Human Skin Segmentation in Videos with Fully Convolutional Neural Networks

Gabriel Birman
Department of Computer Science, Princeton University

Abstract

Convolutional Neural Networks (CNN) have become the state-of-the-art in many computer vision applications, including the task of skin detection in humans. With reference to the task of skin segmentation, this paper not only improves the performance of the current state-of-the-art CNN in the spatial dimension, but also demonstrates increased performance when considering the temporal dimension – relevant for improved skin detection in video formats. The proposed multi-frame CNN refines the output of an arbitrary pre-trained skin detector yielding a small performance boost under specific conditions, and can readily be to pre-recorded and real-time skin segmentation.

Introduction

Skin detection is the problem of labelling pixels in images as either belong to human skin or not belonging to human skin. This often serves as an intermediate processing step for image enhancement, face and human detection, gesture analysis, pornographic contents filtering, surveillance systems, etc.

This paper extends the state-of-the art CNN for skin detection extension for the task of skin detection in color images, where the effects of a Euclidean loss and Cross-Entropy loss on training the network, and the effects of the evaluation methodology are considered. Then, this paper incorporates this CNN into a full-fledged Temporal Refinement Network that uses a set of images temporally centered around an input image to detect skin in that image. This network demonstrates improved results in some facets, yet needs further improvement in other facets.

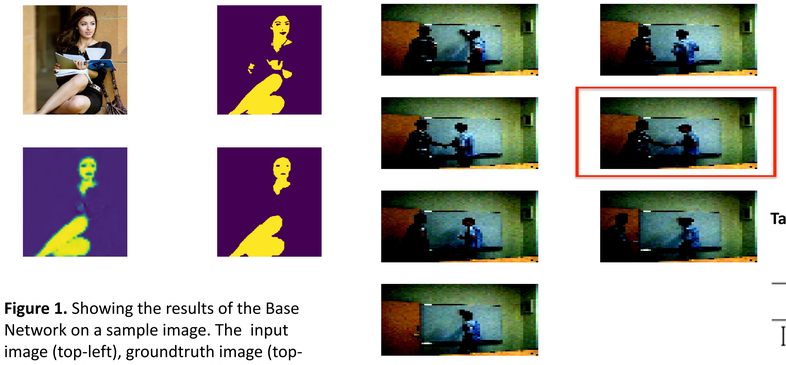

Figure 1. Showing the results of the Base Network on a sample image. The input image (top-left), groundtruth image (top-right), prediction (bottom-left), and binarized image (bottom-right) are all showcased.

Figure 2. The input image (in red) alongside its temporal image set for $T = 7$. Images are chronologically sorted in left to right, top to bottom order.

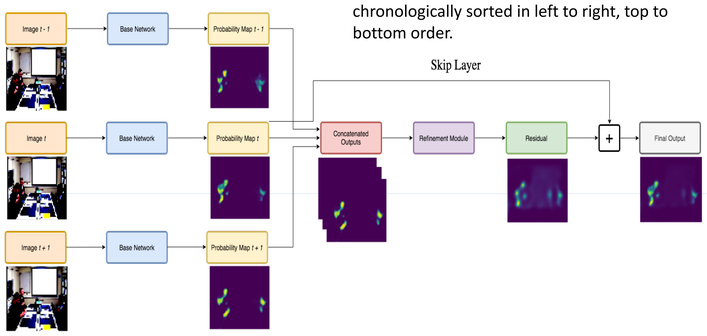

Figure 3. The Temporal Refinement Network Architecture ($T = 3$) with a skip layer. The no-skip analogue would have added a zero image to the summation operation instead of the output of the Base Network on the input image.

Methods

The Temporal Refinement Network consists of two modules: the base network , and the refinement network. The image-based NiN network of Kim, Hwang, Cho was chosen as a template to be modified for the base network. Some modifications included channel-wise normalization for input pre-processing, batch normalization, adding bias to the last convolutional layer, and a different method for weight initialization. The rest of the network and training hyperparameters were kept as in the original, though the batch size and learning rate were chosen through cross-validation. Two loss functions were considered: the original Euclidean loss, and a cross-entropy loss with sigmoid activation.

The modified network was trained on the FSD dataset, and evaluated on the facial component of the Pratheepan dataset, to directly compare with the image-based NiN network. The network with the best performance was then chosen to serve as the base network for the Temporal Refinement Network.

The Temporal Refinement Network was trained on the LIRIS dataset and evaluated on the LIRIS and AMI datasets alongside the base network, which contain videos of human activity. The VDM dataset contained labelled groundtruth images for a subset of frames of the LIRIS and AMI videos. All input-groundtruth pairs from VDM were matched to the corresponding frames in LIRIS/AMI. Using this information 10 sets of T=1,3,5,7 temporal images were collected centered around each input-groundtruth pair, with each set containing images sampled at a fixed interval uniformly selected in (0.25, 5) seconds.

A separate Temporal Refinement Network, with and without the skip layer, was trained for each T value. The temporal image sets were used as inputs to the base network, and the outputs were concatenated into a feature map that was then inputted into the refinement module, which had the same structure/hyperparameters as the base network (except for weight decay = 0.01 chosen through cross-validation). During training, the weights of the base network were not retrained.

