture only.

***Index Terms***— skin detection, deep learning, convolutional network, CNN

## 1. INTRODUCTION

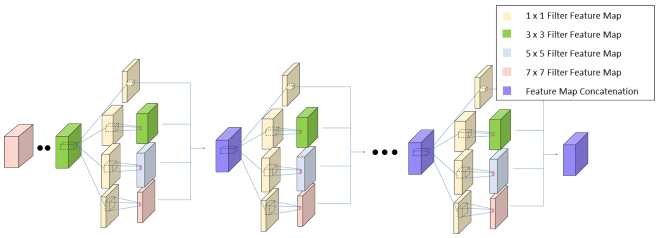
Skin detection is to find skin pixels or regions, which is one of the important pre-processing steps in many image processing and computer vision tasks. Specifically, it is effectively used for image enhancement [1], face and human detection [2], gesture analysis [3], pornographic contents filtering [4], surveillance systems [5], etc. Thus a large number of skin detection algorithms have been proposed, which is well summarized in [6–8]. According to these, the conventional methods usually locate skin pixels that fit to parametric models [2, 9], nonparametric models [4, 10], or some skin-cluster defined regions in certain color spaces [11]. There are also some methods that first detect shape features of human (hands, faces, and body) for finding skin pixels [8, 12], while most of the works detect skin pixels for finding human. More recently, a graph based method was proposed in [13] that represents the image by a multi-layer graph and then propagate the skin probability over the graph. In addition, there are neural network methods such as adaptive neural networks [14], self-organizing maps [15], and deep-learning based method [16] that adopts the auto-encoder. Even though there have been so many algorithms, the skin detection is still considered a challenging problem due to diverse variations such as change of illumination, skin color variations of races and makeup, and skin-like backgrounds.

Recently, the CNN has been shown to provide good performance on many image classification and processing problems, but we could find only a few researches that apply deep learning to skin detection problem [16]. In this paper, we propose two CNN architectures and also two training strategies for the skin detection. Specifically, the first architecture is a typical stacking of convolutional networks, and the second is the layers of NiN architecture inspired by the Inception [17]. For each of these CNNs, we apply two training strategies: one is the patch-based training where input and output to the CNN are the sets of image patches and corresponding skin labels respectively, and the other is the whole-image-based training where the input and output are whole images and corresponding label maps respectively. The result of inference is considered the skin probability at each pixel which is very useful in many image processing and computer vision tasks. Also, the skin probability map is thresholded to obtain a binary map of skin pixels.

It is expected that the whole-image-based training finds human related features such as shape of eyes and mouth in addition to local color/texture features that are well found by the patch-based method. Extensive experiments validate this expectation in that the whole-image-based training shows higher accuracy than the patch-based method. On the other hand, the patch-based method well suppresses skin-colored but non-skin objects as it concentrates on local texture information. Also, all the CNN methods proposed in this paper show better performance than the conventional methods.

## 2. PROPOSED METHOD

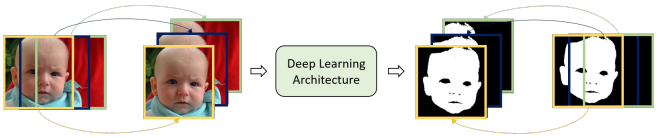
We propose two kinds of networks that are derived from widely used architectures: one is VGG-based network and




**(a)** NiN architecture.

**(b)** VGG based architecture.

**Fig. 1.** Illustration of CNN architectures for skin detection.




**(a)** Whole-image-based training.


**(b)** Patch-based training.

**Fig. 2.** Proposed training strategies.

the other is NiN architecture. For each of these networks, we propose two training strategies, namely whole-image-based training and patch-based training.

### 2.1. Network Structure

We have previously proposed an NiN structure in [18], which is applied to several image restoration and labeling problems such as compression artifacts reduction, semantic segmentation and skin detection. This NiN architecture is shown in Fig. 1a, where it can be seen that each layer consists of the cascades of $1 \times 1$ and larger convolutions. Since the architecture exploits $1 \times 1$ convolution like the Inception network, it requires less number of parameters than other general CNNs with the same depth.

