Deep Skin Detection on Low Resolution Grayscale Images

Marco Paracchini[a,**], Marco Marcon[a], Federica Villa[a], Stefano Tubaro[a]

[a]*Dipartimento di Informazione, Elettronica e Bioingegneria, Politecnico di Milano, Italy*

ABSTRACT

In this work we present a facial skin detection method, based on a deep learning architecture, that is able to precisely associate a skin label to each pixel of a given image depicting a face. This is an important preliminary step in many applications, such as remote photoplethysmography (rPPG) in which the hearth rate of a subject needs to be estimated analyzing a video of his/her face. The proposed method can detect skin pixels even in low resolution grayscale face images (64x32 pixel). A dataset is also described and proposed in order to train the deep learning model. Given the small amount of data available, a *transfer learning* approach is adopted and validated in order to learn to solve the skin detection problem exploiting a colorization network. Qualitative and quantitative results are reported testing the method on different datasets and in presence of general illumination, facial expressions, object occlusions and it is able to work regardless of the gender, age and ethnicity of the subject.

1. Introduction

Skin detection is an important preliminary task in a wide range of image processing problems. In particular this work is driven by the development of a remote PhotoPlethysmoGraphy (rPPG) application. This kind of applications aims at solving the problem of estimating the heart rate of a subject given a video stream of his/her face. Typically a signal, representing the time variation of the light intensity reflected by the skin, which is caused by the transition of blood in the vessel underneath the skin, is extracted averaging the color/intensity value on some selected pixels in each frame. This signal is then consequently analyzed in order to estimate the heart rate of the subject and/or other bio-medical measurements. Many rPPG applications (Rouast et al., 2017) estimate the face regions in which to extract the signal using a combination of classical face detection methods, such as Viola and Jones (2001), and fixed proportions in order to select specific parts of the face, e.g. typically the forehead. This procedure is not optimal since the skin in preselected parts of the face could not be visible due to occlusions of hair, wearable objects or other elements. Furthermore skin segmentation based on a predefined template suffers from errors of the face detection phase and/or due to intrinsic variance of face shapes. Moreover due to the high variability of the subject pose, motion blur, age, ethnicity, hair, facial hair, wearable objects, etc., the first step of a rPPG application (i.e. selecting the face region in which to extract the signal) is not trivial and errors in this step could heavily compromise the final hearth rate estimation. The majority of rPPG applications (Rouast et al., 2017) utilize a standard RGB camera, based on CMOS or CCD technologies, in order to acquire the video stream. The goal of this work is to propose a skin detection algorithm able also to work when applied to images acquired using SPAD (i.e. Single-Photon Avalanche Diode) array cameras. This kind of cameras is capable to detect even a single photon (Bronzi et al., 2016a), has extremely high frame rate (Bronzi et al., 2014) and has proved to be useful in a very large range of applications (Bronzi et al., 2016b), such as 3D optical ranging (LIDAR), Positron Emission Tomography (PET) and many others. In some rPPG works (Paracchini et al., 2019) SPAD cameras are used instead of traditional ones, where their high precision are useful in measure accurately the skin intensity fluctuations produced by the blood flow. On the other hand, due to the complexity of the SPAD sensor, this kind of cameras has a very small spatial resolution, 64x32 in Bronzi et al. (2014), and produces grayscale intensity image, since the low spatial resolution does not allow the use of Bayer filters. In this work we propose an automatic method, based on deep learning, with the aim of solving the task of detecting skin pixels in face images. Furthermore, the proposed method is de-

**Corresponding author:
e-mail: marcobrando.paracchini@polimi.it (Marco Paracchini)

signed to work with low resolution grayscale images such the one obtained using a SPAD array camera (Bronzi et al., 2014). The rest of the paper is organized as follows: in Sec. 2 a brief state of the art review on skin detection is reported highlighting the peculiarity of the problem addressed in this work; in Sec. 3 the proposed method is described while in Sec. 4 the training procedure that exploit transfer learning is illustrated; qualitative and quantitative results are shown in Sec. 5 and finally in Sec. 6 the contribution of this work are highlighted.

