Colourisation of black and white photos

Haaris Hayat

20/12/2022

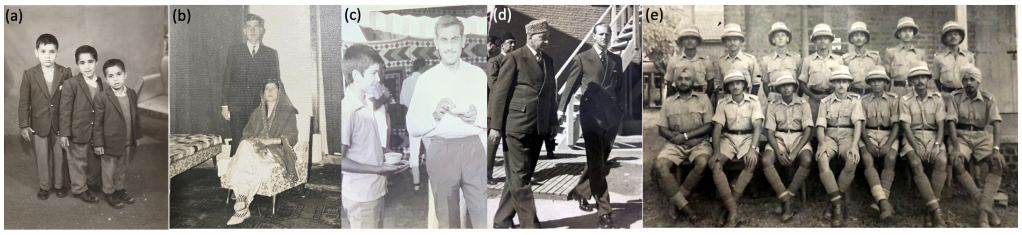

Figure 1: Old family photos: (a) Picture of my father and his siblings; (b) Wedding photo of grandparents; (c) Grandfather; (d) Prince Phillip's visit to Pakistan with a relative (e) Grandfather during WWII.

1 Introduction

In this paper, I will explore the use of various image processing and machine learning techniques to enhance and add colour to old black and white family photos (a process I will refer to as "colourisation").

Old photos contain a lot of history and memories, and can be a means to explore the lives of our ancestors. However, given the age of the photos, the quality of the equipment used to capture them, and the impact of time, it can be difficult to fully appreiciate the pictures. This is the challenge I wish to overcome through the course of this project.

2 Datasets

I will be using two datasets for the project. The first contains digitised versions of old black and white photos of various family members. This dataset has not been pre-processed in any way apart from digitising the original hard copies. The pictures in this dataset are shown in Figure 1.

The seond dataset contains a set of personal photographs which I have captured. The aim of this dataset is to provide a way to assess how well the various colourisation techniques work since I have the original colour photographs.

It should be noted that I will not be training any models myself (although this is an interesting task worth exploring) for the colourisation and will instead use readily available models and assess their effectiveness.

Many of these pre-trained models have been tested extensively on a range of photos (usually simple photos), I therefore wanted to test them on a

personal dataset that had never been used to train on test the models by anyone. For this reason, I have handpicked photos for the second dataset that will push the models to their limits. They contain a range of different terrains, colours, times of days, and buildings. In many old family photos, we are not at a liberty to choose the backgrounds and the model therefore needs to be able to cope with complex scenes.

Since the second dataset was captured using a high resolution camera, each photo was around 30mb each which was unnecessarily large for our purpose. Therefore I had to lower the resolution to 2400x1800 pixels. No other preprocessing has been performed.

3 Methodology

In this section, I will discuss some of the techniqies that can be used to process the photos as well as providing a brief overview of the general process.

3.1 Denoising

Given the age of the pictures, and the fact that they were physical copies subject to the environment, they have suffered quality defects which should be fixed before proceeeding to the colourisation phase. Most of the pictures contain a lack of detail, specifically around faces. However this is due to the limitation of the cameras used and cannot be enhanced through the use of a filter. There have been interesting developments in this field, most notably by Wan et al [1], however this is an entirely different problem and requires its own project even though it would be nice to have an end to end pipeline to both enhance and colourise old photos.

Image (a) from Figure 1 is the most recent and has the best quality. The other images are decent enough to be used straight away apart from (b) and (c). Image (b) looks very washed out and desaturated, and Image (c) has been digitised form a canvas painting resulting in a grid of white dots on the image. There are several commonly used filters I will apply to remove the noise artifacts such as mean/median filters. For the washed out picture, it is unlikely that filters will help (even the minimum filter) due to the fact that even the surrounding pixels would be desaturated. We will therefore apply gamma correction to this photo. This process adjusts the pixel values by a given gamma factor using the formula $O = I^{gamma}$ where I and O are the pre and post adjustment pixel values respectively.

