

Multi-task Convolutional Neural Network for Patient Detection and Skin Segmentation in Continuous Non-contact Vital Sign Monitoring

Sitthichok Chaichulee¹, Mauricio Villarroel¹, João Jorge¹, Carlos Arteta², Gabrielle Green³,
Kenny McCormick³, Andrew Zisserman², and Lionel Tarassenko¹

¹ Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

² Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK

³ Neonatal Unit, John Radcliffe Hospital, Oxford University Hospitals Trust, Oxford, UK

Abstract—Patient detection and skin segmentation are important steps in non-contact vital sign monitoring as skin regions contain pulsatile information required for the estimation of vital signs such as heart rate, respiratory rate and peripheral oxygen saturation (SpO₂). Previous methods based on face detection or colour-based image segmentation are less reliable in a hospital setting. In this paper, we develop a multi-task convolutional neural network (CNN) for detecting the presence of a patient and segmenting the patient's skin regions. The multi-task model has a shared core network with two branches: a segmentation branch which was implemented using a fully convolutional network, and a classification branch which was implemented using global average pooling. The whole network was trained using images from a clinical study conducted in the neonatal intensive care unit (NICU) of the John Radcliffe hospital, Oxford, UK. Our model can produce accurate results and is robust to changes in different skin tones, pose variations, lighting variations, and routine interaction of clinical staff.

I. INTRODUCTION

The process of using a video camera to continuously compute estimates of vital signs such as heart rate, respiratory rate and peripheral oxygen saturation (SpO₂) in a real hospital environment presents several challenges. The detection of the presence of a patient in the video frame and the selection of a region of interest (ROI) from which vital signs can be estimated are of particular interest. The performance of these tasks is essential to the successful estimation of vital signs from the video camera [1, 2].

Over the past decade, most research in camera-based vital sign monitoring has been conducted in controlled environments in which a subject remains relatively still during video recordings [3–10]. Several research teams derived vital signs from the ROIs that were manually selected and fixed across the video sequence [3–7]. Many studies employed the automatic selection of ROIs using well-established face detection methods to define a face region and then use a feature tracker or an image segmentation algorithm to derive the ROIs in consecutive video frames [1, 8–10].

SC and JJ acknowledge the RCUK Digital Economy Programme, grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation). SC was supported by the National Science and Technology Development Agency, Thailand. MV was supported by the Oxford Centre of Excellence in Medical Engineering funded by the Wellcome Trust and EPSRC under grant number WT88877/Z/09/Z. GG and KM were supported by the NIHR Biomedical Research Centre Programme, Oxford. This work was supported by the EPSRC Programme Grant Seebibyte EP/M013774/1.

In a neonatal intensive care unit (NICU) environment, pre-term infants are active and clinical staff constantly interact with them. Their postures and positions are spontaneous. Many routine activities are undertaken several times a day such as feeding, changing infant position, recording vital signs, checking core temperature, administering medication, checking equipment and changing skin probes. Clinical staff or parents regularly take the infants out of the incubator for kangaroo care (cuddling to provide skin-to-skin contact). Lighting conditions change not only throughout the day but also depend on the season of the year, with long and bright days during summer and short and dark days during winter. Artificial light sources, such as room and overhead lights, cause reflections on the skin and change how the cameras record colour. Shadows occlude the baby when people walk by the incubator in which the infant is nursed. These scenarios pose challenges to the detection of the presence of a subject and the selection of ROIs for vital sign estimation, and so far conventional processing pipelines have not been suitable for these tasks.

An ideal method needs to correctly detect the presence of a subject in the video frame and segment skin pixels without relying only on skin-like colour. The better the skin segmentation algorithm processes, the greater the signal-to-noise ratio is when estimating vital sign signals. In addition, the estimation of vital signs is more reliable when fusing the signals obtained from different skin regions of interest [1, 2].

This paper presents a multi-task convolutional neural network (CNN) model that automatically detects the presence or absence of a patient and segments the patient's skin regions if the patient is found in front of the camera. The entire network was trained jointly for both tasks using a dataset of images with annotated skin regions provided by three human annotators. The network demonstrated fast and accurate results in challenging scenarios such as mixed-race subjects and low-light settings.

This paper begins by the introduction of a training dataset with its annotation protocol in Section II. Our proposed multi-task network and its training procedures are described in Section III. The performance of the model is reported in Section IV, followed by discussion in Section V. A conclusion is finally presented in Section VI.