2. Related work

The skin detection problem is usually tackled using color information and exploiting the fact that skin-tone colors share some common properties defined in particular color spaces (Kawulok et al., 2014). After applying the optimal color space transformation it is possible to define rules to discriminate between skin pixels and other materials. Since this kind of methods are based on color information, they obviously require color images (RGB) to be applied to. As stated in Sec. 1, due to the choice of developing a method able to work with SPAD camera output (grayscale), this class of methods could not be applied in this specific problem. Moreover, they have no way to discriminate between the face and other body parts and this could be a problem in rPPG in which, due to the blood flow dynamic in the body, different body parts could carry different information (i.e. time-shifted signal). An extensive review of color based skin segmentation methods could be find in Kakumanu et al. (2007). Some skin detection methods able to work with grayscale images exist, e.g. Sarkar et al. (2017) , but they achieve good results only working with high resolution images since they learn local texture characteristics. Another problem, related to the one described in Sec. 1, is face parsing or face segmentation, which is the problem to analyze an input image of a face and densely segment it in different regions corresponding to different face parts and the background (Zhou et al., 2017). This is performed by labeling pixels in a dense fashion, i.e. to each pixel a label is assigned. In recent years, many deep learning methods have been proposed to solve this kind of problems, e.g. Liu et al. (2017), Nirkin et al. (2017) and Zhou et al. (2017), exploiting the promising results achieved by neural network based methods in semantic segmentation (Guo et al., 2018). Even though this problem is very similar to the one tackled in this paper (e.g. this last could be view as a simplified segmentation problem with just two classes, i.e. skin and other) some differences exist in the definition of the two problems. In fact, in face parsing methods, wearable objects such as glasses and sunglasses, or facial hair are not separated from the face region in which they are present, making this kind of methods not suitable for the skin detection problem. Moreover, methods such the ones proposed in Zhou et al. (2017) and Liu et al. (2017) work on high resolution color images. To the best of our knowledge no other method specifically designed to solve the skin detection problem on low resolution grayscale images exists in the state of the art.

Table 1. Pretraining network architecture (Baldassarre et al., 2017). Blue layers are used for transfer learning.

Encoder		Decoder	
Layer	Kernels	Layer	Kernel
Conv. (str. 2x2)	64x3x3	Conv.	128x3x3
Conv.	128x3x3	Upsamp. 2x2	
Conv. (str. 2x2)	128x3x3	Conv.	64x3x3
Conv.	256x3x3	Conv.	64x3x3
Conv. (str. 2x2)	256x3x3	Upsamp. 2x2	
Conv.	512x3x3	Conv.	32x3x3
Conv.	512x3x3	Conv.	2x3x3
Conv.	256x3x3	Upsamp. 2x2	
fusion			
Conv.	256x1x1		

3. Proposed method

As described in Sec. 2, deep learning based methods represent the state of the art for segmentation problem and they usually require a massive amount of data. On the other hand, due to the uniqueness of the skin detection problem, the amount of data available is very limited. For this reason a transfer learning (Pan and Yang, 2010) procedure was adopted. In particular, a colorization network (Baldassarre et al., 2017) was adapted for the skin detection task. The main reason behind the choice of exploiting a colorization method as a starting point for the proposed network is the empirical observation that a method of this kind applied to a grayscale image that depicts a face correctly colorize each skin pixel with the proper skin color. This means that the network must have learned a way to discriminate between skin pixels and pixels that depicts other objects. Furthermore, both the skin detection problem described in Sec. 1, and the colorization problem share the same kind of input, i.e. grayscale image. Moreover collecting training data in order to train a colorization network is trivial and the problem could be seen as a self supervised one. The driving idea is to propose slight changes to the colorization network in order to be able to transfer as much knowledge as possible from the colorization task to the skin segmentation one and then use a fine tuning approach.

3.1. Network topology

The colorization network presented in Baldassarre et al. (2017), whose architecture is reported in Tab 1, is based on a convolutional autoencoder with an auxiliary parallel branch. This additional branch, starting from the input image, exploits the first layers of a pretrained Inception-ResNet-v2 (Szegedy et al., 2016) in order to extract a vectorized representation of the image semantic. This vector is then merged to the encoded representation of the main branch before performing the decoding part. In particular this operation is performed to help the colorization method better understand the scene depicted in the input image, in order to colorize more precisely a large variety of objects and scenes. On the other hand, this auxiliary branch

