Leveraging histogram equalization for expert photo enhancement
IACV 2019 Project

Luca Cavalli
Politecnico di Milano
luca3.cavalli@mail.polimi.it

Gianpaolo Di Pietro
Politecnico di Milano
gianpaolo.dipietro@mail.polimi.it

Abstract

Human expert photo enhancement applies arbitrary color transformations on different semantic areas of the photo, making it an unstructured problem to be approached directly. Histogram equalization, on the other side, is a simple algorithm that can easily be analysed to suggest structural constraint on a deep architecture to improve learning of meaningful color transformations. In this paper we design deep architectures for color transformations specialized in histogram equalization and we show their generalization capabilities to the expert photo enhancement task. Experimental results suggest that one of the proposed architectures can achieve slightly better results with respect to traditional architectures using much less parameters.

1. Introduction

Images are a major mean of communication in modern society, ranging from social media usage to advertising; as a consequence photo enhancement is a widespread need. High quality enhancements in digital photos are traditionally done by hand labor of well-trained professionals and do not only include photographic defects, but also artistic visual impressions.

In this paper we focus on the problem of learning automatic photo enhancement from the style of a human expert. The enhancement applies different color transformations on different regions of the image, depending both on the global appearance of the image and on the semantics of what the image depicts, leading to a highly nonlinear transformation which is specific of the artistic style of an expert. In addressing this problem we must also take into consideration that the real transformation applied by a human expert cannot be strictly considered as a function as there is significant variance in the artistic production of a human.

Histogram equalization, on the other side, is a simple enhancing algorithm which remaps colors to equalize the global distribution of colors in the image. This algorithm for automatic enhancement provides far inferior results with respect to human artistic enhancement but it is extremely simple and it can be taken as an intermediate step to better understand the task of learning expert photo enhancements. In this paper we select a suitable deep architecture to perform well on the simple task of histogram equalization, and then show their generalization capabilities to the expert photo enhancement task. Experimental results suggest that one of the proposed architectures can achieve slightly better results with respect to traditional architectures using much less parameters.

(a) Original (b) Ground Truth (c) Ours

Figure 1: Expert photo enhancement example. Different transformations are applied to different regions of the image

1.1. Problem definition

In this section we give a formal definition of the problem of photo enhancement. Given a photo $\mathcal{P}$ we define the ideal photo enhancement (with fixed style and no noise) as a function $\mathcal{F}$ that maps $\mathcal{P}$ into its enhanced image $\mathcal{Q} = \mathcal{F}(\mathcal{P})$. This function $\mathcal{F}$ of the whole image can be decomposed as a pixelwise function $\mathcal{F}_{pw}$ that maps each pixel $\mathcal{P}^i$ of $\mathcal{P}$ and local features $\mathcal{L}(\mathcal{N}_i(\mathcal{P}))$ around a neighborhood $\mathcal{N}$ and global features $\mathcal{G}(\mathcal{P})$ into the corresponding pixel of image $\mathcal{Q}$. In short for each image position i we have:

$$\mathcal{Q}^i = \mathcal{F}_{pw}(\mathcal{P}^i, \mathcal{L}(\mathcal{N}_i(\mathcal{P})), \mathcal{G}(\mathcal{P}))$$

Both local and global features in general can encode complex semantic information about the depicted content or the setting and meaning of the photo.

The histogram equalization is an ideal photo enhancement (it is indeed a deterministic function) that can be modeled within our formalism by taking no local features $\mathcal{L}$ and using the global histogram of the image as global feature $\mathcal{G}$.

2. Related work

The widespread commercial solution to the problem of image enhancement is typically a set of tools for human experts to apply color transformations to images, such as Adobe Photoshop or Adobe Lightroom. Some of the commercial tools also include automatic enhancement procedures which usually apply global color adjustments without taking into consideration the semantics of the image, thus leaving high quality enhancements to humans.