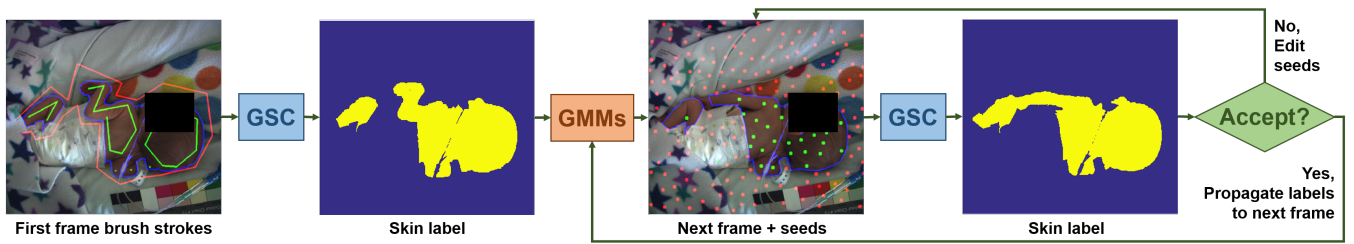


Fig. 1. Skin annotation workflow. For each recording session, an annotator was asked to draw *brush strokes* on non-skin and skin regions in the first image. The brush strokes were then used to compute a skin label using the GSC algorithm [11]. The skin label was propagated to the next frame using GMMs to generate *seeds* or simulated brush strokes (green and red circles). The annotator can interact with the seeds to modify the skin label.

II. DATASET

The main focus of this paper is to develop methods for continuous non-contact vital sign monitoring of pre-term infants [2]. We first describe the clinical study, followed by our approach to creating the dataset for training the algorithms.

A. Clinical study

The clinical study involves the double monitoring of 30 pre-term infants of less than 37 gestational weeks using both conventional monitoring and video recording. The video recordings were carried out in the high-dependency area in the NICU at the John Radcliffe Hospital in Oxford, UK. Each pre-term infant was recorded during daytime under ambient light for up to four consecutive days. No constraints were imposed on the infant's posture, position and orientation. This clinical study did not affect normal patient care.

The recordings were performed using a 3CCD video camera with three separated optical sensors for red, green and blue channels (JAI AT-200CL, JAI A/S, Denmark) mounted on top of a Giraffe Omnibed incubator (General Electric, USA) through a specially-drilled hole positioned approximately 30 centimetres away from the infant. The camera acquired 24-bit colour images with a resolution of 1620×1236 pixels at 20 frames per second. The study was approved by the Medical Research Ethics Committee under the reference number 13/SC/0597 (MONITOR Study). The recordings were carried out from February 2014 to May 2015. See [2] for more details about the clinical study.

The study protocol requires the development and training of algorithms on half of the dataset only, with the other half of the dataset to be used for testing and evaluating the performance of vital sign estimation algorithms. The work described in this paper uses the data from 15 pre-term infants chosen from balanced demographics. The patient demographics of these 15 infants and the total number of hours recorded are listed in Table I of Section III-C.

B. Semi-automatic skin annotation

Skin annotation is a process of labelling skin regions in images. It is not necessary to annotate every frame in the video recordings since multiple consecutive frames contain similar information. Although changes in lighting conditions during daytime can affect skin colour, sudden changes rarely happen. Thus, three annotators were asked to label skin regions in a subset of frames from the videos.

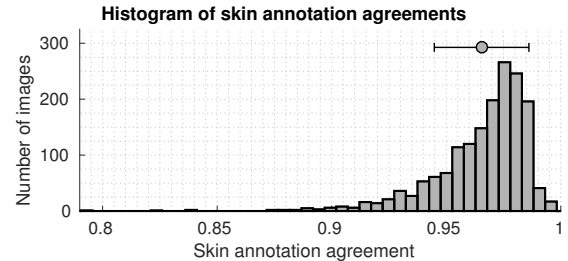


Fig. 2. Histogram of skin annotation agreements from three annotators on the positive images (mean agreement = 96.54%)

The skin annotation was performed using a semi-automatic approach to reduce the effort required for labelling the skin regions. For each video recording, the annotators were required to label the first frame, and the annotation was propagated to the next annotation frame. The algorithm for skin annotation was implemented based on the graph cut segmentation with geodesic star convexity (GSC) proposed by Gulshan *et al.* [11]. The flowchart describing the algorithm is shown in Fig. 1.

The GSC segmentation algorithm [11] exploits both colour and image gradient information for representing a shape of an object of interest. It requires *brush strokes* of non-skin and skin pixels in order to represent hard constraints for a segmentation procedure. These brush strokes were used to create colour models and exploit the object's structure. The final segmentation was obtained using energy minimisation subject to colour, boundary and shape constraints.

The propagation of the annotation to the next image was performed using Gaussian mixture models (GMMs), which learned the colour properties of non-skin and skin patches from previously segmented images. For the subsequent image, the models generated *segmentation seeds*, which are simulated brush strokes. The skin seeds were generated in regions with high skin probabilities, whereas the non-skin seeds were generated in regions with high non-skin probabilities. The location of the seeds was defined by Mitchell's best candidate algorithm [12]. The annotation can be altered by modifying the seeds if the annotators were not satisfied with the result.