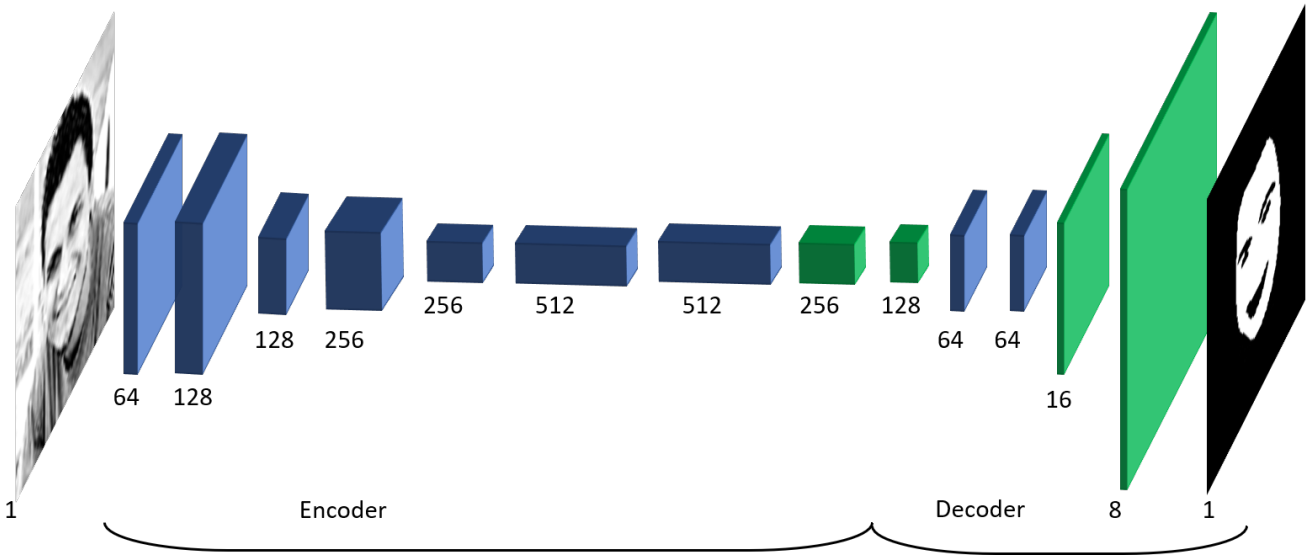

Fig. 1. Proposed network topology. The green layers and the last one are trained from scratch while for the blue ones the knowledge is transferred from a colorization network. The number under each layer indicate the dimension of its output (number of filters).

is totally unnecessary in the case that the input images are a-priori known to contain just a single human face. However its role was crucial in the Baldassarre et al. (2017) approach, in the proposed network this additional branch was completely removed, providing us with a suitable architecture. Another major difference between the proposed network topology and the one proposed in Baldassarre et al. (2017) resides in the output layer dimension. In particular, the original colorization network outputs a two channels image relative to the a* and b* channels of a L*a*b* (Robertson, 1976) color representation of the image (L being the luminance of the image, i.e. the grayscale input image). On the other hand, the proposed network needs to output a single channel image (called mask in the rest of the paper) with each pixel value $\hat{y}_{ij} \in [0, 1]$. This is achieved substituting the last activation function with a sigmoid function. In particular, for each pixel of the output mask, its value represents the probability attributed by the network of the input image having a skin pixel in that particular location. As reported in Fig 1, the encoding part of the network is composed by 8 convolutional layers, with 3x3 kernels and ReLu activation functions, and 3 max pooling layers in order to reduce the spatial dimension in the last encoding layer to 1/8 of the original input dimension. On the other hand, the decoding part is composed by 6 layers with 3x3 kernels and ReLu activation functions (except the last one, which is a sigmoid function in order to output values in $\in [0, 1]$) coupled with upconvolutional layers to increase back the spatial dimension to the input one. In Fig 1 the layers colored in blue are the ones trained propagating the colorization network knowledge while the other ones are trained from scratch. The central ones (with output depth 256 and 128) need to be trained with no prior information due to the removal of the colorization fusion layer. The next two have input and output shapes as in Baldassarre et al. (2017) so their weights value is propagated. Finally, the last ones, since are introduced to solve the skin detection problem, are randomly initialized.

4. Training procedure

As stated in Sec. 3 the training procedure, adopted to estimate the optimal network parameters, is based on a transfer learning approach. This implies that, before training the proposed network, the colorization network described in Baldassarre et al. (2017) needs to be trained on an appropriately chosen dataset. The same loss function and optimization algorithm as the ones described in the original work were used, i.e. mean square error, between the ground truth color image and the one reconstructed by the network, and Adam (Kingma and Ba, 2014) respectively; further information on the colorization network training are reported in Baldassarre et al. (2017). As a training set, on which to perform the colorization training step, a collection of color images depicting faces downloaded from the internet was added to the Labeled Face in the Wild (LFW) dataset (Huang et al., 2007) reaching 20000 images, doubled using data augmentation (horizontal flipping). In order to perform the colorization network training step, the colored images were used as the desired ground truth output while a grayscale representation of them were used as the network input. The model was trained exploiting the implementation described in the original paper (Baldassarre et al., 2017) which is implemented using synergistically Keras (Chollet et al., 2015) and Tensorflow (Abadi et al., 2015). After performing the training of the colorization method, the shared layers between the two networks (the ones colored in blue in Fig 1) were frozen (i.e. set to not trainable) and their weights value set to the corresponding one obtained from the colorization training step described above. The other ones were randomly initialized. The network was trained using Keras (Chollet et al., 2015), with Tensorflow (Abadi et al., 2015) as backend, with the Adam optimization algorithm (Kingma and Ba, 2014) and a learning rate of 0.0005. The loss function and the datasets used are described in Sec. 4.1 and Sec. 4.2 respectively. After a sufficient amount of epochs, 50, a fine tuning step was finally performed in which

all the layers were trained on the whole dataset and using the same training conditions, for an additional 100 epochs on the same training set.