However, in the research literature we find much activity in the field of quality automatic photo enhancement. One of the first researches towards including semantics in automated enhancement is from Kaufman *et al.* [2] who detects typical semantic content (such as skies and faces) and applies pre-determined transformations based on the semantic class of the image patch. Subsequent works considered also the variability in styles, like Yan *et al.* [7]. They learn context-aware quadratic color transformations by providing a set of global and local features to a multi-layer deep neural network, and they show how the selection of the right features is crucial for effective learning of an enhancement style.

A completely different approach has been proposed by Kinoshita *et al.* [4] who enhance low dynamic range (LDR) images by training on LDR training samples synthesized from high dynamic range (HDR) images. The set of useful features in this task is not clear a priori as both input and ground truth are generated algorithmically, so they use a convolutional neural network inspired on the U-Net [6] to automatically learn both local and global significant features.

With our approach we neither explicitly select the features to be used like Yan *et al.* [7] nor let a general purpose architecture to select the relevant features free of prior like Kinoshita *et al.* [4]. Instead, we embed a prior on the relevant features into the structure of the deep architecture, based on the experimental results for histogram equalization.

3. Proposed approach

In this section we will analyse the problem of learning histogram equalization keeping in mind that the final objective is generalization for the more complex task of expert photo enhancement.

Histogram equalization is an algorithmic enhancement that aims at mapping pixel intensity values so that the histogram of the output image results as flat as possible, covering the whole available color range. It can be thus decomposed into two problems: computing the histogram of the input image and finding an adequate intensity mapping to equalize it. We do not make further hypotheses on how this mapping is computed to avoid an excessive bias of the designed architectures towards histogram equalization only.

In the following we will consider the structural hyperparameters:

- $\mathcal{H}$: the number of bins of the histogram, which is its size.
- $\mathcal{D}$: the depth of the deep architecture.
- $\mathcal{C}$: the channels of the input and output images, has either value 1 (grayscale) or 3 (RGB).

3.1. Pixel-wise fully connected

The direct approach towards learning histogram equalization is using the input image histogram itself as an input feature and learning the intensity mapping with a fully connected neural network as a universal function approximator. This setting guarantees that the features provided are fully informative for the histogram equalization task, but it would require selecting new features to generalize to the photo enhancement task.

The parametric architecture of our fully connected mapper is represented in Figure 2. As the number of possible intensity mappings grows exponentially with the size $\mathcal{H}$ of the histogram, the complexity of the network is parameterized by the size of the input histogram feature as well. The depth $\mathcal{D}$ of the mapper and the input pixel channels $\mathcal{C}$ are other hyperparameters that change the complexity and the domain (RGB or grayscale) of the mapper. The input vector is a vector of size $\mathcal{I} = \mathcal{C} + \mathcal{H}$ including the normalized histogram and the normalized pixel value to be mapped. The $\mathcal{D}$ layers are all fully connected layers with relu activation to prevent the gradient from shrinking (except the second last layer which is tanh and the last layer which is sigmoid to output normalized pixel intensities); the first layer reduces the inner representation size to $\mathcal{I}/2$, which is kept constant for $\mathcal{D}-3$ layers and finally reduced to $\mathcal{I}/4$ before the output of size $\mathcal{C}$.

3.2. Histogram convolution

A typically successful approach in image related learning is letting the architecture learn to extract the most significant features and their representation. In our case this may not only result in an improvement in efficiency to represent the significant features with respect to the raw histogram, but it presents also much better generalization capa-

