Subpixel Dense Refinement Network for Skeletonization

Sohom Dey
Kalinga Institute of Industrial Technology
Bhubaneshwar, India
sohom21d@gmail.com

Abstract

Skeletonization is the process of reducing a shape image to its approximate medial axis representation while preserving the topology and geometry of the image. Skeletonization is an important step for topological and geometric shape analysis. In this paper a novel skeleton extraction architecture - Subpixel Dense Refinement Network is introduced which is trained and evaluated on the Pixel SkelNetOn Challenge dataset. The proposed architecture is a three-stage encoder-decoder network with dense interconnections between the decoder networks of each stage. The architecture replaces general up-sampling layers and transposed convolution layers with subpixel convolutions for minimizing the information loss during up-sampling of the encoded features. The deep network is trained end-to-end with intermediate supervision in each stage. The proposed single architecture achieved an F1-score of 0.7708 on the validation set of the Pixel SkelNetOn Challenge dataset.

1. Introduction

Over the years, deep learning has made significant advancement in the three main fields of computer vision – image recognition, object detection, and image segmentation. Although state-of-the-art deep learning approaches are capable of competing with human-level performance in numerous tasks in these fields, not much research is done on topological and geometric shape analysis. In this paper, a deep learning based approach is proposed to advance the state-of-the-art in skeletonization task for shape understanding and abstraction.

Skeletonization is the process of extracting or generating an approximate geometric representation (skeleton) of a shape by reducing it to clean skeleton pixels which preserves the extent and connectivity of the original shape. Skeletonization incorporates a fusion of both local and global knowledge of the shape. A skeleton is a compact and intuitive medial axis representation of a shape which retains the topology and geometry of the shape. This representation of the shape is used for various purposes such as modelling, manipulation, synthesis, matching, registration, compression, and analysis.

Image Processing based computational skeletonization algorithms are sensitive to boundary noise and require human intervention for manual parameter tuning for decent skeleton extraction from shapes. This is a time-consuming process and requires a lot of human effort and skills. Deep neural networks can automate this task and learn to output better skeleton representations directly without being susceptible to noise. A few deep learning based approaches in the literature are explored in the following section. Most of them use a segmentation approach to solve this problem. The task of skeleton segmentation is much difficult than standard image segmentation tasks as the extracted skeleton is expected to be of 1-pixel width and must retain the topology and geometry of the shape.

We pose this problem of skeleton extraction as a segmentation task where a semantic segmentation network learns to classify the pixels of a shape into skeleton pixels or background. In this paper, a three-stage encoder-decoder network – Subpixel Dense Refinement Network is introduced. It consists of three key components to improve its capability of extracting skeleton pixels from pre-segmented shape images – dense interconnection among the decoder networks, subpixel convolution layers for efficient up-sampling and intermediate supervision for stable learning. The proposed model is trained end-to-end with intermediate supervision and currently achieves the highest F1-score in the literature. The model is tested on the validation set of the Pixel SkelNetOn Challenge dataset [1] and it achieves an F1-score of 0.7708. In this paper, we provide a detailed description of the model and a comparative analysis of the previous skeletonization approaches on the Pixel SkelNetOn Challenge dataset.

2. Related Work

In this section, a few published approaches for the Pixel SkelNetOn-2019 competition are explored. There are other works on skeletonization but most of them are not related to the domain of geometric shape understanding. The following are the top-ranked methods on the Pixel SkelNetOn competition dataset and thus serves as competitors for consistent benchmark comparisons.

Demir I. et al. [1] in 2019 introduced a baseline model for the Pixel SkelNetOn competition in which they used a vanilla pix2pix network for image translation from distance transformed binary shapes to their approximate skeleton. Their model achieved an F1-score of 0.6244 on the Pixel SkelNetOn validation dataset.