The other network is inspired from VGGNet [19] which demonstrates the efficiency of consecutive $3 \times 3$ convolutions. More specifically, using small filter kernels in deep CNNs is effective and efficient in that the number of parameters is small, while the receptive field is sufficient when the architecture is deep. We use 20 hidden layers of $3 \times 3$ convolutions as shown in Fig. 1b, which will be called *VGG-based* architecture in the rest of this paper. In both of NiN and VGG-based networks, the number of hidden channels is set to 64. For these numbers of depth and channels, the NiN requires 236K parameters and VGG requires 660K regardless of input image size (or patch), because it is solely composed of convolution layers.

### 2.2. Image based training

Fig. 2a illustrates the whole-image-based training, where the *Deep Learning Architecture* in the figure is either of NiN or VGG-based network introduced above. This method is to feed the whole image (not the patches) as the input, for learning the human-related shape features as well as the color and texture of skin. However, it needs to be noted that not every image has the same size or aspect ratio, and it is difficult to train a very large network when the input image is so large. Hence, we fix the size of the network input to $N \times N$, and decimate the images so that the larger side (horizontal or vertical) is set to $N$ and the image is stridden into the direction of other side (vertical or horizontal).

The details of input and output to the networks and their sizes tried in this paper are summarized as follows:

**Input for the training**: Decimate the image such that smaller side (horizontal/vertical) is reduced to 50. Then the inputs to the network are the $50 \times 50$ stridden images into the other (larger side) direction where the stride size is 10 pixels.

**Label image for the training**: Decimated and stridden ground truth binary map (skin vs. non-skin) in the same way as the input.

**Input for inference**: Image that is decimated such that smaller side (horizontal/vertical) is reduced to 50.

**Final output**: Probability maps for the stridden inputs are obtained as the output, which are merged and interpolated to the original size. We can obtain a binary decision map by thresholding the probability map.

### 2.3. Patch based training

Fig. 2b illustrates the patch-based training, where the *Deep Learning Architecture* is also either of NiN or VGG-based architecture. The networks are trained with image patches at the input and their corresponding label patches at the output. Patches are extracted from the stridden windowing of input images, thus we can prepare hundreds of patches per one image. Unlike the whole-image-based training, the input is not decimated so that it is expected that the network learns high frequency texture of skin region.

The details of input, output and their sizes are summarized as follows:

**Input for the training**: Image patches with the size of $31 \times 31$ are extracted from the training images where the stride size is 20 pixels.

**Label image for the training**: Corresponding patches are extracted from the ground truth binary map.

**Input for inference**: Stridden patches from the input image.

**Final output**: Output patches from the network are merged

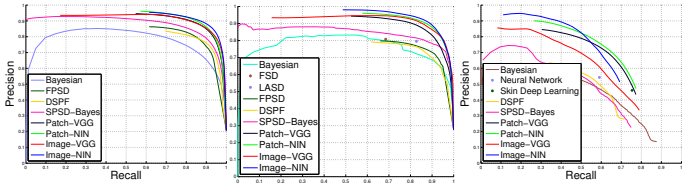



**Fig. 3.** Comparison of PR curves on ECU, Pratheepan and VT-AAST datasets (from left to right).

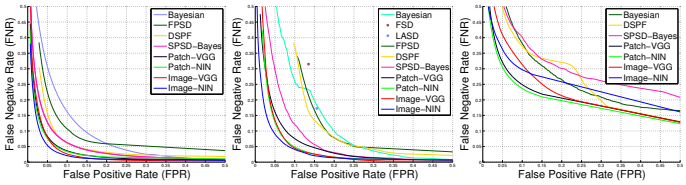



**Fig. 4.** Comparison of ROC curves on ECU, Pratheepan and VT-AAST datasets (from left to right).

together, and we obtain a probability map. We can also obtain a binary map by thresholding the probability map.