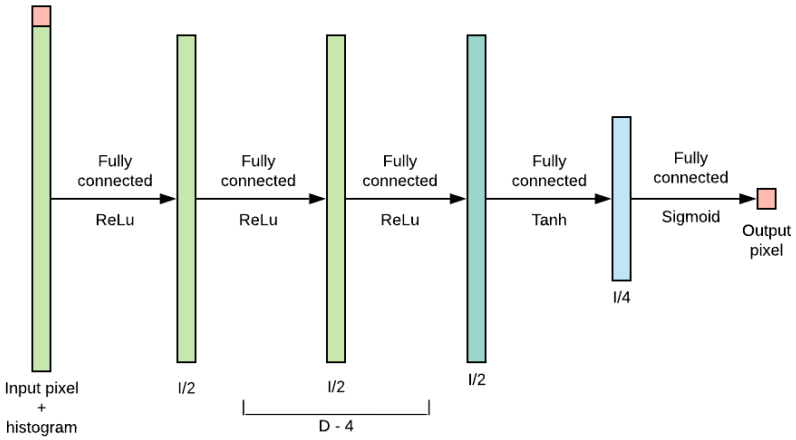


Figure 2: The The Pixel-wise Fully connected parametric architecture depending on hyperparameters $\mathcal{H}$, $\mathcal{D}$, $\mathcal{C}$.

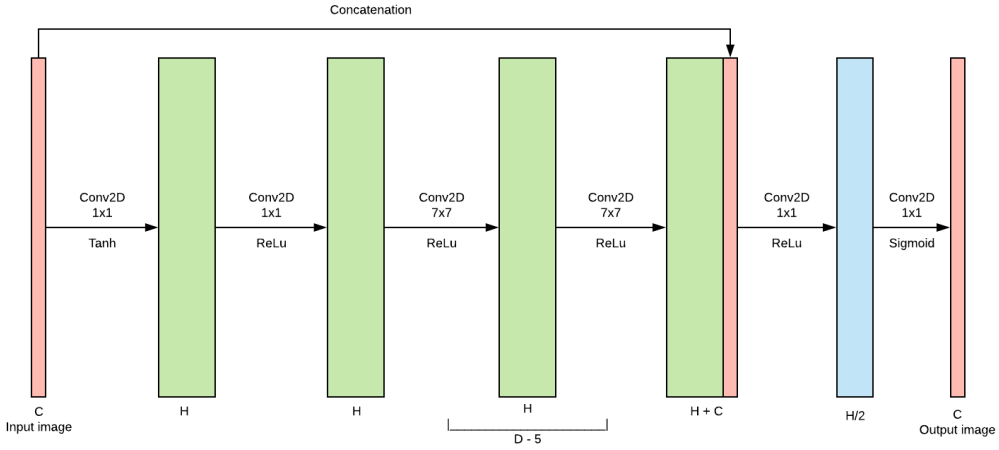


Figure 3: The Histogram CNN parametric architecture depending on hyperparameters $\mathcal{H}$, $\mathcal{D}$, $\mathcal{C}$.

bilities to learn different color transformations, eventually in a context-aware manner which is typical of expert made photo enhancements. The particular convolutional architecture is designed to learn the (local) histogram feature more directly and easily by pruning paths which are not needed.

We designed the architecture in Figure 3 by breaking the problem of histogram equalization in three steps:

- **Pixel bin classification**: the fist step is classifying each pixel independently to assign it to a specific bin out of the $\mathcal{H}$ possible bins. This classification needs to identify convex intervals on a line, so it will require only two layers. Moreover, each pixel needs to be classified independently, thus unitary sized kernels are preferred. As a result the input image with $\mathcal{C}$ channels is passed through two convolutions both with $\mathcal{H}$ filters with unitary kernels.

- **Local features aggregation**: once the pixels have been independently classified we need to aggregate this information. A sequence of large kernel convolutions is needed at this step, allowing to collect the local histograms and leaving space to learn different local features that can be useful in the context of expert photo enhancement. Consequently the $\mathcal{H}$ wide histograms are passed through $\mathcal{D}-4$ convolutions with square kernels of size 7 and relu activations to sum up the local histograms.

