Deep Colorization for Facial Gender Recognition

Jonathan Hogervorst, Emmanuel Okafor, and Marco Wiering

Institute of Artificial Intelligence and Cognitive Engineering,
Faculty of Science and Engineering, University of Groningen, The Netherlands
jonathan@hogervorst.info,{e.okafor,m.a.wiering}@rug.nl

Abstract. Recent research suggests that colorization models have the capability of generating plausible color versions from grayscale images. In this paper, we investigate whether colorization prior to gender classification improves classification performance on the FERET grayscale face dataset. For this, we colorize the images using an existing Lab colorization model, both with and without class rebalancing, and our novel HSV colorization model without class rebalancing. Then we construct gender classification models on the grayscale and colorized datasets using a reduced GoogLeNet convolutional neural network. Several models are trained using different loss functions (cross entropy loss, hinge loss) and gradient optimization solvers (Nesterov's Accelerated Gradient Descent, Stochastic Gradient Descent), initialized using both random and pre-trained weights. Finally, we compare the gender classification accuracies of the models when applied to the face image color variants. The best performances are obtained by models initialized using pre-trained weights, and models using colorization without class rebalancing.

Keywords: Deep Convolutional Neural Networks · Gender recognition · Face image analysis · Computer vision · Automatic colorization

1 Introduction

Gender recognition is an interesting problem with various applications, like video surveillance and authentication. Furthermore, gender plays an important role in human interaction. Gender recognition can therefore also be a useful method of improving human-computer interaction [20]. While gender recognition is a simple task for humans, it is difficult to create a system that can perform this task with sufficient performance. Additionally, face images may only be available in grayscale. For example, simple surveillance cameras and night vision cameras often do not provide color images, and historic photo and video material is also not available in color. Gender recognition may be more difficult on grayscale images because color images contain more information that may aid recognition.

Much research has been done into face gender recognition systems, applying various kinds of techniques [4, 24]. A successful approach is the use of support vector machines on face images [9, 15]. Recently, convolutional neural networks (CNNs) have been tried as an approach to gender classification as well [7, 20].

Research has examined the development of systems that can predict color information given a grayscale image using different approaches. One such approach is the transfer of color from a source image [13, 19], requiring one or more color source images featuring a scene similar to the image to be colorized. Another approach is colorization of grayscale images using color scribbles [8, 21]. While such systems do not need similar source images, they require the user to scribble appropriate colors onto the grayscale image. Recent research yielded colorization models which use large datasets of color images [3, 23]. These systems colorize images automatically, without any user input.

Contributions. In this paper, we investigate the influence of colorization prior to gender classification on the FERET grayscale face dataset [12]. The images are colorized using a colorization CNN [23] in Caffe [6]. We use the provided Lab colorization model, which predicts CIE *Lab* color values given a grayscale input image. We use the provided models with and without class rebalancing, which compensates for the fact that some color areas in the *ab* output space occur less often in photos. Furthermore, we adapt the CNN to predict HSV color values and train our own HSV colorization model without class rebalancing on the ImageNet dataset [14]. We then train gender classification models on the various face image color variants. For this, we use a reduced version of the GoogLeNet CNN [16] in Caffe. We compare two classification loss functions: cross entropy loss and hinge loss. Additionally, we compare two gradient optimization solvers: Nesterov's Accelerated Gradient Descent (NAGD) [10] and Stochastic Gradient Descent (SGD). Gender classification models are trained both from scratch, using random Xavier initialization [5], and fine-tuned from a pre-trained GoogLeNet model. We compare classification accuracies of the eight resulting gender model variants for the four image color variants using ten-fold Monte Carlo cross-validation (MCCV).

Outline. In Section 2, we present our method: first, we train an HSV colorization model in addition to the existing Lab colorization models (Section 2.1); then, we colorize the grayscale images to new, colorized versions of the dataset using the colorization models (Section 2.2); and lastly, we use the colorized images in our gender classification models (Section 2.3). In Section 3, we provide our results. Finally, in Section 4, we discuss our conclusions.

2 Methods

2.1 Colorization

To perform colorization, we used a colorization CNN [23] implemented in the deep learning framework Caffe [6]. This CNN was created with the goal of predicting plausible color versions of grayscale images given as input. The CNN employs the CIE *Lab* color space: it predicts the a and b channels of an image, given its L channel. The L channel contains information about the image's lightness or brightness — it thus provides a grayscale representation of the image.