C. Annotation protocol

The annotation was performed using three human annotators to ensure high-quality ground truth data. Considering the trade-off between annotation effort and lighting variations, images were sampled every 6 minutes from each video

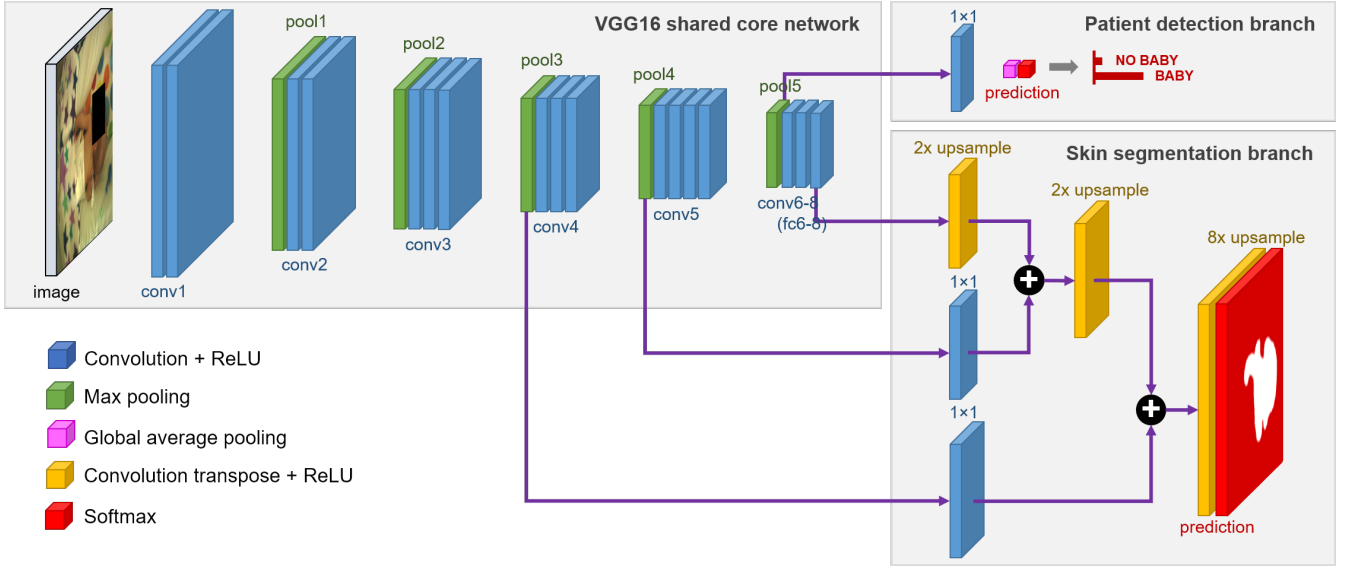


Fig. 3. The proposed CNN model extended the VGG-16 network [13] with two branches: the segmentation branch implements a fully convolutional network as described in [14]; the patient detection branch implements global average pooling over feature maps, similarly to [15, 16].

recording. Hence, a total of 2,269 images were obtained. The annotators were asked to annotate the same set of images at their original resolution.

For each session, the annotator was required to annotate the first image by providing brush strokes on non-skin and skin regions, and the skin annotation was then computed. For the next image, the annotator was automatically provided with segmentation seeds and skin annotation. The annotator could move on to the next image if the provided annotation was satisfactory. If not, he or she could interact with the seeds to modify the annotation. The annotator was asked to skip an image if: an infant was not present, the image contained the skin of clinical staff or parents, the scene was too dark to segment the skin or the infant was receiving phototherapy.

Even though the annotation was performed using the semi-automated approach, all images were reviewed and all skin labels were confirmed by the annotators.

D. Construction of the dataset

Our dataset consists of positive images with infant presence and pixel-level skin annotations and negative images without the infant present.

Positive Images: The skin annotations from three annotators were combined to form positive images. Images were regarded as positive if more than two annotators provided skin annotations without skipping the images. With this criterion, 1,718 out of 2,269 images (76%) were regarded as positive. For each image, pixels were regarded as skin if at least two annotators agreed, otherwise the pixels were regarded as non-skin. The skin annotation agreement was calculated using the ratio of the intersection of skin labels provided by at least two annotators to the union of skin labels provided by all annotators. The mean agreement is 96.54%. Fig. 2 shows the histogram of annotation agreements.

Negative Images: The negative images were also annotated by the same annotators using the images that were collected from periods, as reported in clinical records, during

which the infants were unlikely to be in the videos such as kangaroo care, camera covered, and quiet period (lights were dimmed or the incubator was covered to allow the infants to rest). For the 15 pre-term infants, these periods account for 23.5 of 226.4 hours of recordings. Images were taken every 20 seconds or 180 images per hour from these periods, which generated 4,227 images. All images were presented to three annotators. The annotators were required to identify each image as infant or non-infant. The images with two or more agreements on the non-infant class were regarded as *negative*. With this strategy, 2,885 negative images were obtained with only 12 disagreements.