- **Inference**: this last step is learning the pixel-wise function $\mathcal{F}_{pw}$. The previous steps could extrapolate the local features $\mathcal{L}$ and potentially global features $\mathcal{G}$ but we have no guarantee that the corresponding pixel information $\mathcal{P}^i$ is preserved. We structurally ensure that $\mathcal{P}^i$ is preserved by concatenating the original image to-

gether with the available features. As $\mathcal{F}_{pw}$ is computed independently for each pixel position, the convolutions at this step have again unitary kernel size. On the architecture the feature aggregation output with size $\mathcal{H}$ is concatenated with the $\mathcal{C}$ channels of the original pixels, resulting in a $\mathcal{H}+\mathcal{C}$ channels inner image features representation. This is finally passed through two convolutions with unitary kernel size and filters $\mathcal{H}/2$ and $\mathcal{C}$ respectively with relu and sigmoid activation function to output the final image in normalized pixel values.

4. Experiments

The experimental setup for this paper is divided into two main phases:

- **Learning histogram equalization**: in the first phase we evaluate the proposed models for the task of histogram equalization. Both the pixel-wise fully connected and the histogram convolution are evaluated on the task and compared with the performances of a generic plain convolutional neural network with 3x3 filters.

- **Learning expert photo enhancement**: in the second phase the best performing architecture for the histogram equalization task is tested onto the task of expert photo enhancement. At this step we compare it with a plain convolutional neural network and a U-Net inspired [6] convolutional neural network. The latter is particularly interesting in this phase as it has been shown to be well suited to catch semantic features, which are very significant for this task.

All comparative architectures are reported in Figure 5 for clarity.

4.1. Histogram equalization

The first experiments have been run on the task of learning histogram equalization. The used dataset is cifar10 [5], comprising 60000 RGB images of size 32x32 pixels. The images have been preprocessed to be grayscale to allow for lighter models and better qualitative assessment of results, thus setting hyperparameter $\mathcal{C} = 1$. The ground truth of equalized cifar10 images have been generated using the standard histogram equalization algorithm with 128 bins; the loss function used is the classical mean squared error with respect to the ground truth images, which has been used to quantitatively assess the results as well. All models have been trained with ADAM optimizer [3] and learning rate of 10^{-4}. In Table 1 we show the pixel mean squared error for 500 samples of the cifar10 validation set for both pixel-wise fully connected and convolutional architectures (CNNs). The plain CNN is intended as a classical convolutional neural network with square kernels of size 3 as shown in Figure 5.

The Histogram CNN far outperformed both the Plain CNN and the Pixel-wise FC, being resilient to overfitting even with highly complex models. Notice that the best performing value of network depth $\mathcal{D} = 12$ is the smallest value such that the receptive field of the network encloses the full size of the used images. The Pixel-wise FC instead showed much worse results even though it was fed with perfect features in input. It must be underlined that extending each pixel with its histogram feature greatly increased the memory requirements of training and made it much more lengthy in time, thus forcing us to train the model on 10000 images rather than the full 50000 training images used for the other models. Moreover, the pixel-wise FC model proved to be extremely prone to overfitting and required heavy regularization with an additional 0.8 dropout layer before the output layer.

Architecture	$\mathcal{H}$	$\mathcal{D}$	Parameters	MSE
*Pixel-wise FC	64	10	10k	1.45e-2
Histogram CNN	128	12	6.4M	**1.6e-3**
Histogram CNN	32	12	400k	4.1e-3
Plain CNN	128	10	1.2M	1.1e-2

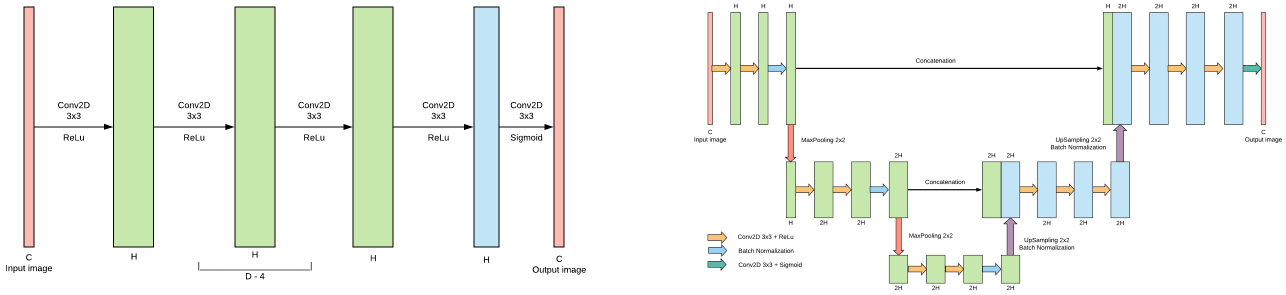
Table 1: Metrics of experimental results on the Histogram Equalization task. Starred models have been trained with less training samples.