Jiang N. et al. [2] in 2019 proposed their Feature Hourglass Network (FHN) for skeleton extraction. Their model decreases the residual between the prediction and ground truth by integrating side-outputs hierarchically in a deep-to-shallow manner. Their model achieved an F-score of 0.6325 on the Pixel SkelNetOn validation dataset.

Nathan S. et al. [3] in 2019 proposed a custom U-Net architecture with a redesigned decoder in the format of HED architecture. They used 4 side layers fused to one dilated convolution layer for increased performance. Their model achieved an F-score of 0.7480 on the Pixel SkelNetOn validation dataset.

Panichev O. et al. [4] in 2019 proposed their custom U-Net network with residual blocks in encoder and decoder and trained their model with focal-loss to minimize the class imbalance problem. Their model achieved an F-score of 0.7500 on the Pixel SkelNetOn validation dataset.

3. Proposed Method

3.1. Dataset

The Pixel SkelNetOn Challenge dataset is used for the experiment. The dataset contains 1725 single-channel segmented binary shape images and their corresponding binary skeleton images. The images are in portable network graphics format each having a dimension of 256x256. The dataset is split into a training set of 1218 samples, a validation set of 241 samples and a test set of 266 test samples. The ground truth skeleton images are provided only for the training set. The validation set and the test set are used for evaluation on the CodaLab evaluation server of SkelNetOn challenge. Since only the ground truth images are available only for the training set, the original training set is further divided into a training split of 1000 samples and a validation split of 218 samples. The model is trained on this training split and tuned on the validation split. The model is tested on the original validation set on the evaluation server. Since the test set is not made available until the last phase of the competition, it is not used for comparison. All the comparisons provided for each method in this paper are the validation set scores.

The images are normalized between 0-1 and that's the only pre-processing done. Since only 1000 samples are available for training, we use data augmentation to increase the number of samples to 3000. Spatial-level transforms simultaneously applied on both the input images as well as the skeleton images is used for this purpose. This is an essential step to prevent the model from overfitting. A combination of random flip, random rotation, random transpose, random shift, random scaling, elastic transform, grid distortion and piecewise affine transform is used for augmentation.

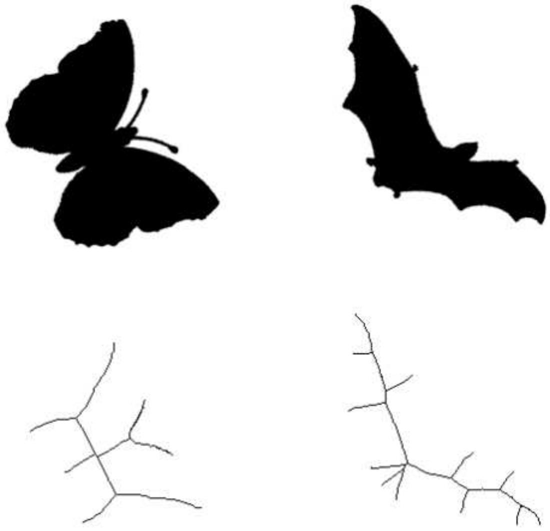

Figure 1. Sample images and their ground truth skeletons

3.2. Architecture

The proposed Subpixel Dense Refinement Network, shown in Fig. 2, is a three-stage segmentation architecture which builds upon the U-Net [5, 6] design. The concept of stacking architectures [7] is not new, but the simple stacking strategy doesn't help much in the skeletonization task. We propose a novel stacked architecture design specializing in skeleton extraction which outperforms all the previous methods of skeleton extraction.