Table 1. Base Network Evaluation Results on Pratheepan (Facial) Dataset at Peak F-measure (upsampling)

Methods	Accuracy	Precision	Recall	F-measure
Image-NiN [1]	0.9484	0.9003	0.8912	0.8957
ResNet [5]	0.9499	0.8480	0.8981	0.8678
Ours (CE)	0.9312	0.8409	0.8875	0.8613
Ours (Eucl.)	**0.9553**	**0.9022**	**0.9108**	**0.9045**

References

Y. Kim, I. Hwang, and N. I. Cho. Convolutional neural networks and training strategies for skin detection. In 2017 IEEE International Conference on Image Processing (ICIP) , pages 3919–3923, Sep. 2017.

S. L. Phung, A. Bouzerdoum, and D. Chai. Skin segmentation using color pixel classification: analysis and comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(1):148–154, Jan 2005.

Wei Ren Tan, Chee Seng Chan, Pratheepan Yogarajah, and Joan Condell. A fusion approach for efficient human skin detection. IEEE Transactions on Industrial Informatics, 8(1):138–147, 2011.

J. SanMiguel and S. Suja. Skin detection by dual maximization of detectors agreement for video monitoring. Pattern Recognition Letters, (34):2102–2109, Dec 2013.

A special thanks to professor Olga Russakovsky and preceptor Felix Yu for their assistance!

Results

Results were evaluated across Accuracy, Precision, Recall, and F-measure at the peak F-measure (i.e. where a chosen binarization threshold maximized F-measure).

The modified base network with Euclidean loss showed improvement over the original network when evaluated by upsampling the prediction to its original size using bilinear interpolation, however the Cross-Entropy loss performed better than the Euclidean loss when downsampling the groundtruth using nearest neighbor interpolation (same F-measure as original network).

The Temporal Refinement Network generally performed worse than the Base Network when considering F-measure, except when it was evaluated on the LIRIS dataset for $T = 7$, where marginal improvements were had. The skip architecture performed slightly worse than without.

Figure 4. Comparison of the Base Network output and the Temporal Refinement Network ($T = 7$, skip) output showing marked improvement. Images shown are input, groundtruth, base network, refinement network from left to right.

Table 2. Temporal Refinement Network Results for LIRIS (top) and AMI (bottom) Datasets at Peak F-measure

Methods	Accuracy	Precision	Recall	F-measure	Threshold
Base Network	0.9741	0.3249	**0.3212**	0.2719	0.17
Skip (T=1)	0.9705	0.2312	0.3104	0.2180	-0.39
Skip (T=3)	0.9797	0.3315	0.2973	0.2713	-0.33
Skip (T=5)	**0.9825**	**0.4362**	0.2526	0.2580	-0.30
Skip (T=7)	0.9795	0.3622	0.3119	0.2816	-0.37
No Skip (T=1)	0.9723	0.2523	0.2926	0.2283	0.14
No Skip (T=3)	0.9801	0.2906	0.2809	0.2484	0.24
No Skip (T=5)	0.9803	0.3214	0.2723	0.2515	0.23
No Skip (T=7)	0.9810	0.4130	0.3042	**0.2871**	0.16
Base Network	0.9712	**0.5113**	0.5461	**0.5030**	0.26
Skip (T=1)	0.9695	0.4347	0.4263	0.4133	-0.27
Skip (T=3)	0.9714	0.4692	0.4781	0.4563	-0.29
Skip (T=5)	0.9717	0.4772	0.4744	0.4558	-0.30
Skip (T=7)	0.9728	0.4821	0.5007	0.4729	-0.33
No Skip (T=1)	0.9686	0.4398	0.4300	0.4170	0.27
No Skip (T=3)	0.9727	0.4915	0.5180	0.4867	0.28
No Skip (T=5)	**0.9730**	0.5010	0.4390	0.4492	0.30
No Skip (T=7)	0.9723	0.4865	**0.5463**	0.4975	0.22

Discussion & Conclusion

By modifiying the Image-based NiN network of Kim, Hwang, Cho and some steps involved in training it on the FSD dataset, this paper has achieved state-of-the-art results on the Pratheepan dataset. Moreover, it has been shown that the evaluation method can lead to large variation in outcomes. A clearer standard for evaluation skin detection is needed to better compare network performance.

The Temporal Refinement Network's performance was overall similar to the Base Network, which was unsurprising given the construction of the architecture to include the base network output as input. The complexity of the video datasets proved a limiting factor when assess the Temporal Refinement Network, as the refinement module's performance was constrained by the Base Network's poor performance on these data.

Nevertheless, the Temporal Refinement Network ($T = 7$) outperformed the Base Network on the LIRIS dataset, and slightly underperformed the Base Network on the AMI dataset, hinting that it can be an effective tool for improved segmentation at inference time when guided by qualitative supervision.

Despite the poor performance of the Temporal Refinement Network with a smaller number of temporal input frames, the positive correlation between T and F-measure is promising as it suggests that the refinement module could very well outperform the base network when the base network gives a reasonable initial guess.

Overall, this paper has brought to light many of the concerns of using convolutional neural networks for the problem of skin detection, both in the spatial and temporal dimensions, yielding promising results and avenues for future exploration.