4.2. Expert photo enhancement

The generalization capability of the best architectures on the histogram equalization task have been tested on the task of expert photo enhancement. The MIT-Adobe FiveK dataset [1] has been used for the evaluation: this includes 5000 quality images of a broad range of scenes and their ground truth version freely enhanced by an expert using Adobe Lightroom. Five different ground truths from five different experts are available, we took ExpertB only as a reference ground truth. We randomly split the dataset into training and validation set with a 5:1 proportion. As a preprocessing step all images have been resized to a standard 500x332 pixels format (rotating landscapes to keep proportions); color has been kept as it is usually an essential element in photo enhancement, thus setting hyperparameter $\mathcal{C} = 3$. All models have been trained with ADAM optimizer [3] and learning rate of 10^{-4}, using only 1000 training images due to hardware limitations. In Table 2 we show the pixel mean square error for the first 500 validation samples.

The Histogram CNN did not perform significantly better than the other architectures, but still it obtained comparable

Figure 4: Samples of grayscale histogram equalization output results. First row: original. Second row: ground truth. Third row: Histogram CNN. Fourth row: Plain CNN. Fifth row: Pixel-wise FC.



(a) Plain CNN

(b) U-Net

Figure 5: The comparative classical architectures used in our experiments.

performances with much less parameters and did not perform better by adding further complexity. The two architectures tested for comparison required much complexity to obtain comparable performances, but in the end no architecture could clearly outperform the others.

Moreover, from a qualitative assessment of the output of all networks as in Figure 6, we can notice that high frequencies in the input image are usually not preserved by any of the tested architectures, resulting in a slight blur effect.

Architecture	$\mathcal{H}$	$\mathcal{D}$	Parameters	MSE
Histogram CNN	64	7	611k	**9.5e-3**
Plain CNN	128	10	1.2M	1.0e-2
U-Net	256	11	5.3M	9.7e-3

Table 2: Metrics of experimental results on the expert photo enhancement task. For each architecture the best performing hyperparameter configuration has been reported.

5. Conclusion

We designed a novel convolutional architecture with a structural prior to learn histogram equalization and showed that it greatly outperforms classical architectures at this task. We showed that the new architecture achieves similar results on the task of learning expert photo enhancement with respect to classical alternatives while using much fewer parameters, but we still lack clear evidence that it can achieve better performance.

All architectures tested on high resolution suffered a slight blur effect. This problem could be addressed in future work by adding further skip connections directly from the input image to the last layer; alternatively, as Yan *et al.* [7] suggest, learning a transformation of pixels instead of mapping pixels directly can help preserving high frequencies. Moreover, as the histogram CNN is clearly efficient in learning color transformations, its performances could be improved by coupling it with an architecture specialized in understanding semantics, such as the U-Net.