III. METHODS

A. Patient detection and skin segmentation network

We formulated the problems as a joint task of image classification and image segmentation built into a single CNN model. Our multi-task network (see Fig. 3) has a shared core network with two branches: the patient detection branch implemented using global average pooling; and the skin segmentation branch implemented using hierarchical upsampling of feature maps across the shared core network.

Shared core network: Our multi-task network was built up on the VGG16 network proposed by Simonyan and Zisserman [13]. The VGG16 network is constructed of a stack of 3×3 convolution layers (`conv`) with non-linearities (`ReLU`) and periodically followed by 2×2 max-pooling layers (`pool`) for downsampling. The structure repeats until the output has a small spatial size and a decision is made on that output by fully-connected layers (`fc`) and a softmax layer for calculating class probabilities. Since the network is equipped with 5 max-pooling layers, the spatial size was reduced by a factor of 2^5 or 32 before reaching the fully-connected layers. The network was previously trained for image classification on 1.3 million images of the ImageNet dataset. It is recognised as a generic feature extractor and demonstrated good generalisation in transfer learning on other datasets [13].

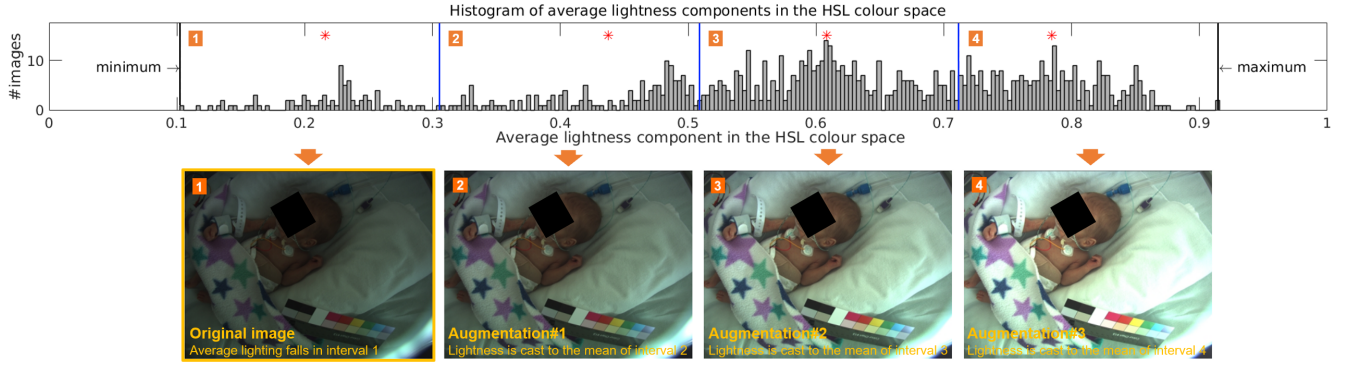


Fig. 4. Lighting augmentation was applied to generate more training images with different lighting conditions. (top) The histogram of average lightness components of all original images is divided into four uniform sections. The mean of each section is marked with a red asterisk (*). (bottom) For each original image, if its average lightness had fallen in one section, three more images are generated by scaling the lightness component using the scales calculated from the three other sections. The scale is defined as the ratio of the mean of each section to the image’s average lightness component.

Long *et al.* [14] proposed a modification to the VGG16 network. The original VGG16 network took a fixed-size image and produced softmax probability estimates over classes. Several adaptations were needed to enable the model to perform pixel-level segmentation. The three fully-connected layers in the VGG16 network (`fc6-8`) were converted into convolution layers (`conv6-8`) by having them perform convolution operations instead of inner products; therefore, the layers are able to produce a spatial output map with spatial coordinates preserved (see Fig. 3). The last convolution layer (`conv8`) was modified to produce 2-class outputs for non-skin and skin classes.

Patient detection branch: The patient detection branch was implemented using global average pooling for classification similar to Lin *et al.* [15] and Szegedy *et al.* [16]. The use of global average pooling significantly reduces the number of model parameters and forces feature maps to be meaningful confidence maps of each class before prediction [15]. A 1×1 convolution layer with 2 outputs was added on top of the `pool5` layer for reducing the dimensionality of the feature maps and enabling the prediction of the two classes, followed by a global average layer which averages out the spatial information resulting in an output vector fed to a softmax layer for calculating class probability estimates.

Skin segmentation branch: The skin segmentation branch was implemented using a fully convolutional network [14], which performs a series of spatial upsampling from cross-network feature maps in order to produce pixel-level labelling of skin regions (see Fig. 3). The implementation, again, followed that of Long *et al.* [14]. The feature maps of the `conv8` layer contain a coarse prediction or confidence maps for each class at a subsampling factor of 32. A finer prediction can be obtained by combining the prediction of the `conv8` layer with that of shallower layers. Firstly, a 1×1 convolution layer with 2 outputs was added on top of the `pool4` layer and the `pool3` layer in order to produce two additional predictions at higher resolutions (at a subsampling factor of 16 and 8 respectively). Next, the prediction from the `conv8` layer was spatially upsampled through a convolution transpose layer with a upsampling factor of 2 and then fused with the prediction of the `pool4` layer. Similarly, in the same

manner, the prediction fused from the `conv8` and `pool4` layers spatially was upsampled with a factor of 2 and then fused with the prediction of the `pool3` layer resulting in a combined prediction at a subsampling factor of 8. Finally, a convolution transpose layer with an upsampling factor of 8 was added in order to obtain a final prediction at the same spatial size as the input image. The network ended with a softmax layer that produces per-pixel class probability estimates, which are later thresholded to a skin label.