3.2 Colourisation

Cheng at al [2] proposed the very first deep learning based image colouriser, which later techniques were based on. According to this paper, there are two broad categories of image colourisers:

1. Scribble based colourisers - These require users to input their expectations of the colours on certain parts of the pictures. These expectations are then propogated to the rest of the picture. There are two problems with this approach for our use case. The first is the time and effort required by the user. The second problem is the fact that we may not know the colours ourselves unless there is another reference photograph.

2. Example based colourisers - These colourisers use similar photographs to infer the colours. This can be through the user providing a photo that is similar (again may not be suitable for our use case) or through the use of a large data base of photos to train a model.

We are going to be mostly interested in the example based colourisers based on deep learning models. All of the approaches work on the LAB channels rather than the standard RGB channels that are commonly used for most image classification models. In greyscale images, we have the L-chnnel (lightness), and so need to predict the two colour channels A and B. The A channel represents the balance between green and magenta while the B channel represents the balance between green and yellow.

Note that whilst we can infer the colour of the sky, it is difficult to know what colour a certain object is. For example given an image containing clothes of a certain lightness, the same lightness could be achieved by several different clothes and unlike the sky, we may not know the fashon preferences of the people in the images. These models therefore provide plausible colourisation and not necessarily the ground truths for these objects.

3.2.1 CNN based models

The deep learning model trained in [2] is a fairly simple CNN with 3 hidden layers which predicts the values in the A and B channels. The model was trained uing a Eucliden loss function which seems intuitve given the problem can be thought of as a regression problem for each pixel and each channel. If we let $Y_{i,j}$ be a two dimensional vector representing the ground truth values

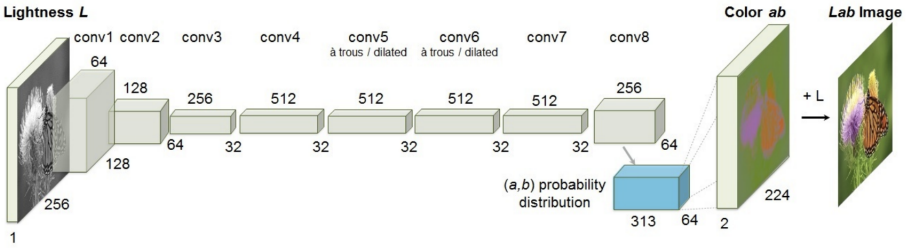

Figure 2: CNN architecture used by Zhang et al.

of the AB channels for the pixel with coordinates (i,j), and let $\hat{Y}_{i,j}$ be our prediction for that pixel, then the L2 loss function can be written as:

$$L_2(Y, \hat{Y}) = \frac{1}{2} \sum_{i,j} \left\| Y - \hat{Y}_{i,j} \right\|_2^2 \tag{1}$$

What makes this model interesting is actually the process before the CNN. Their model first clusters images into 47 categories, and trains a different model for each category. The models themselves don't just take the L-channel values, they also take information about the surrounding patch (e.g. the surrounding pixel values and a high level descriptor of these values) in addition to a symantic discriptor of the pixel (for example the type of object the pixel belongs too). The idea behind the approach is that the L-channel by itself cannot provide enough context about the pixel in order to predict the colour.

The model proposed by Zhang et al. [3] builds a more complex CNN as shown in Figure 2. The image is intially compressed to a resolution of 256x256 before being processed by each convolutional layer. Each convolutional layer actually consists of series of 2 or 3 sub-convolutional layers as well as a ReLu layer. In addition, BatchNorm layers are used in between each hidden layer, which aim to prevent internal covariate shift through normalising the inputs for the next layer and therefore stabilising the network during training. The resolutions are adjusted between the layers through spatial down-sampling and up-sampling.