The colorization CNN is unique due to its tailored loss function. Image color prediction is a multimodal problem [3]: objects in an image often have multiple

plausible colors. By using a loss function like the Euclidean loss, the optimal solution would be the mean of all plausible color values, resulting in grayish colors [23]. To prevent this, the colorization CNN predicts a probability distribution over the possible colors. For this, the in-gamut ab output space is divided into 313 bins, over which the probabilities are calculated. In addition, the loss function applies class rebalancing. This is done to compensate for the fact that some color values in the ab output space occur less often in photos. Class rebalancing reweighs the loss of pixels based on the rarity of the color during training, ensuring that the model can still predict less common color values.

The Lab colorization model was trained on the training set of the ImageNet dataset [14], containing 1.3M color images, in 450k iterations [23]. The ImageNet dataset consists of color images of many different types of objects collected from the Internet. We used the trained Lab models with[1] and without class rebalancing[2]. More information about the loss function and training parameters of the Lab colorization model can be found in [23].

Moreover, we adopted the colorization CNN to create our own HSV colorization model. In this CNN, we predict the H and S channels of an image given its V channel. Similar to the L channel in the Lab color space, the V channel in the HSV color space provides a grayscale representation of an image. We applied binning over the HS output space using 18 equally sized bins in both directions, yielding 324 bins in total. We did not apply class rebalancing. We trained our HSV colorization model on the training set of the ImageNet dataset for 300k iterations using the unmodified training parameters of the Lab colorization CNN. This took $\sim$ 14 days on a single core of an NVIDIA Tesla K40 GPU.

2.2 Dataset

We employed a selection of 1242 images of the grayscale face image dataset FERET [12]. This dataset consists of photos of subjects from several angles with different facial expressions. Our selection consisted of frontal images and images in which subjects had their face slightly turned. Our images were cropped to squares with the faces aligned in the middle and contained one color channel. From our selection of 1242 images, we put 255 aside as test set. We randomly split the remaining images into a training set of 789 images (80%) and a validation set of 198 images (20%). This random splitting procedure was repeated to create ten training/validation set pairs for MCCV.

We then colorized the images. To colorize an image using the Lab colorization models, we loaded it using the Caffe [6] function `io.load_image()`, automatically converting it to three-channel RGB data in grayscale. Then we converted it to Lab using the scikit-image [18] function `color.rgb2lab()` and provided the L channel as input to the colorization model. The output of the model, being the predicted ab color values, was combined with the calculated L channel. The resulting Lab image was then converted to RGB using `color.lab2rgb()` and

[1] `colorization_release_v2.caffemodel`
[2] `colorization_release_v2_norebal.caffemodel`

stored. We repeated this for all images. Similarly, each image was colorized using our HSV colorization model: we converted it to HSV using `color.rgb2hsv()`; provided the V channel to the model; combined the V channel with the predicted HS channels; converted it to RGB using `color.hsv2rgb()`; and stored it.

We thus had four variants of the dataset: original grayscale, Lab colorized with class rebalancing, Lab colorized without class rebalancing, and HSV colorized without class rebalancing. Some example images are shown in Fig. 1.

Fig. 1. Example images from the FERET face dataset [12] used in our research. From left to right: original grayscale, Lab colorized with class rebalancing, Lab colorized without class rebalancing, and HSV colorized without class rebalancing.

2.3 Gender Recognition

Network We performed gender classification using a reduced version of the GoogLeNet CNN [16] implemented in Caffe [6]. GoogLeNet contains Inception modules, each consisting of six convolution layers and one pooling layer, outputting the concatenation of its contained layers. The original CNN is 27 layers deep and contains nine Inception modules. It yielded an excellent performance on the ImageNet Large-Scale Visual Recognition Challenge 2014 [16].

We noticed that the extensive GoogLeNet structure would not be required for our gender classification task. Therefore we reduced it, removing six of the nine Inception layers. The resulting reduced GoogLeNet still provided good performance for our task, while saving $\sim 40\%$ computation time during training. The structure of our reduced GoogLeNet is shown in Fig. 2.

Loss Function Loss functions indicate the difference between predicted values and target values in supervised learning. During training, the loss is minimized by the CNN to get the predictions closer to the target values, improving the model's performance. The original GoogLeNet contains a softmax classification layer which employs the cross entropy loss function (`SoftmaxWithLossLayer` in Caffe). We also applied this in our reduced CNN. Additionally, we created a variant of our CNN using the hinge loss function (`HingeLossLayer`). More information can be found in [1].