B. Network training

Data preprocessing: All training images and ground truth labels were resized to 512×512 pixels with aspect ratio maintained by allowing black spaces at the top and bottom of each image. This was done to ensure that the network and intermediate data could fit into the graphic card memory during the training of the network. Data augmentation, as explained below, was applied to introduce more variations in the training set. Finally, mean subtraction was performed to center the data around the origin.

Data augmentation: To reduce overfitting and improve the generalisation of the network, different variations of each original training image were generated. In our problems, the position and orientation of the infant were changed according to the clinical staff’s discretion and the camera’s position. In addition, the infant’s skin colour was varied by external lighting conditions. We employed three augmentation techniques:

- **Rotation:** To encourage the network to be robust to different rotations, seven more images were generated for each image by rotating the image at 45-degree increments between 0 and 360 degrees without resizing.
- **Flipping:** Since the human body is symmetric, two additional images were generated by flipping each original image horizontally and vertically.
- **Lighting variations:** To let the network learn about the variations in illumination, we generated three more images for each original image by scaling the lightness component in the HSL colour space (see Fig. 4).

Loss functions: Each branch was associated with an individual function. The patient detection branch was equipped with a multinomial logistic loss as in [15, 17]. The skin

TABLE I
SUMMARY OF PATIENT DEMOGRAPHICS IN THE D₁ AND D₂ SETS

Set	Subjects	Sessions	Hours	Gender ¹		Ethnicity ²						
				M	F	W	B	A	WB	WA	Other	
D ₁	8	22	118.7	5	3	5	1	1	—	—	1	
D ₂	7	21	107.7	3	4	5	—	—	1	1	—	
Total	15	43	226.4	8	7	10	1	1	1	1	1	

¹ M = Male, F = Female; ² W = White, B = Black, A = Asian, WB = Mixed White and Black, WA = Mixed White and Asian

segmentation branch was equipped with a multinomial logistic loss summed across the whole spatial pixel outputs and normalized with respect to the number of ground truth pixels, similarly to [14]. Since, in a video frame, the number of non-skin pixels is larger than that of skin pixels, we weighted the loss of the skin class according to the ratio of the number of true non-skin pixels to that of true skin pixels. The overall loss function can be seen as the weighted sum of these individual losses.

Model initialisation: Prior to the training phase, batch normalisation layers were added between a convolution layer and a ReLU layer in the shared core network. Batch normalisation can reduce overfitting and improve the generalisation capabilities of the network [18]. It can also facilitate the training by enabling the use of a high learning rate [18].

The model was initialised with the original VGG16's weights, which hold the accumulated knowledge on edges, patterns and shapes. All new weight layers, except for convolution transpose layers, were initialised using the Xavier algorithm [19] with zero bias. The convolution transpose layers were initialised with bilinear filters with no bias as suggested in Long *et al.* [14].

Training procedures: Our network was implemented on the MatConvNet framework [20]. The training is performed using standard Stochastic Gradient Descent (SGD) in two stages. In the first stage, the network was trained for the skin segmentation task using only the images containing the infant with annotated skin regions. The learning rates were scheduled to start at 10^{-2} and reduced by a factor of 10 for every two epochs until convergence, with momentum of 0.90 and batch size of 20. In the second stage, the network was trained jointly for both the patient detection and skin segmentation tasks using the whole dataset. The individual loss function of each task was weighted equally. The learning rate started at 10^{-4} and was decreased by a factor of 10 for every two epochs until convergence with momentum of 0.90 and batch size of 20.

After training, a computation sequence was arranged to evaluate the patient detection branch first, and then evaluate the skin segmentation branch only when the infant is present.

C. Evaluation protocol

An approach to obtaining predictive performance is to use cross-validation on two independent folds. The dataset was firstly divided into two sets, D₁ and D₂, such that one set has 8 subjects and the other set has 7 subjects. The assignment to different sets was based on the balance of skin phenotype,

gestational ages and the number of positive images. For each set, positive images were taken directly from the positive pool, and negative images were randomly sampled without replacement from the negative pool so that the number of positive and negative images were equal.

Table I shows the statistics of these two folds. A model was first trained on D₁ and validated on D₂. Then, another model was trained on D₂ and validated on D₁. The validation results from both models were combined to produce an overall predictive performance.