4.1. Loss function

Regarding the loss function, the mean square error could be sufficient in order to train the network to perform the skin segmentation task. On the other hand, considering the main motivation that drives the building of this network (i.e. the rPPG application described in Sec. 1), false negative and false positive errors should not have the same weight in the loss function computation. In particular, in order to estimate the heart rate of a subject, it is not strictly necessary to consider all visible skin pixels whilst, on the other hand, labeling as skin a pixel depicting other tissues or materials could have an important negative impact on the final estimation. For this reason, given a predicted mask $\hat{y}$ obtained applying the proposed network to an input grayscale image x having a ground truth mask y with elements $y_{ij} \in \{0, 1\}$, we define the loss function as:

$$E(\hat{y}, y) = \sum_{ij} (y_{ij} - \hat{y}_{ij})^2 (\alpha \cdot y_{ij} + (1 - \alpha)(1 - y_{ij})) \qquad (1)$$

Where $\alpha \in [0, 1]$ is a parameter introduced in order to make E asymmetric. We choose a value for α smaller than 0.5, e.g. 0.4, in order to penalize false positive errors (i.e. $\hat{y}_{ij} = 1$ with $y_{ij} = 0$).

4.2. Datasets

To the best of our knowledge, there is no dataset available specifically created for the purpose of solving the facial skin segmentation problem. Some skin detection dataset exists, e.g. Tan et al. (2014), but they features images with multiple people and annotations with other body parts, which made them not usable for this particular problem. Moreover the number of images in this dataset is extremely low, e.g. 78 images are present in Tan et al. (2014), and insufficient to train deep methods. For this reason, we choose to adapt two already existing datasets, i.e. MUCT (Milborrow et al., 2010) and Helen (Zhou et al., 2013), consisting of RGB face images annotated with landmark locations, in order to produce facial grayscale images associated with skin masks. In particular, both datasets provide diversity in lighting, pose, age and ethnicity of the subjects. Moreover, the ones appertaining to the MUCT dataset are acquired in a controlled environment whilst the Helen ones are captured in the wild. A more detailed description on the processing performed on the two dataset is described in the following sections (Sec.4.2.1 and Sec.4.2.2 for the MUCT and Helen datasets respectively).

4.2.1. MUCT dataset

As described in Milborrow et al. (2010), the MUCT dataset consists of 3755 images (each one with a resolution of 640x480 pixels) captured from 276 subjects. Each image depicts a single face with a homogeneous blue background and it is associated with the pixel coordinates of 76 manually annotated facial landmarks. During the photo acquisition, in order to increase the dataset variety, five different camera views and three different lighting sets were used. The landmarks provided are relative to the lower face contour, eyes, eyebrows, nose and mouth. Starting from these landmark positions, for each image, a mask is produced considering a filled polygon shape with corners given by the jaw/chin contour points and the eyebrows upper contour. The eyes', eyebrows' and mouth's regions are consequently removed from the mask using the corresponding given contours. Unfortunately, as in the majority of facial landmark datasets, no upper face contour annotation is provided in this dataset (skin/hair contour). In order to extend the obtained masks to the forehead region a color similarity method has been used, exploiting the RGB channels information (the color information is indeed available in this preprocessing step for the creation of the dataset but is not available in the network training step). In particular, a rectangular region above the eyebrows is considered; each pixel in that region is clusterized in 3 sets using a K-means algorithm and using the Euclidean distance in the RGB space the pixel belonging to the hair or other occluding objects are rejected. This method, being automatic and based on color similarity, inevitably introduces some errors in the pixel labeling and produces worse results compared to manual annotation, which is unfortunately unavailable. Moreover in this dataset a binary information on the presence of glasses is provided although their position inside the image is not available. In order to remove the glasses region from the mask two rectangle of fixed size and centered around the eyes are subtracted from the mask. Lastly, in the original dataset facial hair is not labeled and in order to remove it from the mask a similar approach to the one adapted for the forehead region is performed in the lower part of the face and only on male subjects (gender labels are available in the original dataset).

4.2.2. Helen dataset

The Helen dataset features 2330 high quality, real world photographs of a large variety of people. Each image, each one with a different resolution (from less than 1 Mpixel up to 12 Mpixel), is densely annotated with landmarks locations. Moreover it has been used for face parsing works (Smith et al., 2013) in which an accurate face segmentation annotation for different part of the face has been provided. The masks need for the skin segmentation problem are simply built combining different segmentation regions. Unfortunately, in the Helen dataset many images feature more than one visible face while just one face is annotated in each image. Training a CNN on this data could compromise its performances due to a not consistent annotation. In order to avoid this problem a simple state of the art face detector algorithm (Viola and Jones, 2001) is run on each image of the dataset. Since even in presence of multiple faces the ground-truth annotation is always related to just one of faces a new images were created cropping the original images in regions centered around the annotated face. This step, being performed automatically, introduces inevitably some errors. Lastly, also in this dataset, a facial hair annotation is unavailable and the same method used for the MUCT dataset was implemented in order to remove beard regions from the masks.