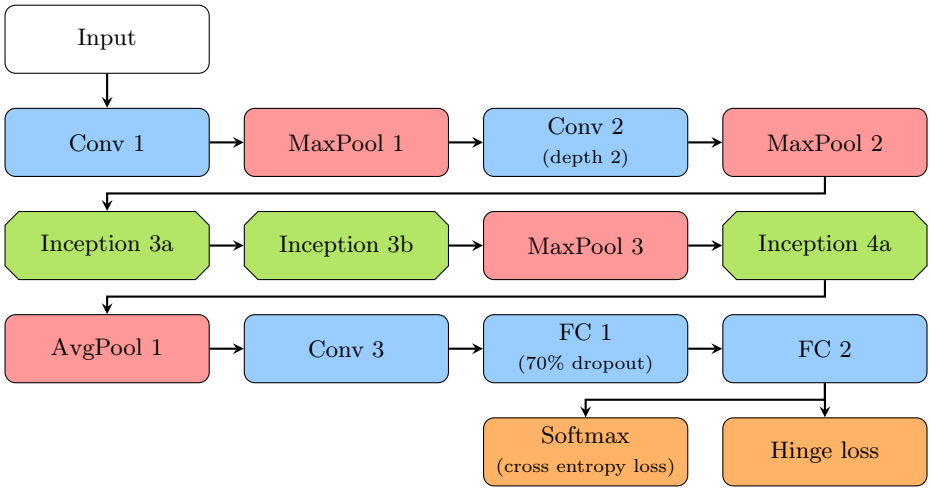


Fig. 2. Structure of our reduced GoogLeNet CNN for gender classification

Gradient Solver During training, the model's parameters are updated to minimize the loss. The gradient solver specifies how the CNN updates its parameters. We employed two different gradient optimization solvers: NAGD [10] and SGD. More information can be found in [2].

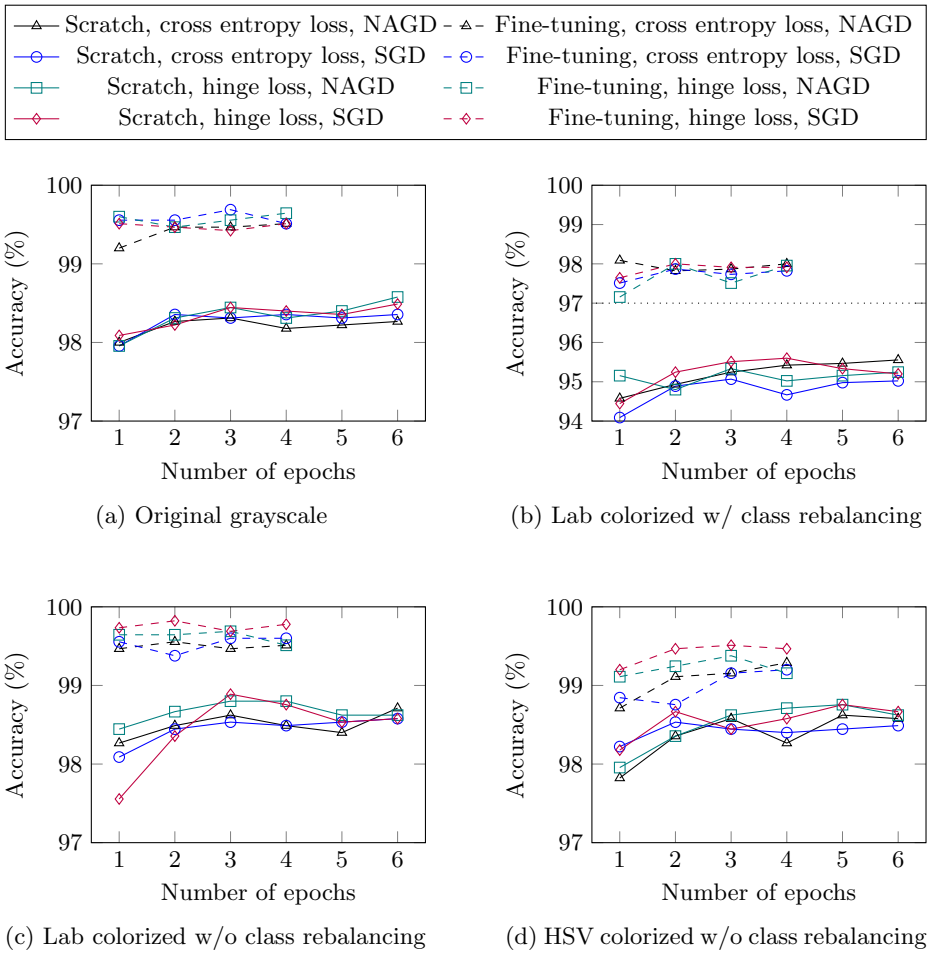
Initialization The Caffe implementation of GoogLeNet employs Xavier initialization [5]. This algorithm initializes the model's weights from a distribution around zero, with a variance based on the number of input/output neurons in the CNN. We employed this technique when training our models from scratch.