Evaluation metrics: In patient detection, the classifier's performance is described using accuracy, precision and recall values. In skin segmentation, a pixel-wise intersection over union (IOU), which is the standard metric for evaluating a segmentation algorithm, is used to describe segmentation performance. The IOU is defined as $\text{IOU} = (y_p \cap y_g) / (y_p \cup y_g)$ where y_p denotes a predicted segmentation result and y_g denotes a ground truth label. Since our work concerns skin segmentation, the IOUs of non-skin labels were ignored.

IV. RESULTS

We compared our model with colour-based skin colour filters, which were adopted in several non-contact vital sign estimation techniques [21, 22]. The filters were trained with three classifiers: Naïve Bayes, Random Forests and GMMs. They were trained on images in the HLS colour space with white balance correction applied to reduce variations in skin colour [23]. The skin filters classify each pixel as skin based solely on skin colour. Patient detection is performed using the ratio and the average probability of predicted skin pixels to make a decision, similarly to [24].

Table II summarises the performance of the patient detection and skin segmentation algorithms. The multi-task model trained with data augmentation outperformed all other models. For skin segmentation, compared to the best of the colour-based skin filters (Random Forests), the multi-task model with data augmentation gave results which were 3.1% better for pixel accuracy and 12.7% better for the IOU score. Fig. 5 shows the skin segmentation results of several images in the dataset for both approaches.

V. DISCUSSION

Qualitative assessment: Even though Random Forests yielded the highest performance among the other colour-based skin filters, the classifiers performed poorly in low light conditions (see Fig. 5). The multi-task CNN model correctly segmented skin regions under normal ambient lighting thanks to a reasonably large skin area in the image (15 – 20% of the image is skin pixels). In most of the negative images, the models did not produce skin labels when an infant was not present. The CNN model did not produce noisy and grainy skin labels as it processes the whole image at once.

The network had some difficulties in very low-light conditions, however its results were better than that of colour-based skin filters. It should be noted that the non-contact estimation of vital signs is also challenging under low ambient lighting due to low signal-to-noise (SNR) ratios [1, 10].

TABLE II
PERFORMANCE ON PATIENT DETECTION AND SKIN SEGMENTATION OF THE COLOUR-BASED SKIN FILTERS AND THE CNN MODELS

Model	Patient Detection			Skin Segmentation					
	Accuracy	Precision	Recall	Pixel Accuracy			Intersection over Union		
				Mean (SD)	Min	Max	Mean (SD)	Min	Max
Colour-based skin filters									
Naïve Bayes	98.60	97.82	99.42	89.49 (8.31)	32.70	98.94	61.34 (17.38)	4.29	92.88
Random Forests	97.67	97.45	97.90	95.00 (4.57)	57.68	99.32	75.87 (16.11)	6.83	95.40
GMMs	97.09	97.76	96.39	93.41 (5.20)	47.47	99.05	71.20 (14.16)	16.83	94.68
CNNs without data augmentation									
CNN for Patient detection	97.98	96.09	100.00	—	—	—	—	—	—
CNN for Skin segmentation	—	—	—	92.16 (3.37)	71.10	74.42	57.37 (15.21)	0.00	84.48
Multi-task CNN	98.22	96.57	100.00	96.16 (2.01)	75.89	98.89	77.21 (9.89)	4.84	92.86
CNNs with data augmentation									
CNN for Patient detection	97.09	96.02	98.25	—	—	—	—	—	—
CNN for Skin segmentation	—	—	—	97.92 (1.20)	88.71	99.46	87.82 (5.95)	49.37	96.54
Multi-task CNN	98.75	97.56	100.00	98.05 (1.90)	75.57	99.57	88.57 (7.45)	38.95	97.00