Fig. 2. Example of some images in the proposed dataset with the skin mask superimposed, originally in the MUCT and the Helen datasets respectively.

4.2.3. Complete dataset

The complete dataset is built merging the two datasets obtained as described in Sec. 4.2.1 and Sec. 4.2.2 resulting in roughly 6000 grayscale face images (converted from the original RGB images) each associated with a skin labeling mask. Two examples of images in the final dataset are reported in Fig. 2, on the left an image originally in the MUCT dataset and on the left one from the Helen one; the ground truth skin mask is superimposed in pink. Moreover, in order to better approximate the test conditions (images coming from low spatial resolution devices, such as SPAD cameras) the grayscale images were downsampled to 64x64 (adding black border if necessary) and then upsampled to 128x128 using bicubic interpolation. The training/testing data split was obtained selecting 100 images (50 for each original dataset, randomly selected from MUCT and selected in the same way as Zhou et al. (2013) for Helen) for building the testing set. In order to ensure fair skin detection results, all the images belonging to the test set were checked manually and the annotations were corrected if needed. Subsequently, a horizontal flipped version of each training image is added in order to perform data augmentation. Finally, a validation set is created randomly selecting the 10% of the training set.

5. Results

The proposed method was trained as described in Sec. 4 using the training set described in Sec. 4.2.3. In this section some results are reported highlighting the necessity for the transfer learning approach and the accuracy of the obtained method, in Sec. 5.1 and Sec. 5.2 respectively.

5.1. Training with transfer learning

The learning curves of the skin detection network, obtained following the training procedure described in Sec. 4, are reported in Fig. 3. In particular, red curves are related to the loss error calculated on the whole training set at each epoch while blue ones are obtained on the validation set. As described in Sec. 4, following a transfer learning approach, in the first part of the training (first 50 epochs) the majority of the layers are kept frozen, as described in Sec. 3.1, preserving the weights value inherited from the colorization network, trained in a preliminary step as described in Sec. 4. This allow the network to

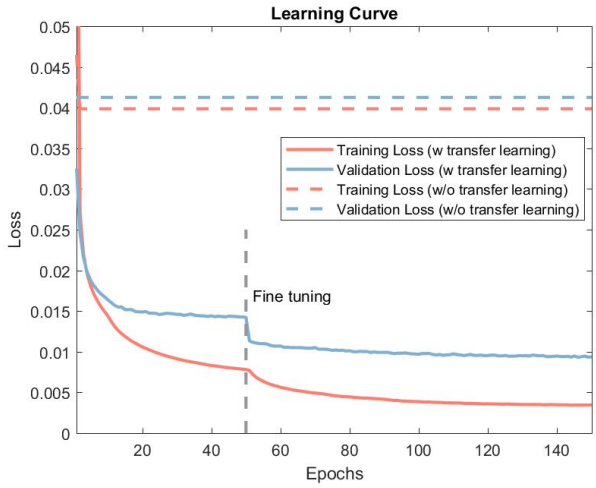

Fig. 3. Loss values during the training. Red lines represent loss values in each epoch on the training set while blue ones are obtained on the validation one. Dashed lines are the related to training directly on the skin detection problem with random initialization.

quickly adapt to the skin detection problem as can be seen in the first part of the solid red and blue curves reported in Fig. 3. On the other hand, since the colorization and skin detection problem are related but different, an additional fine tuning step is necessary in order to further specialize the network to solve the skin detection problem. The effect of the fine tuning is clearly visible in Fig. 3, in which both the solid curves have a sharp decay after the dashed vertical gray line (fine tuning begin point). The importance of the transfer learning approach could be also observed in Fig. 3, in which the red and blue dashed lines represents respectively the training and the validation loss obtained without using the colorization network wights as the initialization. In this case the training almost immediately collapses to the trivial solution of producing a masks with just zero values. Once the model reaches this point the training is not able to converge to other more interesting solutions. The same trivial result is obtained in all the training runs executed, regardless of the random initialization and hyperparameter settings. As can be observed from Fig. 3, a two steps approach, is able to drive the model training to a non trivial solution reaching a more interesting minimum point of the loss function.

5.2. Skin detection accuracy

5.2.1. Quantitative Results

The proposed method was tested on the 100 images test set described in Sec. 4.2.3 resulting in a test loss value of 0.012 between the output masks (values $\in [0, 1]$) and the ground truth ones (values $\in \{0, 1\}$). ROC curves for the per pixel skin classification task are reported in Fig. 4. In particular the proposed method achieved the best results on the test images originally belonging to the MUCT dataset (green line) given the less variability of the data, as described in Sec. 4.2.1. Considering the complete test set curve (red line), the best work point have a true positive rate (i.e. recall) of 89.8% with just 3.0% of false positive rate (i.e. fallout), thanks to the asymmetrical loss func-

Table 2. Comparison between the proposed method and Nirkin et al. (2017) based on intersection over union and F-score results obtained on MUCT, Helen and complete test set. The second line show results obtained combining Nirkin et al. (2017) and ground-truth masks in order to exclude eyes, eyebrows and mouth regions.