There are several differences between this model and the one proposed in [2], the first being the complexity of the CNN. The secod difference is the loss function used to train the CNN. While a Euclidean loss function makes sense in theory, in reality, an object may take a variety of different AB values and the value that should be used to assess predictions should be the mean of

these values, otherwise the model starts to make prudent predictions leading to grey/desaturated images. Zhang et al. therefore propose a different approach, where the AB values are split into 313 bins and the problem is treated as a multinomial classification instead. Instead of predicting the pixel values, we predict $\hat{Z}_{i,j,q} \in [0, 1]$, the probability of a pixel (i,j) belonging to bin q. The ground truth values, $Y_{i,j}$ are transformed into a vector $Z_{i,j}$ using on-hot encoding for each of the bins. These values are then used to form the multinomial cross-entropy loss function:

$$L_{CL}(Z, \hat{Z}) = -\sum v(Z_{i,j}) \times Z_{i,j,q} \times log(\hat{Z}_{i,j,q}) \tag{2}$$

where the term $v(Z_{i,j})$ is a rebalancing factor for each pixel colour. The purpose of this balancing is to prevent the loss function being dominated by colours with low AB values (such as colours of clouds, water, sky etc.) which tend to be common in pictures. This is equivalent to resampling from the colour space to ensure rare colours are also featured in the training.

Lastly, Zhang et al don't apply the feature engineering (such as getting patch/semantic information). Instead, the more complex CNN, coupled with a much larger training set (1.3m compared to the 3k used in [2]) appears to be able to handle the task more competently. See [4]for a comparison of results.

I will be applying the model produced by Zhang et al. The reason for this is the improvement in performace and the fact that Cheng et al focus on scenic and environmental pictures rather than unnatural objects (such as cars, clothes etc.) or people for training, and thus it did not suit my use case. There are two variations of the model, ECCV-16 and Siggraph-17, both have the same architecture described above but slighlty different weights. Siggraph-17 can also perform user guided colourisation.

3.2.2 GAN based models

Generative Adveserial Networks ("GANs") are a state of the art technique that can be used for image to image translations. For example, the Pix2Pix model created by Isola et al [5] can not only be used to convert black and white pictures to colour, but talso to change the style of the photo from night to day or generate an image from just edges. Note that we will only provide a high level overview of GANs in this paper.

GANs consist of two neural networks. The first is the generator, which generates images in an attempt to fool the second network. The second network, the discriminative network, learns a loss function to predict whether

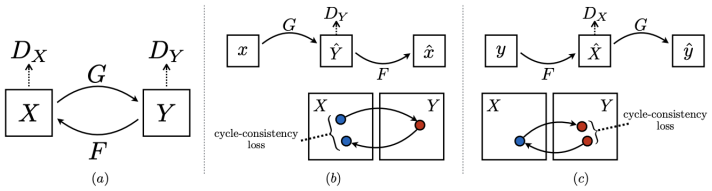

Figure 3: Cycle-GAN architecture deployed by Zhu et al.

the generated image is real or generated. The first network tries to minimise this loss and a tug of war ensues between the two models. the idea behind using GANs is to uncover an inherent mapping between two picture spaces (for example greyscale and colour image spaces) without the need for explicit pairs of images that would be used in a typical supervised learning approach.

The Pix2Pix model usese conditional GANs, or cGANs where the networks are conditioned on an image. For example, given a black and white image, the generator tries to generate colour version while the discriminator examines the output. This results in a learned mapping from the black and white image space to the colour image space.

Zhu et al [6] deploy a variation called Cycle-Consistent Adverserial Networks or cycle-GANs as shown in Figure 3. This involves the use of two cGANs between two image spaces X and Y. The first cGAN, generates a mapping G from X to Y while being opposed by the discriminator D_Y as shown by (a) in Figure 3. Similarly, the second cGAN generates a mapping F from Y to X with associated discriminator D_X. To further regularise the mappings, a cycle consistency loss is used. The idea is that if we take $x \in X$, then $F(G(x)) \in X$ should be close to x. The cycle consistency loss measures how far apart these two objects are. Similarly we could take $y \in Y$ and compare this to $G(F(x)) \in Y$. The concept of cycle consistent losses is shown in (b) and (c) of Figure 3.

Unfortunately, these two models don't offer any pre-trained weights. I have therefore used the DeOldify library [7] which provides the option of using pre-trained model weights. DeOldify also used cycle-GANs, however it uses a novel approach called No-GANs to train the model. In this approach, the Generator is trained first, followed by the discriminator, before they are trained together in the usual way. The author claims that this results in a much lower training requirement.