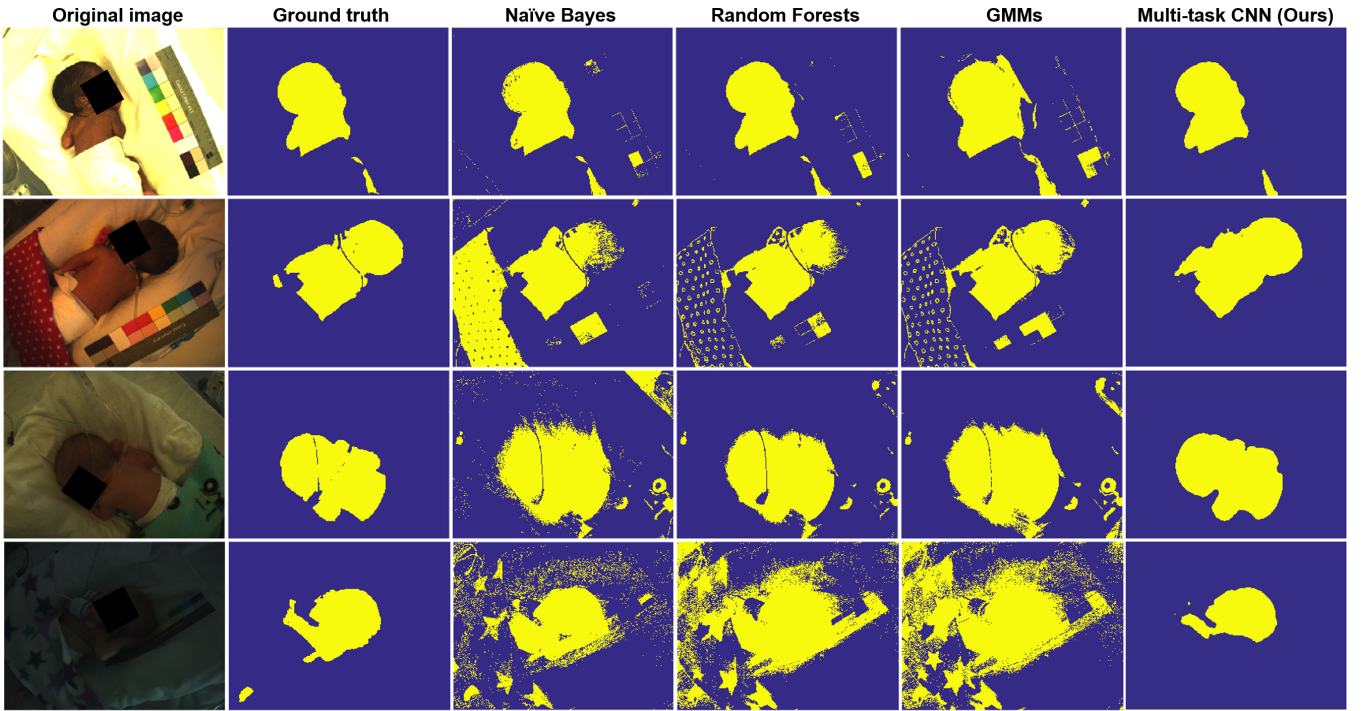


Fig. 5. Example results of skin segmentation using both colour-based skin filters and CNNs. All colour-based skin filters over-segmented skin regions by picking up false positives whose colours are similar to skin colour. The multi-task CNN model produced more accurate skin labels. As expected, the model cannot pick up small skin regions due to the CNN algorithm that periodically downsamples feature maps in the network. In low-light scenarios, the colour-based skin filters produced grainy segmentation results, whereas the CNN model under-segmented skin regions.

Effect of multi-task learning: The multi-task model exhibited slightly higher performance than the single-task models. However, the improvement may be due to the fact that the multi-task network was fine-tuned further with lower training rates for several epochs. As expected, the joint network did not show a bias towards an individual task. The multi-task network can perform both tasks two times faster than a cascade of two single-task networks.

Effect of data augmentation: Data augmentation was found to substantially improve the performance of the network (see Table II) as well as the quality of segmentation results (See Fig. 6). With data augmentation, 30.5% and 11.4% increases in the segmentation’s mean IOU score

were observed in the single-task and the multi-task models respectively. Without data augmentation, the CNN model only produced coarse segmentation results. This might be due to our dataset being too small for the network to learn a generic structure of the object of interest.

The performance of the patient detection algorithms trained with and without data augmentation was found to be similar. This is likely to be because patient detection is a much simpler task than skin segmentation.

Global average pooling: We examined two variants of the patient detection branch using traditional fully-connected (FC) layers and global average pooling layer. The first variant achieved an accuracy of 99.01% in patient detection;

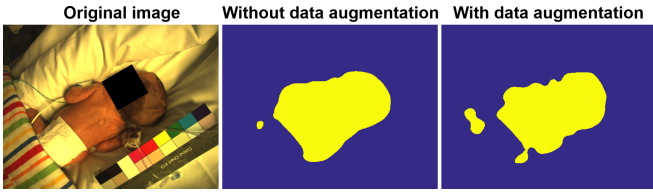


Fig. 6. Image segmentation on the CNN models trained without and with data augmentation applied to training images. The model trained with data augmentation produced skin labels with better segmentation detail.

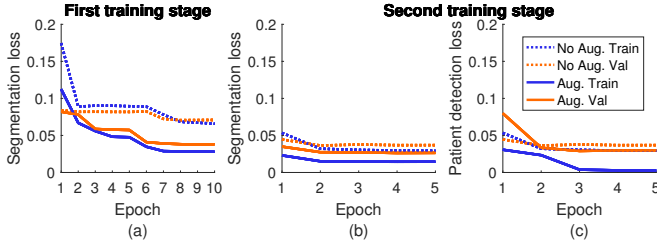


Fig. 7. The decay in the training and validation loss over epochs in (a) the first training stage and (b–c) the second stage. The models trained with data augmentation achieved lower training and validation loss.

however, its IOU score for skin segmentation was 85.23%, which was lower than for the single-task model (see Table II). We did not find this performance decrement in the second variant. This may be because the global average pooling forces feature maps to form confidence maps of each class [15], which is similar to how segmentation predictions are created from feature maps; in addition, the implementation of global average pooling required far fewer parameters.