In this three-stage architecture, the feature map from the penultimate layer of the first stage is passed on to the second stage. The feature map of the penultimate layer contains more information than the last layer output. This feature map is concatenated with the original shape image and is passed to the input of the second stage. Concatenating the original shape image with the previous feature map improves the refinement of the predictions. Similarly, the feature map from the penultimate layer of the second stage is passed on to the third stage and concatenated with the original input image. The final predictions are obtained from the output layer of the third stage

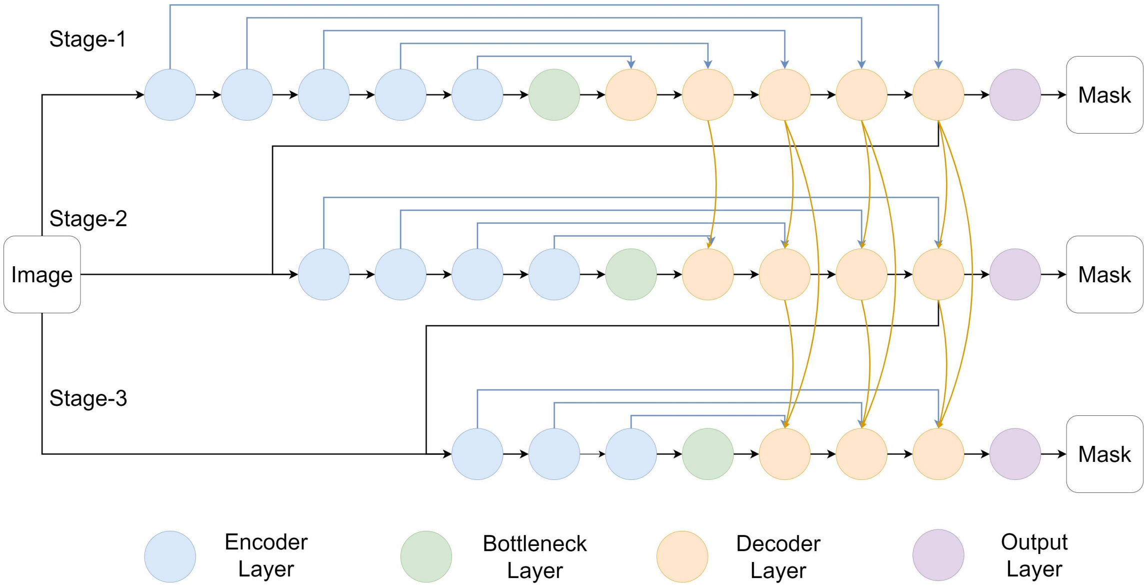

Figure 2. Proposed SDRNet Architecture

The interpolation-based up-sampling method or the learnable transposed convolution method is replaced by the efficient subpixel convolutions [8]. A subpixel convolution layer, shown in Fig. 3, is just a standard 1x1 convolution layer followed by a pixel shuffling operation which rearranges the pixels from depth dimension to the spatial dimension. Subpixel convolution unlike the previous interpolation methods or the transposed convolution method minimizes the loss of information during the up-sampling of images in the decoder network.

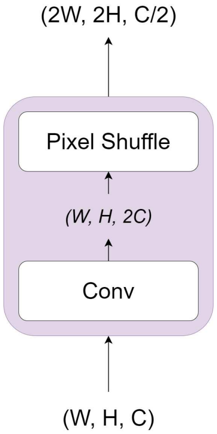

Figure 3. Subpixel Convolution Layer

The parallel layers of the three decoder networks are connected via dense connections [9, 10, 11]. This improves the spatial knowledge transfer through the model by allowing the decoders of each stage to use the feature maps from the decoders of previous stages. Each subsequent stage is shallower than its previous stage as it uses a lot of prior knowledge from the earlier stages. Each stage has its own output layer and the model is trained end-to-end with intermediate supervision in each stage which helps in efficient training and improves convergence.