In addition to training from scratch, we also trained models by fine-tuning from pre-trained weights. In this case, the weights in our models were initialized using weights from the pre-trained Caffe GoogLeNet model[3], which was trained on the ImageNet dataset similar to [16].

Training Parameters When training our models, we set the base learning rate to 0.001 (`base_lr` parameter in Caffe's `solver.prototxt` file). After each step of 10k training iterations (`stepsize`; `lr_policy=step`) we lowered the learning rate by multiplying it with 0.96 (`gamma`). Furthermore we used a momentum of 0.9 (`momentum`) and a weight decay factor of 0.0002 (`weight_decay`).

We trained gender classification models on the ten training sets for each color variant of our dataset. We trained each scratch model for 30k iterations, which took $\sim$ 1.5 hour on a single core of an NVIDIA Tesla K40 GPU. We trained each fine-tuned model for 20k iterations, which took $\sim$ 1 hour on the same GPU.

[3] `bvlc_googlenet.caffemodel`



(a) Original grayscale

(b) Lab colorized w/ class rebalancing

(c) Lab colorized w/o class rebalancing

(d) HSV colorized w/o class rebalancing

Fig. 3. Average gender classification accuracy of ten-fold MCCV on the test set after training for a number of epochs (1 epoch ≡ 5k iterations). The scratch CNN models were trained for 30k iterations, the fine-tuned CNN models for 20k iterations.

3 Results

The gender classification learning curves of our models after various numbers of training iterations are shown in Fig. 3. The scratch models, which used Xavier weight initialization, show significant improvements after the first and second training epoch; after that, the accuracies slightly improve further. The same applies to some of the fine-tuned models, which were initialized using pre-trained weights. Other fine-tuned models only show slight improvements from the first training epoch on. Because the models keep improving until the final training epoch, the models at the final training epoch will be used for further reporting.

Table 1. Average gender classification accuracy in % (± standard deviation) of ten-fold MCCV on the test set after training for 30k (scratch) or 20k (fine-tuning) iterations.

Training	**Loss Function**	**Gradient Solver**	**Grayscale**	**Lab Color'ed** w/ class rebal.	**Lab Color'ed** w/o class rebal.	**HSV Color'ed** w/o class rebal.
From scratch (Xavier init.)	Cross entropy loss	NAGD	98.27 ± 0.50	95.56 ± 0.66	98.71 ± 0.46	98.58 ± 0.55
		SGD	98.36 ± 0.40	95.02 ± 0.48	98.58 ± 0.62	98.49 ± 0.80
	Hinge loss	NAGD	98.58 ± 0.56	95.24 ± 0.63	98.62 ± 0.58	98.62 ± 0.42
		SGD	98.49 ± 0.41	95.20 ± 0.91	98.58 ± 0.62	98.67 ± 0.40
Fine-tuning (pre-trained weights)	Cross entropy loss	NAGD	99.51 ± 0.24	$\mathbf{98.00 \pm 0.46}$	99.51 ± 0.31	99.29 ± 0.22
		SGD	99.51 ± 0.42	97.82 ± 0.37	99.60 ± 0.24	99.20 ± 0.33
	Hinge loss	NAGD	$\mathbf{99.64 \pm 0.27}$	97.96 ± 0.80	99.51 ± 0.42	99.16 ± 0.42
		SGD	99.51 ± 0.31	97.91 ± 0.56	$\mathbf{99.78 \pm 0.30}$	$\mathbf{99.47 \pm 0.27}$

The gender classification test accuracies of our various models are reported in Table 1. The results show that the maximum classification accuracy is obtained by the fine-tuned model using hinge loss and SGD, applied on the Lab colorized version of the dataset without class rebalancing, yielding 99.78%.

When we compare the accuracies of our models on the Lab and HSV colorized images without class balancing, they show a significant increase over the models with class rebalancing. This implies that colorization without class rebalancing works better for this study. Furthermore, the models without class rebalancing only show a slight improvement over some of the original grayscale image models.