## 3. EXPERIMENTS

### 3.1. Experiment Setup

We use images in the ECU dataset [25] for training and validation, because it is the largest reliable set for the skin detection research. This set is composed of photos of people from various ethnic groups, and there are some photos with illumination variation. Also, there are some photos with complex non-skin but skin-colored regions. The dataset is composed of 4,000 images, where we use half of them for training and the others for the evaluation. We adopt the precision recall (PR) and Receiver Operating Characteristic (ROC) curves for evaluating the performance. Furthermore, we investigate the quality of binarized skin probability map in terms of four metrics: *Accuracy*, *Precision*, *Recall*, and *F-measure*. We also evaluate the methods with two additional skin datasets: Pratheepan [26] and VT-SSAT [27].

As we propose the image and patch based training methods for each of VGG-based and NiN structure, we have four approaches which are named as: *Patch-VGG*, *Patch-NiN*, *Image-VGG*, and *Image-NiN*. We compare our methods to the state-of-the-art methods like: Bayesian [4], FSD [20], LASD [21], Neural Network [24], Skin Deep Learning [16], FPSD [22], DSPF [23] and SPSD [13]. FSD uses the mix of dynamic threshold and Gaussian model, and LASD is a method based on luminance adapted color space. Neural Network is based on multi-layer skin classifier and Skin Deep Learning is based on the auto-encoder network. FPSD, DSPF and SPSD are based on the graph representation of an image, where initial skin (and non-skin) seeds are propagated over the graph in the manner of semi-supervised learning.

Our network is implemented with the Caffe tool box [28], and GTX 980 with 4GB memory is employed both for training and testing. There are some hyper parameters for the proposed network which are decided empirically. We set the learning parameters as: initial learning rate = 0.001, weight decay = 0.0002.

Table 1: Evaluation on ECU and Pratheepan datasets at peak F-measure.

| Methods | ECU dataset | | | | Pratheepan dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *Precision* | *Recall* | *F-measure* | *Accuracy* | *Precision* | *Recall* | *F-measure* |
| Bayesian [4] | 0.8910 | 0.7292 | 0.8220 | 0.7728 | 0.8237 | 0.6881 | 0.8972 | 0.7788 |
| FSD [20] | - | - | - | - | 0.8255 | 0.8077 | 0.6851 | 0.7414 |
| LASD [21] | - | - | - | - | 0.8361 | 0.7954 | 0.8275 | 0.8111 |
| FPSD [22] | 0.9106 | 0.7948 | 0.8534 | 0.8231 | 0.8419 | 0.7387 | 0.8991 | 0.8070 |
| DSPF [23] | 0.9190 | 0.7713 | 0.8864 | 0.8249 | 0.8521 | 0.7543 | 0.8436 | 0.7964 |
| SPSD [13] | 0.9306 | 0.8085 | 0.8805 | 0.8430 | 0.8782 | 0.7659 | **0.9328** | 0.8412 |
| Patch-VGG | 0.9479 | 0.8577 | 0.8913 | 0.8742 | 0.9299 | 0.8563 | 0.8750 | 0.8655 |
| Patch-NiN | 0.9492 | 0.8696 | 0.8923 | 0.8808 | 0.9334 | 0.8802 | 0.8972 | 0.8886 |
| Image-VGG | 0.9486 | 0.8499 | 0.9037 | 0.8760 | 0.9313 | 0.8577 | 0.9069 | 0.8816 |
| Image-NiN | **0.9562** | **0.8720** | **0.9122** | **0.8917** | **0.9484** | **0.9003** | 0.8912 | **0.8957** |

Table 2: Evaluation on the VT-AAST dataset at peak F-measure.

| Methods | *Accuracy* | *Precision* | *Recall* | *F-measure* |
|---|---|---|---|---|
| Bayesian [4] | 0.8798 | 0.4857 | 0.5740 | 0.5262 |
| Neural Network [24] | 0.8856 | 0.5426 | 0.5907 | 0.5656 |
| Skin Deep Learning [16] | 0.8881 | 0.4605 | **0.7538** | 0.5717 |
| FPSD [22] | 0.8918 | 0.5333 | 0.5583 | 0.5455 |
| SPSD [13] | 0.8871 | 0.5136 | 0.5477 | 0.5301 |
| Patch-VGG | 0.9243 | 0.6850 | 0.6455 | 0.6647 |
| Patch-NiN | **0.9272** | 0.7001 | 0.6539 | **0.6762** |
| Image-VGG | 0.9103 | 0.6225 | 0.5798 | 0.6004 |
| Image-NiN | 0.9249 | **0.7183** | 0.5826 | 0.6434 |