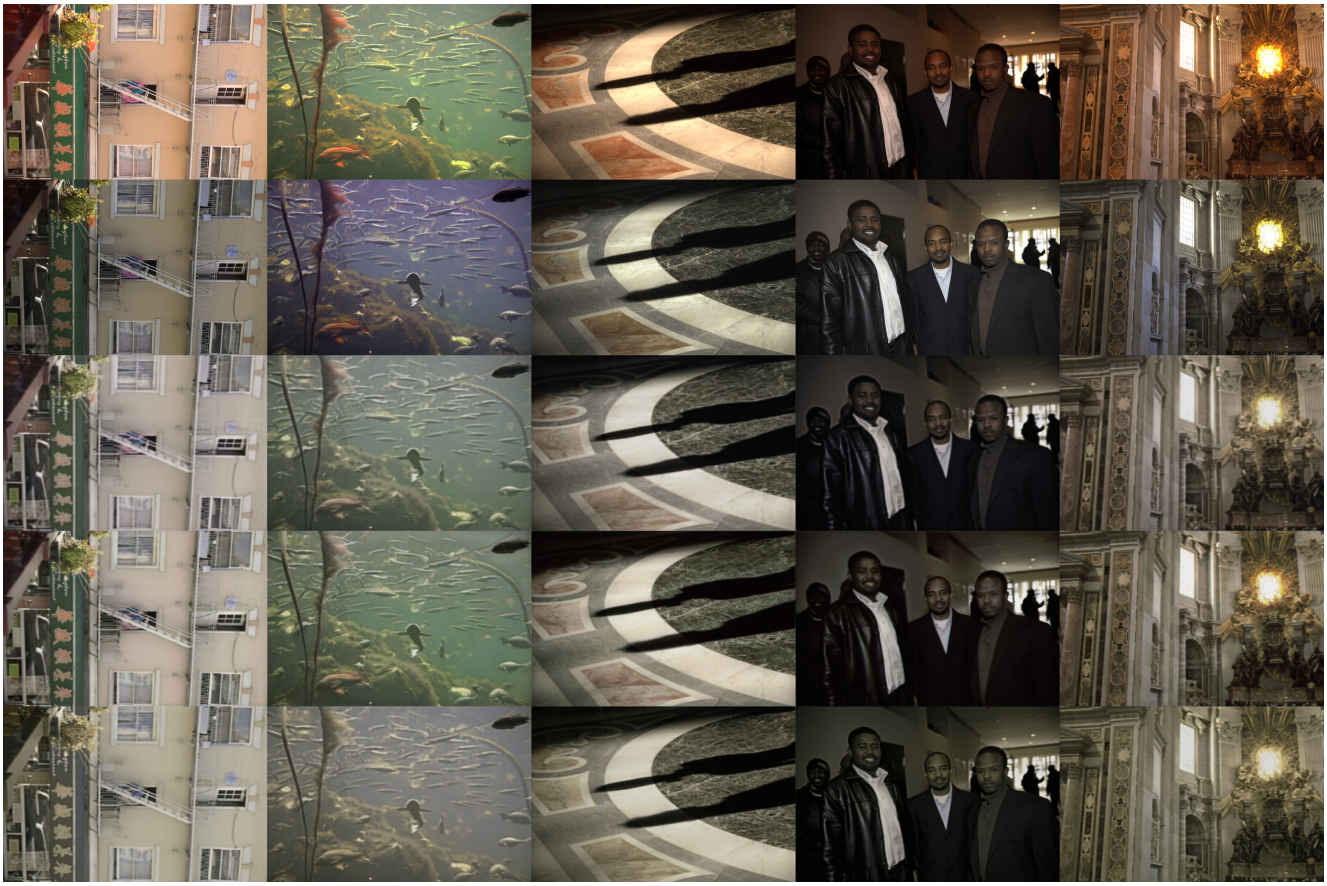


Figure 6: Samples of expert photo enhancement output results. First row: original. Second row: ground truth from ExpertB. Third row: Histogram CNN. Fourth row: Plain CNN. Fifth row: U-Net.

References

[1] V. Bychkovsky, S. Paris, E. Chan, and F. doDurand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 4

[2] L. Kaufman, D. Lischinski, and M. Werman. Content-aware automatic photo enhancement. In *Computer Graphics Forum*, volume 31, pages 2528–2540. Wiley Online Library, 2012. 2

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[4] Y. Kinoshita and H. Kiya. Image enhancement network trained by using hdr images. *arXiv preprint arXiv:1901.05686*, 2019. 2

[5] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 4

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[7] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):11, 2016. 2, 5