Network training: In the first training stage, the network converged after epoch 6. The training took 12 hours. In the second training stage, the patient detection branch was overfitted by epoch 5. This is likely to be due to small variations in the negative images. The decay of training and validation loss is shown in Fig. 7. The training took 16 hours. Thus, the whole network took around 28 hours to train. The training was performed on a Nvidia GTX Titan 6GB GPU.

Runtime evaluation: The network can process 512×512 images at a rate of 9.9 images per second if a subject is present in the image, and at a rate of 13.4 images per second if a subject is absent. Real-time performance can be achieved on video processing with dual GPUs, a lower sampling rate or a smaller spatial size.

VI. CONCLUSION

We have presented a multi-task CNN model for patient detection and skin segmentation. A combination of global average pooling for patient detection and the fully convolutional network for skin segmentation yielded high performance, while still keeping a manageable number of parameters with low training time and real-time processing. The model demonstrated robustness with respect to skin tones, pose variations, subject locations and illumination variations. Overall, this approach is capable of determining measurement areas for continuous non-contact vital sign monitoring. Accurate skin segmentation is essential for the successful estimation of vital signs in a clinical context.

REFERENCES

- [1] L. Tarassenko *et al.*, “Non-contact video-based vital sign monitoring using ambient light and auto-regressive models,” *Physiological Measurement*, vol. 35, no. 5, pp. 807–31, 2014.
- [2] M. Villarroel *et al.*, “Continuous non-contact vital sign monitoring in neonatal intensive care unit,” *Healthcare Technology Letters*, vol. 1, no. 3, pp. 87–91, 2014.
- [3] F. P. Wieringa *et al.*, “Contactless multiple wavelength photoplethysmographic imaging: A first step toward “SpO2 camera” technology,” *Annals of Biomedical Engineering*, vol. 33, no. 8, pp. 1034–41, 2005.
- [4] W. Verkruysse *et al.*, “Remote plethysmographic imaging using ambient light,” *Optics Express*, vol. 16, no. 26, pp. 21 434–45, 2008.
- [5] H.-Y. Wu *et al.*, “Eulerian video magnification for revealing subtle changes in the world,” *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.
- [6] L. A. M. Aarts *et al.*, “Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit,” *Early Human Development*, vol. 89, no. 12, pp. 943–48, 2013.
- [7] L. Kong *et al.*, “Non-contact detection of oxygen saturation based on visible light imaging device using ambient light,” *Optics Express*, vol. 21, no. 15, pp. 17 464–71, 2013.
- [8] M.-Z. Poh *et al.*, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation,” *Optics Express*, vol. 18, no. 10, pp. 10 762–74, 2010.
- [9] M.-Z. Poh *et al.*, “Advancements in noncontact, multiparameter physiological measurements using a webcam,” *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 7–11, 2011.
- [10] M. Kumar *et al.*, “DistancePPG: Robust non-contact vital signs monitoring using a camera,” *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–88, 2015.
- [11] V. Gulshan *et al.*, “Geodesic star convexity for interactive image segmentation,” in *Proc. CVPR*, 2010, pp. 3129–36.
- [12] D. P. Mitchell, “Spectrally optimal sampling for distribution ray tracing,” in *Proc. SIGGRAPH*, 1991, pp. 157–64.
- [13] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *ICLR*, 2015.
- [14] J. Long *et al.*, “Fully Convolutional Networks for Semantic Segmentation,” in *Proc. CVPR*, 2015, pp. 3431–40.
- [15] M. Lin *et al.*, “Network In Network,” in *ICLR*, 2014.
- [16] C. Szegedy *et al.*, “Going Deeper with Convolutions,” in *Proc. CVPR*, 2015, pp. 1–9.
- [17] K. He *et al.*, “Deep Residual Learning for Image Recognition,” in *CVPR*, 2016.
- [18] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *Proc. ICML*, 2015, pp. 448–56.
- [19] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. AISTATS*, 2010, pp. 249–56.
- [20] A. Vedaldi and K. Lenc, “MatConvNet – Convolutional Neural Networks for MATLAB,” in *Proc. ACM MM*, 2015, pp. 689–92.
- [21] K.-Z. Lee *et al.*, “Contact-free heart rate measurement using a camera,” in *Proc. CRV*, 2012, pp. 147–52.
- [22] F. Bousefsaf *et al.*, “Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate,” *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 568–74, 2013.
- [23] P. Kakumanu *et al.*, “A survey of skin-color modeling and detection methods,” *Pattern Recognition*, vol. 40, no. 3, pp. 1106–22, 2007.
- [24] M. J. Jones and J. M. Rehg, “Statistical color models with application to skin detection,” *International Journal of Computer Vision*, vol. 46, pp. 81–96, 2002.