### 3.2. Experimental Results

Figs. 3 and 4 show PR and ROC curves of the methods stated above, where it can be seen that our four CNNs show better performance than others. Also, we can see that it is difficult to tell which one is the best among our four skin detection schemes. Tables 1 and 2 compare the methods by Accuracy, Precision and Recall when the threshold is set to maximize the F-measure, which also show that the proposed CNNs yield better performances than the conventional methods. From the figures and tables, it can also be seen that the results on VT-AAST dataset are worse than the results on other datasets, because the VT-AAST is composed of 4K images while we train the networks with the image in the ECU set that contains only $640 \times 480$ images, which is further reduced in the case of image-based training. Hence, in the case of whole-image approach, the 4K images are unduly decimated such that small regions in the images disappear. On the other hand, since the patch-based method does not decimate the image, it shows better performance in the case of VT-AAST dataset. For the ECU and Pratheepan dataset, which contain the same or similar sized images as the training data, the image-based training

**Fig. 5.** Visual comparison with other methods on the ECU2 dataset: (from left top to right bottom) input, Bayesian, FPSD, DSPF, SPSD, Patch-VGG, Patch-NIN, Image-VGG, Image-NIN, and ground truth images.

usually shows better performance than the patch-based because it learns the shapes as well as color features.

Subjective comparisons are also presented in Figs. 5 and 6, which also show that the proposed methods provide more plausible results regardless of illumination variation (for example see the illumination variation over the baby's face in the third row of Fig. 5), and the occlusion by skin-colored materials (see the false positive errors in the first and second row images). More specifically, the image based method is robust to illumination variation because it learns the shapes of human factors such as face, eye and mouth. On the other hand, since the patch based method learns the skin texture better, it can reject skin-colored but non-skin regions better than the image based method as evidenced by the first and second row of Fig. 5. It shows that the skin colored towel is rejected by the patch based methods but not by the image based methods.

## 4. CONCLUSIONS

We have proposed four skin detection schemes based on the CNN. Specifically, we have proposed two deep neural network architectures: VGG and NiN based ones, and we have also proposed two training schemes: patch-based and whole-image-based methods. From the extensive experiments, we found that the proposed CNNs outperform the conventional methods that are based on color models, graph representation, and auto-encoder based deep learning method. Also, we could see that the NiN architecture generally works better than the VGG network for the skin detection, though the NiN needs less parameters than the VGG. Comparing the training methods, the whole-image-based and patch-based training have their own strong and weak points. Specifically, the whole-image-based training finds the human shape features better so that it is robust to illumination and color variations, while the patch-based method finds skin texture very well so that it can reject the skin-colored background when it has different texture from the skin.

**Fig. 6.** Visual comparison with other methods on the Pratheepan dataset: (from left top to right bottom) input, Bayesian, FSD, LASD FPSD, DSPF, SPSD, Patch-NIN, Image-NIN, and ground truth images.


## Acknowledgement

This research was supported by Projects for Research and Development of Police science and Technology under Center for Research and Development of Police science and Technology and Korean National Police Agency (PA-C000001).


# References


[1] B Zafarifar, EB Bellers, and PHN de With, "Application and evaluation of texture-adaptive skin detection in tv image enhancement," in *IEEE International Conference on Consumer Electronics*, 2013, pp. 88–91.

[2] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

[3] Hardy Francke, Javier Ruiz-del Solar, and Rodrigo Verschae, "Real-time hand gesture detection and recognition using boosted classifiers and active learning," in *Pacific Rim Conference on Advances in Image and Video Technology*, 2007, pp. 533–547.

[4] Michael J Jones and James M Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.

[5] Zui Zhang, Hatice Gunes, and Massimo Piccardi, "Head detection for video surveillance based on categorical hair and skin colour models," in *IEEE International Conference on Image Processing*, 2009, pp. 1137–1140.