	IOU			F-score		
Method	MUCT	Helen	Complete	MUCT	Helen	Complete
Nirkin et al. (2017)	70	56	63	82	71	76
Nirkin et al. (2017) + GT	-	62	-	-	76	-
Proposed method	**78**	**69**	**73**	**87**	**81**	**84**

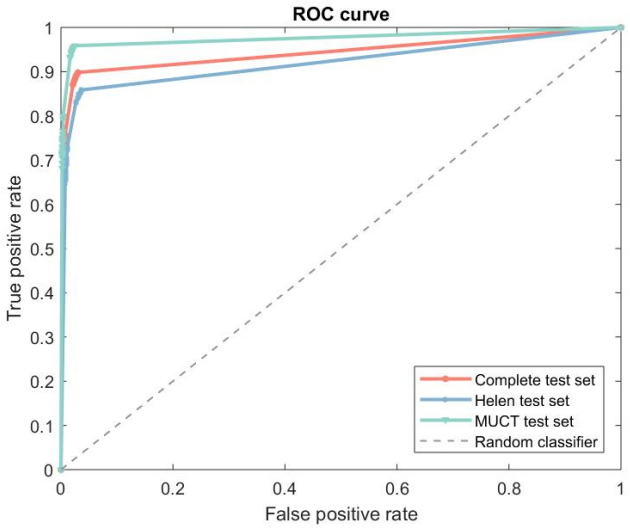

Fig. 4. Skin classification ROC curves obtained with the proposed method on the complete test set (red), MUCT test subset (green) and Helen test subset (blue).

tion defined in Sec. 4.1. As explained in Sec. 2, other methods for facial skin detection on grayscale low resolution images are rare or no existing. However a quantitative comparison between the proposed method and facial segmentation ones could be made. In particular we selected the facial segmentation method proposed in Nirkin et al. (2017) since as ours it can work with occluded faces. Since the method in Nirkin et al. (2017) produces masks that contain the eyebrow, eye and mouth regions we tested also the accuracy of this method combined with ground-truth information of these regions. In particular we removed from the mask obtained from Nirkin et al. (2017) the ground-truth mask of the unwanted regions (assuming so a perfect estimation of them). We compared the three methods (the one proposed by us and the one in Nirkin et al. (2017) not using or using ground-truth information) adopting the Intersection Over Union (IOU) and F-score metrics. In particular the F-score is defined as the harmonic mean between precision and recall. The results obtained are summarized in Tab.2; as can be observed even using the ground-truth information for the eyes, eyebrows and mouth regions, the proposed method produces more accurate results achieving a IOU of 73% and an F-score of 84% on the complete test set.

5.2.2. Qualitative Results

Some qualitative results with various images belonging to the test set are shown in Fig. 5 and Fig. 6, where the returned skin mask is superimposed to the input image using a pink color. Fig. 5 reports some results on images originally belonging to the Helen dataset while Fig. 6 shows other results using images initially in the MUCT dataset. As can be observed, the proposed skin detection method is able to produce qualitatively good results even in presence of non frontal faces, in-plane rotation, different head shapes and sizes, expressions, hair occlusions, glasses and other wearable objects. The beard is not always properly rejected, especially if it has an intensity similar to the subject skin. Moreover, in Fig. 7 some masks obtained with the proposed method are superimposed to some input images acquired by the SPAD array camera. These results are particularly promising since these images belong to a very different dataset in respect to the one used for training, even acquired with a different technology. As can be seen in Fig. 7, the network is able to generalize and produce good quality results even on images acquired in different conditions compared to the training dataset, and even in presence of different expressions, poses and heavy occlusions.

5.3. Real time performance

We evaluated the time performance of our method executing it on the test set described in Sec. 4.2.3 achieving an execution time of 6.6 milliseconds for each image corresponding to 152 fps. We obtain this result with a Tensorflow (Abadi et al., 2015) implementation of the network and executing it on a Nvidia Titan Xp® GPU.

5.4. Hidden layer output visualization

In Fig. 8 a visualization of the knowledge acquired by the network is reported. In particular, Fig. 8 visualizes the output of the decoder's second hidden layer when the network is run on an image acquired by the SPAD camera: the left picture in Fig. 7. As can be observed, after the training some filters specialized in detecting same particular facial feature relevant for the skin detection problem (e.g. eyes, 4th and 6th column on 5th row), the background (6th column on 1st and 2nd row) the face contour (5th column on 1st row) and finally the skin (6th column on 3rd row, 1st column on 5th row and 5th column 7th row).

Fig. 5. Some qualitative results on images in the test set belonging originally to the Helen dataset.