In general, our fine-tuned models showed more promising performance than their scratch counterparts on all versions of the dataset we examined. The results of the models using the different loss functions or gradient solvers show only very small differences — no conclusions can be drawn from this.

4 Conclusions

This study comprised a comprehensive evaluation using different variants of a reduced GoogLeNet CNN to determine the gender classification performance on original and colorized versions of the FERET grayscale face dataset. The results show that class rebalancing in colorization does not improve the gender classification performance in this study since all models without class rebalancing yielded better results than the models with class rebalancing. Additionally, fine-tuning from pre-trained weights seems to be the best approach in this study: all fine-tuned models yielded better performance than their scratch counterparts, while also requiring less training time.

In general, the gender classification accuracies from our models were all very high and similar. While we observed some improvements in models using colorization, they were in the order of tenths of a percentage point. To attribute sig-

nificant improvement to colorization prior to gender classification, there should be a greater difference in accuracy between the grayscale and colorized models.

A reason for the high classification accuracies could be that the model used some individual characteristics of subjects, rather than gender characteristics only. Because subjects were photographed from multiple angles for the FERET dataset, different images of the same face could have ended up in different sets, like our training and test set. While such images were not identical, they may have allowed the model to use face-specific characteristics.

Research has shown that color information improves the performance of face recognition systems [17]. Moreover, psychological research suggests that humans use color information in gender recognition if other information is unavailable [11, 22]. Our dataset consisted of high quality frontal whole face photos. Our approach of colorization prior to gender classification might be better applicable to more complex datasets. In that case, the grayscale baseline results would be worse, and colorization could bring actual improvement.

Future work should examine multi-orientation data augmentation using colorized images, which may improve the performance further.

Acknowledgments. The authors would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

1. Berkeley Vision and Learning Center: Caffe documentation. `http://caffe.berkeleyvision.org/doxygen/` (2017), accessed: 2017-03-17
2. Berkeley Vision and Learning Center: Caffe tutorial: Solver. `http://caffe.berkeleyvision.org/tutorial/solver.html` (2017), accessed: 2017-03-17
3. Charpiat, G., Hofmann, M., Schölkopf, B.: Automatic image colorization via multimodal predictions. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 126–139. Springer, Berlin, Heidelberg (2008)
4. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces: a survey. In: Proceedings of the IEEE. vol. 83, pp. 705–741. IEEE (1995)
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) 13th International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, vol. 9, pp. 249–256. PMLR (2010)
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: 22nd ACM International Conference on Multimedia. pp. 675–678. ACM, New York (2014)
7. Levi, G., Hassncer, T.: Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 34–42. IEEE (2015)
8. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: Hart, J.C. (ed.) ACM SIGGRAPH 2004. ACM Transactions on Graphics, vol. 23, pp. 689–694. ACM, New York (2004)

9. Moghaddam, B., Yang, M.H.: Gender classification with support vector machines. In: IEEE International Conference on Automatic Face and Gesture Recognition. pp. 306–311. IEEE (2000)
10. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady 27(2), 372–376 (1983)
11. Nestor, A., Tarr, M.J.: Gender recognition of human faces using color. Psychological Science 19(12), 1242–1246 (2008)
12. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 16(5), 295–306 (1998)
13. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications 21(4), 34–41 (2001)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3), 211–252 (2015)
15. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recognition Letters 33(4), 431–437 (2012)
16. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9. IEEE (2015)
17. Torres, L., Reutter, J.Y., Lorente, L.: The importance of the color information in face recognition. In: 1999 International Conference on Image Processing. vol. 3, pp. 627–631. IEEE (1999)
18. van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: scikit-image: image processing in Python. PeerJ 2 (2014)
19. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. In: ACM SIGGRAPH 2002. ACM Transactions on Graphics, vol. 21, pp. 277–280. ACM, New York (2002)
20. van de Wolfshaar, J., Karaaba, M.F., Wiering, M.A.: Deep convolutional neural networks and support vector machines for gender recognition. In: 2015 IEEE Symposium Series on Computational Intelligence. pp. 188–195. IEEE (2015)
21. Yatziv, L., Sapiro, G.: Fast image and video colorization using chrominance blending. IEEE Transactions on Image Processing 15(5), 1120–1129 (2006)
22. Yip, A.W., Sinha, P.: Contribution of color to face recognition. Perception 31(8), 995–1003 (2002)
23. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016)
24. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2003)