[6] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[7] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva, "A survey on pixel-based skin color detection techniques," in *Graphicon*, 2003, vol. 3, pp. 85–92.

[8] Simone Bianco, Francesca Gasparini, and Raimondo Schettini, "Adaptive skin classification using face and body detection," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4756–4765, 2015.

[9] Ming-Hsuan Yang and Narendra Ahuja, "Gaussian mixture model for human skin color and its applications in image and video databases," in *SPIE: Storage and Retrieval for Image and Video Databases VII*, 1998, vol. 3656, pp. 458–466.

[10] Rehanullah Khan, Allan Hanbury, and Julian Stoettinger, "Skin detection: A random forest approach," in *IEEE International Conference on Image Processing*, 2010, pp. 4613–4616.

[11] Jie Yang, Weier Lu, and Alex Waibel, "Skin-color modeling and adaptation," in *Asian Conference on Computer Vision*. Springer, 1998, pp. 687–694.

[12] Limin Liao, Hu Mei, Jianfeng Li, and Zhiliang Li, "Estimation and prediction on retention times of components from essential oil of paulownia tomentosa flowers by molecular electronegativity-distance vector (medv)," *Journal of Molecular Structure: THEOCHEM*, vol. 850, no. 1, pp. 1–8, 2008.

[13] Insung Hwang, Yoonsik Kim, and Nam Ik Cho, "Skin detection based on multi-seed propagation in a multi-layer graph for regional and color consistency," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017.

[14] Son Lam Phung, Douglas Chai, and Abdesselam Bouzerdoum, "A universal and robust human skin color model using neural networks," in *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*. IEEE, 2001, vol. 4, pp. 2844–2849.

[15] D A Brown, I Craw, and J Lewthwaite, "A som based approach to skin detection with application in real time systems," in *Proceedings of the British Machine Vision Conference*, 2001, pp. 51.1–51.10.

[16] Mohammadreza Hajiarbabi and Arvin Agah, "Human skin detection in color images using deep learning," *International Journal of Computer Vision and Image Processing (IJCVIP)*, vol. 5, no. 2, pp. 1–13, 2015.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[18] Yoonsik Kim, Insung Hwang, and Nam Ik Cho, "A new convolutional network-in-network structure and its applications in skin detection, semantic segmentation, and artifact reduction," *arXiv preprint arXiv:1409.1556*, 2017.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] Wei Ren Tan, Chee Seng Chan, Pratheepan Yogarajah, and Joan Condell, "A fusion approach for efficient human skin detection," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 138–147, 2012.

[21] Insung Hwang, Sang Hwa Lee, Byungseok Min, and Nam Ik Cho, "Luminance adapted skin color modeling for the robust detection of skin areas," in *IEEE International Conference on Image Processing*, 2013, pp. 2622–2625.

[22] Michal Kawulok, "Fast propagation-based skin regions segmentation in color images," in *IEEE International Conference on Workshops Automatic Face and Gesture Recognition*, 2013, pp. 1–7.

[23] Michal Kawulok, Jolanta Kawulok, and Jakub Nalepa, "Spatial-based skin detection using discriminative skin-presence features," *Pattern Recognition Letters*, vol. 41, pp. 3–13, 2014.

[24] Ming-Jung Seow, Deepthi Valaparla, and Vijayan K Asari, "Neural network based skin color model for face detection," in *Applied Imagery Pattern Recognition Workshop, 2003. Proceedings. 32nd*. IEEE, 2003, pp. 141–145.

[25] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai, "Skin segmentation using color pixel classification: analysis and comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 148–154, 2005.

[26] Pratheepan Yogarajah, Joan Condell, Kevin Curran, Abbas Cheddad, and Paul McKevitt, "A dynamic threshold approach for skin segmentation in color images," in *IEEE International Conference on Image Processing*, 2010, pp. 2225–2228.

[27] Abdallah S Abdallah, Mohamad Abou El-Nasr, and A Lynn Abbott, "A new color image database for benchmarking of automatic face detection and human skin segmentation techniques," in *Proceedings of World Academy of Science, Engineering and Technology*. Citeseer, 2007, vol. 20, pp. 353–357.

[28] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 675–678.