Fig. 6. Some qualitative results on images in the test set belonging originally to the MUCT dataset.

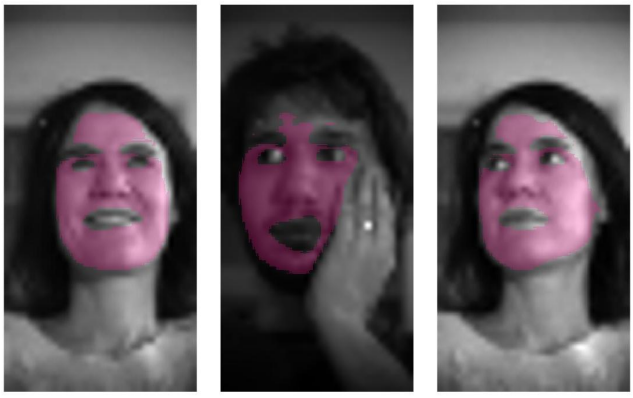

Fig. 7. Qualitative results on three face images acquired by the SPAD camera.

6. Conclusions

In this paper, we presented a deep learning based method proposed in order to solve the facial skin detection problem on low-resolution grayscale images, motivated by a rPPG application (as described in Sec. 1). Analyzing the state of the art of similar problems, in Sec. 2, we showed the peculiarity of the proposed problem and how, to the best of our knowledge, the method described in this work is the first being proposed specifically to solve it. Given the similarity between this problem and a semantic segmentation one, and the good accuracy achieved by neural network methods in this field, a deep learning based method was proposed. On the other hand, these kind of methods need massive amount of data to be trained on. Since the facial skin detection problem is very specific unfortunately not a huge amount of data are available for this specific problem. For this reason a transfer learning approach was adopted in the training phase. In particular, the proposed network architecture was chosen in order to have the majority of layers in common with the convolutional neural network proposed to solve the grayscale images colorization problem (Baldassarre et al., 2017). The similarities between the two problems are described in Sec. 3.1. As described in Sec. 4, the adopted transfer learning strategy was the following: firstly the colorization method was trained on a large dataset of unlabeled face images while the proposed network was subsequently trained starting from the colorization network weights and minimizing an asymmetric loss function, described in Sec. 4.1, on a novel dataset obtained using two freely available datasets (as described in Sec. 4.2). Lastly in Sec. 5.1 this training procedure has been justified showing that, without using it, it would be impossible to train the proposed network with the few data available. In addition, in Sec. 5.2 some quantitative results were reported providing accuracy evaluation for the proposed skin detection method and comparisons with a state of the art face segmentation method. Moreover, many network outputs were shown for both images acquired in similar conditions with respect to the ones used to built the training set and for images completely independent from the training set, acquired with the SPAD camera. Both these results show how the proposed method is able to achieve quantitative and qualitative good results in the skin detection problem even in presence of different poses, ages, expressions, ethnicity, wearable objects and other occlusions. The trained model along with image labels created by the author of this work are made available at the link `https://github.com/marcobrando/Deep-Skin-Detection-on-Low-Resolution-Grayscale-Images`.

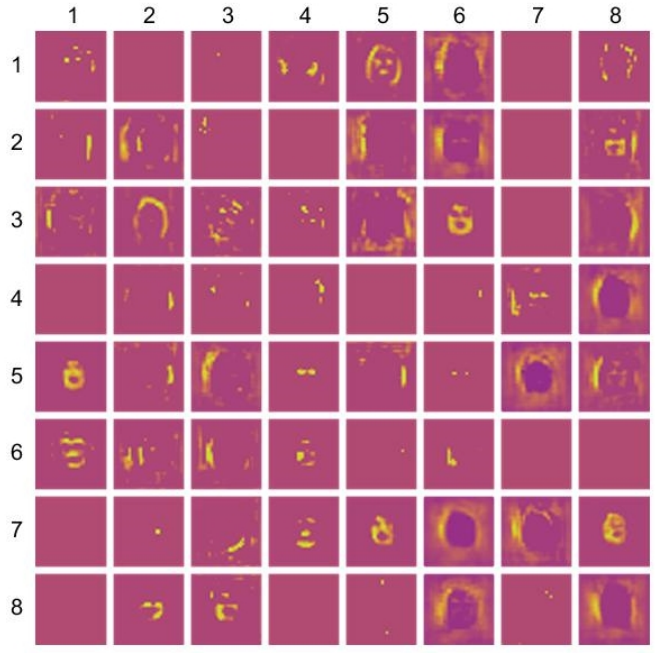

Fig. 8. Visual representation of the activations of the second hidden layer in the decoder stage when tested on a face image acquired by the SPAD camera.

Acknowledgments

Founding for this work was provided by the H2020 European project DEIS (EU grant agreement No 732242).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/. software available from tensorflow.org.