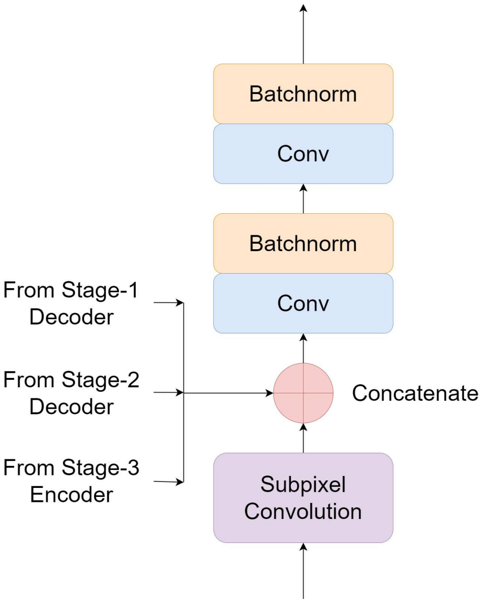

Figure 4. A decoder block from stage-3

3.3. Loss Function

Pixel-wise Binary Cross-entropy is a widely used loss function for semantic segmentation since it evaluates the class predictions for each pixel individually. Another widely used loss function is the Dice loss which measures the overlap between samples. Pixel-wise cross-entropy loss suffers from class imbalance while the Dice loss has a normalizing effect and is not affected by class imbalance. A combination of Binary Cross-entropy and Dice Loss (Bce-Dice Loss) is used to train the model.

$$BCE\,Loss = -[y\,log\,p + (1-y)\,log(1-p)] \quad (1)$$

$$Dice\,Loss = 2 \times \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \quad (2)$$

$$BCE - Dice\,Loss = BCE\,Loss + Dice\,Loss \quad (3)$$

3.4. Evaluation Metric

The spatial representation of the skeleton pixels in each image is much lower than the background pixels. This is why F1-score is used for evaluation which takes into account the class imbalance. This is also the most widely used evaluation metric on the literature and thus allows consistent benchmark comparisons. F1-score is the harmonic mean of precision and recall.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. Experimental Evaluation

The model is trained end to end on 3000 augmented samples. It takes around 3 hours to train the model on Google Colab running a 16 GB NVIDIA Tesla P100. The 218 samples of the validation split that was made from the original training set are used for validating and fine-tuning the model. The model is tested on the original validation set on the SkelNetOn competition evaluation platform on CodaLab, thus guaranteeing consistent evaluation score. The model is trained on original size images (256x256) and the batch size is set to 5 to compensate for the hardware limitations. Adam optimizer is used with the initial learning rate set to 0.001. The learning rate is reduced by a factor of 0.1 on reaching a plateau and similarly, training is stopped when the validation loss doesn't decrease for 7 consecutive epochs.

The model achieves an F1-score of 0.7708 on the original validation set of the Pixel SkelNetOn Challenge dataset which is currently the best score in the literature and the competition. A quantitative analysis of the results is shown in the following tables.

S/N	F1-score	S/N	F1-score
Exp-1	0.770798	Exp-6	0.770595
Exp-2	0.769658	Exp-7	0.77006
Exp-3	0.770756	Exp-8	0.770731
Exp-4	0.770171	Exp-9	0.769901
Exp-5	0.770213	Exp-10	0.770142

Table 1. Results for 10 experiments with random initializations

Models	F1-score
UNet	0.6987
UNet + Subpixel	0.7130
UNet + Subpixel + Stacked	0.7358
UNet + Subpixel + Stacked + Dense Interconnections	0.7577
UNet + Subpixel + Stacked + Dense Interconnections + Intermediate Supervision	0.7708

Table 2. Ablation Study

Method	F1-score
pix2pix (baseline) [1]	0.6244
Jiang [2]	0.6325
Nathan [3]	0.7480
Panichev [4]	0.7500
Proposed method	**0.7708**