There are two sets of pre-trained weights available in the DeOldify library:

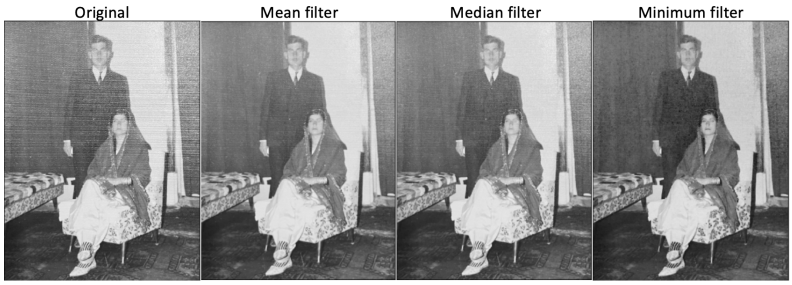

Figure 4: Application of various filters to clean noisy image

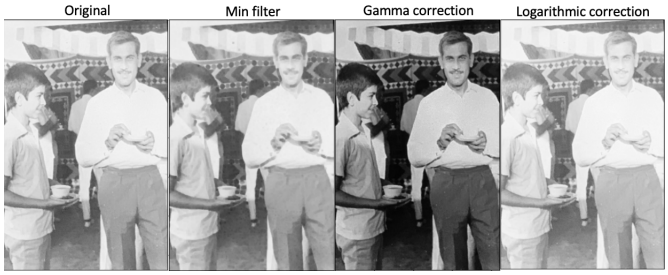

Figure 5: Correction for desaturated picture

Artistic and Stable. Artistic is supposed to provide more vibrant colours while Stable should perform better with scenery. We will deploy both of these and explore the results.

4 Results

4.1 Image Denoising

In Figure 4 we can see the effect of various filters to attempt to clean the picture from the white specks. It appears that the min filter and the mean filter are the most effective.

As expected, we can see in Figure 5 that the application of filters does not fix the issue of desaturation. In fact the minimum filter results in a loss of detail, especially around the eyes. The most effective method appears to be gamma correction. However, with all of these corrections, we will apply the colourisations on both the original and cleaned images since the corrections themselves may have a negative impact on the colour mapping.

4.2 Colourisation of synthetic black and white photos

In this section, we will look at how well the colourisation models discussed in Section 3 work on images which I have synthetically transformed into black and white images. The advantage of this approach is that I have the original images for comparison. Note that the purpose of this project is to create plausible colourisations of historic photos. Therefore I will not be calculating any accuracy metrics to see which model maps images closest to the original as this can result in grey images. Instead, I will be visually inspecting the results to assess the models. Accuracy score can be improved with prudent colour estimates which isn't what we want for the use case. In addition, scores will be penalised more if a car is red when it should be blue than if skin is slightly more orange even though the latter will look more odd.

Note that I have not included all of the pictures that were used for testing in this document, but have included these in the associated GitHub repository for the project.

Figure 6 shows the original colour images (numbered), as well as the colourisations performed by the four (variations of) models which we have used. In general, ECCV-16 had disappointing results, and appeared to have an orange tint in every picture. The only time this helped was in capturing the colour of sunsets for example in image 6. There were certain limitations which all models suffered from such as colourising objects and buildings (see images 5 and 7), however this was expected. There were also some aspects which all the models seemed to be able to tackle easily (apart from ECCV-16) such as the snow in image 1, the Cutty Sark in Image 6 which was surprising given the low light setting.

Siggraph generated the most vibrant colours, for example the grass in images 3 and 5, however the model struggled with colourising people and clothes. images 8 and 9 colourised by Siggraph look very odd.

Despite not having the most vibrant colours, the DeOldify colourisations seemed to be the most realistic, especially when it came to skin tones and clothes (see images 8 and 9). Artistic seemed to perform the best in the night time photo (image 4) and for colouring in skin tones and clothes in images 8 and 9. However, the buildings and scenery were very desaturated. In comparison, the Stable model performed better when it came to scenic colourisations for example in image 2. The grass in image 5 was greener and closer to the original when colourised by the Siggraph model. In fact this model had the best colourisation of the building in image 5. The water in

Figure 6: Comparison of different models

Figure 7: Colourisation of old photos using the Artistic model (row 1) and the Stable model (row 2).

image 2 was also very grey in the Artistic colourisation. Both the Artistic and Stable models performed well on the desert shot in image 10 and the farm in image 3.