Baldassarre, F., Gonzalez-Morin, D., Rodes-Guirao, L., 2017. Deepkoalarization: Image colorization using cnns and inception-resnet-v2. ArXiv:1712.03400 URL: https://arxiv.org/abs/1712.03400.

Bronzi, D., Villa, F., Tisa, S., Tosi, A., Zappa, F., 2016a. Spad figures of merit for photon-counting, photon-timing, and imaging applications: A review. IEEE Sensors Journal 16, 3–12. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7283534, doi:10.1109/JSEN.2015.2483565.

Bronzi, D., Villa, F., Tisa, S., Tosi, A., Zappa, F., Durini, D., Weyers, S., Brockherde, W., 2014. 100 000 frames/s 64 x 32 single-photon detector array for 2-D imaging and 3-D ranging. IEEE Journal of Selected Topics in Quantum Electronics 20, 354–363. doi:10.1109/JSTQE.2014.2341562.

Bronzi, D., Zou, Y., Villa, F., Tisa, S., Tosi, A., Zappa, F., 2016b. Automotive three-dimensional vision through a single-photon counting spad camera. IEEE Transactions on Intelligent Transportation Systems 17, 782–795. doi:10.1109/TITS.2015.2482601.

Chollet, F., et al., 2015. Keras. https://github.com/fchollet/keras.

Guo, Y., Liu, Y., Georgiou, T., Lew, M.S., 2018. A review of semantic segmentation using deep neural networks. International Journal of Multimedia Information Retrieval 7, 87–93. URL: https://doi.org/10.1007/s13735-017-0141-z, doi:10.1007/s13735-017-0141-z.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49. University of Massachusetts, Amherst.

Kakumanu, P., Makrogiannis, S., Bourbakis, N., 2007. A survey of skin-color modeling and detection methods. Pattern Recogn. 40, 1106–1122. URL: http://dx.doi.org/10.1016/j.patcog.2006.06.010, doi:10.1016/j.patcog.2006.06.010.

Kawulok, M., Kawulok, J., Nalepa, J., 2014. Spatial-based skin detection using discriminative skin-presence features. Pattern Recogn. Lett. 41, 3–13. URL: http://dx.doi.org/10.1016/j.patrec.2013.08.028, doi:10.1016/j.patrec.2013.08.028.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980. URL: http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14.

Liu, S., Shi, J., Liang, J., Yang, M.H., 2017. Face parsing via recurrent propagation. CoRR abs/1708.01936.

Milborrow, S., Morkel, J., Nicolls, F., 2010. The MUCT Landmarked Face Database. Pattern Recognition Association of South Africa .

Nirkin, Y., Masi, I., Tran, A.T., Hassner, T., Medioni, G., 2017. On face segmentation, face swapping, and face perception. arXiv:1704.06729.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. on Knowl. and Data Eng. 22, 1345–1359. URL: http://dx.doi.org/10.1109/TKDE.2009.191, doi:10.1109/TKDE.2009.191.

Paracchini, M., Marchesi, L., Pasquinelli, K., Marcon, M., Fontana, G., Gabrielli, A., Villa, F., 2019. Remote photoplethysmography using spad camera for automotive health monitoring application, in: 2019 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE), pp. 1–6. doi:10.23919/EETA.2019.8804516.

Robertson, A.R., 1976. The cie 1976 color-difference formulae. Color Research & Application 2, 7–11. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1520-6378.1977.tb00104.x, doi:10.1002/j.1520-6378.1977.tb00104.x.

Rouast, P.V., Adam, M.T.P., Chiong, R., Cornforth, D., Lux, E., 2017. Remote heart rate measurement using low-cost rgb face video: a technical literature review. Frontiers of Computer Science URL: https://doi.org/10.1007/s11704-016-6243-6, doi:10.1007/s11704-016-6243-6.

Sarkar, A., Abbott, A.L., Doerzaph, Z., 2017. Universal skin detection without color information, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 20–28. doi:10.1109/WACV.2017.10.

Smith, B.M., Zhang, L., Brandt, J., Lin, Z., Yang, J., 2013. Exemplar-based face parsing, in: CVPR, IEEE Computer Society. pp. 3484–3491.

Szegedy, C., Ioffe, S., Vanhoucke, V., 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR abs/1602.07261. URL: http://dblp.uni-trier.de/db/journals/corr/corr1602.html#SzegedyIV16.

Tan, W.R., Chan, C.S., Pratheepan, Y., Condell, J., 2014. A fusion approach for efficient human skin detection. CoRR abs/1410.3751.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, pp. I–511–I–518 vol.1.

Zhou, F., Brandt, J., Lin, Z., 2013. Exemplar-based graph matching for robust facial landmark localization, in: IEEE International Conference on Computer Vision (ICCV).

Zhou, L., Liu, Z., He, X., 2017. Face parsing via a fully-convolutional continuous CRF neural network. CoRR abs/1708.03736.