Table 3. Comparison of results with existing literature

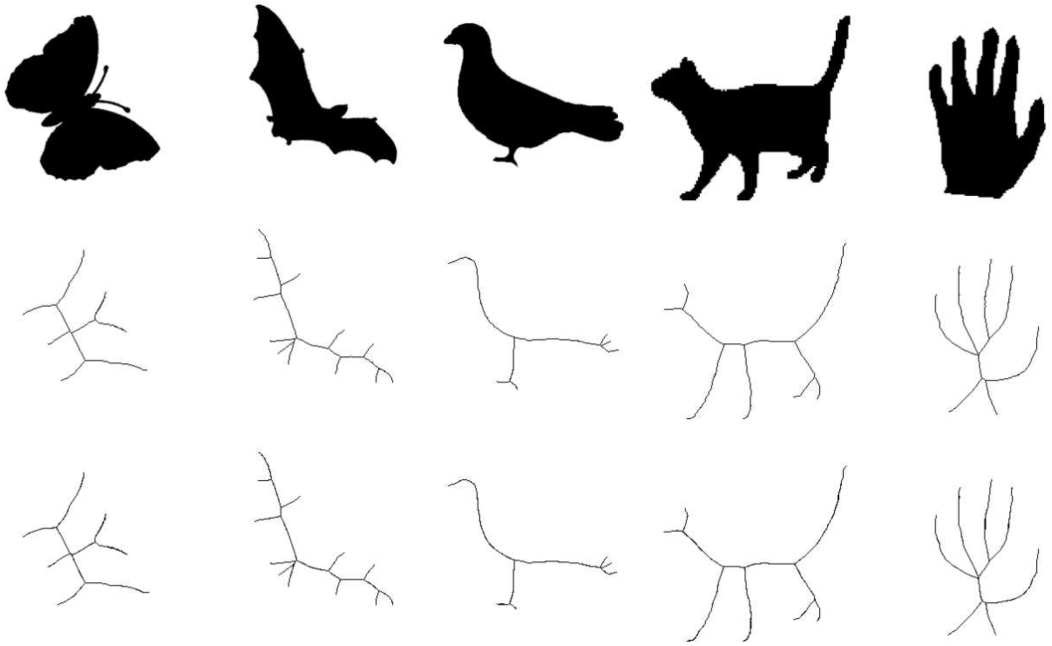

Figure 5. From top to bottom: Images, Ground Truth Skeleton, Predicted Skeleton

Teams	F1-score
acdart	0.6535
sabarinathan	0.7279
HeBowei	0.7358
gro	0.7358
Proposed method	**0.7708**

Table 4. CodaLab Leaderboard Results as on 31st March 2020

5. Conclusion

In this paper, we introduced a novel architecture for skeletal image extraction from pre-segmented shape images. The Subpixel Dense Refinement Network proposed here outperforms all the state-of-the-art methods for skeletal shape extraction till date and achieves an F1-score of 0.7708 on the Pixel SkelNetOn Challenge validation dataset. The three-stage refinement, interstage spatial knowledge transfer through dense connections, lossless up-sampling using subpixel convolution and intermediate supervision, all of these together complements each other producing a superior skeleton extraction network and setting a new benchmark for the skeletonization task.

References

[1] I. Demir, C. Hahn, K. Leonard, G. Morin, D. Rahbani, A. Panotopoulou, A. Fondevilla, E. Balashova, B. Durix, and A. Kortylewski. Skelneton 2019: Dataset and challenge on deep learning for geometric shape understanding. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1143–1151, 2019.

[2] N. Jiang, Y. Zhang, D. Luo, C. Liu, Y. Zhou, and Z. Han. Feature hourglass network for skeleton detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1172–1176, 2019.

[3] S. Nathan and P. Kansal. Skeletonnet: Shape pixel to skeleton pixel. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1181–1185, 2019.

[4] O. Panichev and A. Voloshyna. U-net based convolutional neural network for skeleton extraction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1186–1189, 2019.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015.

[7] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki. Iternet: Retinal image segmentation utilizing

structural redundancy in vessel networks. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3656–3665, 2020.

[8] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[9] Z. Zhou, M.M. Rahman Siddiquee N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.