When it came to colourising people, it was interesting to see that it was easier to change the parameters to colour skin tones well, however where the models struggled were shadows on clothes. The parameters had to be tuned to prevent these shadows showing up as different colour patches on clothes.

4.3 Colourisation of old family photos

Now that we have an idea of the strength of the models, I will apply the DeOldify models to the old family photos from Figure 1 since these models showed the strongest performance. The results were assessed by firstly inspecting if there were any unusual or implausible colours. Next, the pictures were shared with family to see if the colours were in line with their recollection of the people.

Figure 7 shows the results of the two models. The first thing we observe is that the Stable model generally had a slightly greener tint than the Artistic model. The first image was colourised very well, although the sleeves of the blazers begin to turn blue in the Artistic colourisation. The Artistic model

Figure 8: Colourisation of pre-processed images

showed more vibrant colours in the second image, however did not perform as well when it came to colourising by grandfather in the background. The third image was colourised well by the Artistic model, with hints of red being picked up in the background which is accurate.

The stable model did not colourise the fourth image well, with the skin tones showing up as grey. In comparison, the Artistic model performed excellently, although the hands were still grey. In the final image, the stable model performed better, with the greener grass bringing the picture to life. However both models struggled with the darkening of the image on the sides which was a defect with the original image.

Feedback from family was very positive, especially regarding skin tones and generally the colourisation of people (apart from the shortcomings mentioned above). However, there were some interesting differences from reality. For example, my father remembered that the blazers in image 1 should have been blue, whilst the jumpers should have been maroon. In addition, the rug in image 2 has survived the test of time and I can attest that is a dark red in reality!

Figure 8 shows the colourisation applied to the pre-processed images from section 4.1. I found that the gamma-corrected image of my grandfather, coupled with the Artistic colouriser offered the most detailed colourisation of my grandfather, at the expense of the background losing colour. The minimum and mean filters coupled with the Artistic colouriser reduced the noise from the original image. In addition, the minimum filter seems to have increased the contrast of the final details, making the colourisation jump out even more. Similarly the Stable colourisations were improved by the pre-processing, the image of my grandfather has more colour than the Artistic colourisation.

5 Conclusion

Overall the models have shown strong results, despite the limitations around objects/buildings which were expected. Since sharing the results with family, I have been bombarded with requests to colour more photos which perhaps attests to the success of the techniques. However I think that I have only scratched the surface and more could be done to improve results. For example Zhang et al [8] propose a method to colourise with interactive inputs from users. Whilst this may not be appropriate on its own, it could be used to refine the results we have achieved. For example, to fine tune the colour of objects. Using this technique, we could even take the skin tones from the face, and extrapolate to the hands in the picture with Prince Phillip where the hands were grey.

Each model had its own strengths and weaknesses, Siggraph had more vibrant colours, Artistic performed better on people and Stable on scenery. I think that the next step for a colourisation model is to use the approach of Cheng et al in [2] coupled with cGANs to train different models for different objects to get the best of both worlds.

References

[1] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, Fang Wen. Bringing Old Photos Back to Life. Microsoft, 2020.

[2] Zezhou Cheng, Qingxiong Yang, Bin Sheng. Deep Colorization. ICCV, 2015.

[3] Richard Zhang, Phillip Isola, Alexei A. Efros. Colorful Image Colorization. University of California, 2016.

[4] https://richzhang.github.io/colorization/resources/deep_colorization_comparison.html

[5] Isola, Phillip and Zhu, Jun-Yan and Zhou, Tinghui and Efros, Alexei A. Image-to-Image Translation with Conditional Adversarial Networks. Computer Vision and Pattern Recognition (CVPR), 2017.

[6] Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Computer Vision (ICCV), 2017.

[7] Jason Antic. https://github.com/jantic/DeOldify

[8] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, Alexei A. Efros, Real-Time User-Guided Image Colorization with Learned Deep Priors. 2016.