Session 6: Applications, Architecture and Systems Integration

IMVIP 2019: Irish Machine Vision and Image Processing

2019

Spatial Coherency in Colourisation

Sean Mullery
Dublin City University

Paul F. Whelan
Dublin City University

Spatial Coherency in Colourisation

Seán Mullery & Paul F. Whelan

Vision Systems Group, School of Electronic Engineering, Dublin City University, Dublin 9, Ireland

Abstract

Automatic colourisation is the function of inferring colour information from a grey-scale prior and then combining the colour with the grey-scale to form a colourised version of the image. We identify Spatial Coherence as a particular weakness in methods that use Convolutional Neural Networks for colourisation. Generated colours do not adhere to semantic edges and are not consistent within boundaries where we would expect uniform colour. Spatial Coherence, while often evident to the human eye, does not yet have an objective metric. We show, by segmentation of the combined ab channels of the CIEL*a*b* colour space, that a segmentation based on CNN colourisation is poor. We argue the need for the development of metrics to evaluate a colourisation's performance on Spatial Coherence.

Keywords: Colourisation, Segmentation, Computational Photography, GAN.

1 Introduction

Automatic colourisation is the function of inferring colour information from a grey-scale prior and then combining these to form a colourised version of the grey-scale image. Colourisation, therefore, is an ill-posed problem. In the case where we start with an 8-bit grey-scale image **L** channel and wish to infer a 24-bit colour CIEL*a*b*, there are 2^{16} possible values for every grey level in the prior **L**. The difficulty is compounded by the lack of an objective measure of the quality of a colourisation result.

Take the case of a ground truth image in CIEL*a*b*. We can use our chosen colourisation algorithm to predict our **ab** using just **L**. A simple ℓ_2 distance may tell us how far our result is from the ground truth. The CIEL*a*b* space was developed to map perceptually uniform changes in colour but these changes were only measured over small changes in colour, hence ℓ_2 distance is an appropriate measure of the difference between two colours in the CIEL*a*b* space, but only over small distances. In most other colour spaces it is not applicable. Using ℓ_2 distance also assumes there is only one appropriate colourisation for any grey scale image. This is problematic in terms of colourisation. For some natural objects, there will be a constraint on the plausible colour of objects; Grass should be green; Wood should be brown; The sky is most probably blue but can be anything from a distribution of colours. Human-made objects have little constraint in terms of colour. We cannot assume that an item of clothing can only be one colour.

Most importantly, in terms of our discussion here, however, is that ℓ_2 distance takes no account of spatial coherence. There can be a variety of distances between the ground truth colour and other plausible colours. Indeed a subjectively poor colourisation in which a semantic object has a random patch work of colours all close to the ground truth colour will be deemed a better colourisation than a semantic object with a single spatially coherent colour, if that colour is a greater distance from the ground truth colour that the mean distance of the random colours.

2 State of the Art

Before the recent popularity of deep learning, many works on colourisation relied on some user input or a donor colour image which could supply a suggested colour based on the similarity of the grey-scale image statistics.

In general, spatial coherency was considered a guiding constraint on finding good colourisation. Where the grey-scale prior was homogeneous the chrominance applied should also be homogeneous; Discontinuities in the grey-scale prior should likely see discontinuities in the chrominance channels; Areas of similar grey-scale texture should probably have the same chrominance associated with them.

The deep learning approaches concentrated on a per-pixel prediction of colour based on the grey-scale prior. While CNN's can have a large receptive field, due to their depth, long-range spatial dependencies are considered a weakness of the architecture. [Iizuka et al., 2016] introduce an elaborate setup of four networks predicting low, medium and global level features which fed to a colourisation prediction network. They believed the best route to colourisation is to consider details in the grey-scale image at many levels of abstraction.

[Zhang et al., 2016] framed the colourisation problem as a classification task. The goal is to predict plausible colourisations that can fool a human observer. They predict a distribution of colours for each pixel and the loss is re-weighted during training to emphasise rare colours which encourages diversity. [Larsson et al., 2016] concentrated on systems that could learn a histogram (distribution) of colours for a given grey-scale pixel. They consider the problem as semantic composition and localisation. To do so, they take an ImageNet pre-trained VGG-16 network [Simonyan and Zisserman, 2015] and concatenate features from multiple layers into a hyper-column [Hariharan et al., 2015]. This is fed to a fully connected layer and in turn connected to the output predictors. The predictors must predict a binned distribution of **ab** space or a binned distribution of the hue and chroma. The ground truth is the binned histogram of **ab** or hue and chroma from a region surrounding the centre pixel in the ground truth image. The loss is the KL divergence between the ground truth histogram and the predicted one. Zhang et al. extended their work in [Zhang et al., 2017], where the network learns not just a mapping from grey-scale to colour but also allows for sparse user hints at the pixel level as well as global hints at the statistics level. Many colourisation systems will only produce one plausible colour for an object. With the system in [Zhang et al., 2017], the user can direct it to other plausible colours. The main aim is to get from credible to correct colourisations. Following the ideas of [Larsson et al., 2016] they use a hyper-column from throughout the network layers, rather than merely inferring the distribution from the final layer. The Global Hints network works on the global statistics of an image, based on the quantised **ab** space and on the global saturation distribution using the HSV space.

The pix2pix formulation [Isola et al., 2017] is the seminal work in colourisation using Generative Adversarial Networks (GAN), despite colourisation only being one of the demonstrated applications in that paper. Indeed, while Isola et al. make available the weights for many of the other applications, the colourisation weights are not available for testing. Pix2pix splits the job of minimising the error in the generated samples across two loss functions. They claim that for low-frequency information, ℓ_1 pixel error is sufficient. They do this at the output of the GAN's Generator network where they compare, pixel-wise, the real **ab** to the fake **ab**. This produces blurry images, so to enforce quality high-frequency details, they use the GAN's Discriminator, but only for patches of the image. This is motivated by increasing evidence that classification type network architectures are most often looking for texture rather than overall structure. [Nazeri and Ng, 2018] built on the work of pix2pix, focusing solely on colourisation and making a few changes in line with best practice for the training of GANs. They also make the weights available, which gives a reasonable substitute for Isola et al.'s performance on the colourisation task. Deoldify [Antic, 2019] is an unpublished but accessible on-going work that seeks to both restore old images as well as colourise them. Antic started with a pix2pix [Isola et al., 2017] setup but modified the down-sampling side of the generator U-net to a ResNet-101 [He et al., 2015] that was pre-trained on ImageNet. He added many state-of-the-art techniques, such as Spectral Normalisation [Miyato et al., 2018], Two Time-Scale Update Rule [Heusel et al., 2017] and Self Attention [Zhang et al., 2018]. In particular, the addition of Self Attention seemed to bring a significant improvement in spatial coherence. He also introduced his own unpublished GAN training method called NoGAN.

3 Method

We take colour images of size 256×256 and convert to the CIEL*a*b* space. We use trained instances of methods in Section 2 to generate the **ab** channels. The **ab** channels are segmented using a K-means clustering algorithm [Lloyd, 1982] with K=8. We use the OpenCV implementation of K-Means. The **ab** channel of the ground truth colour image is also segmented. While the segments will not be identical in the generated colourisations and hence the false colour applied to the segments will be different, each segmentation should look similar if the colourisation method has achieved good spatial coherence.

4 Results

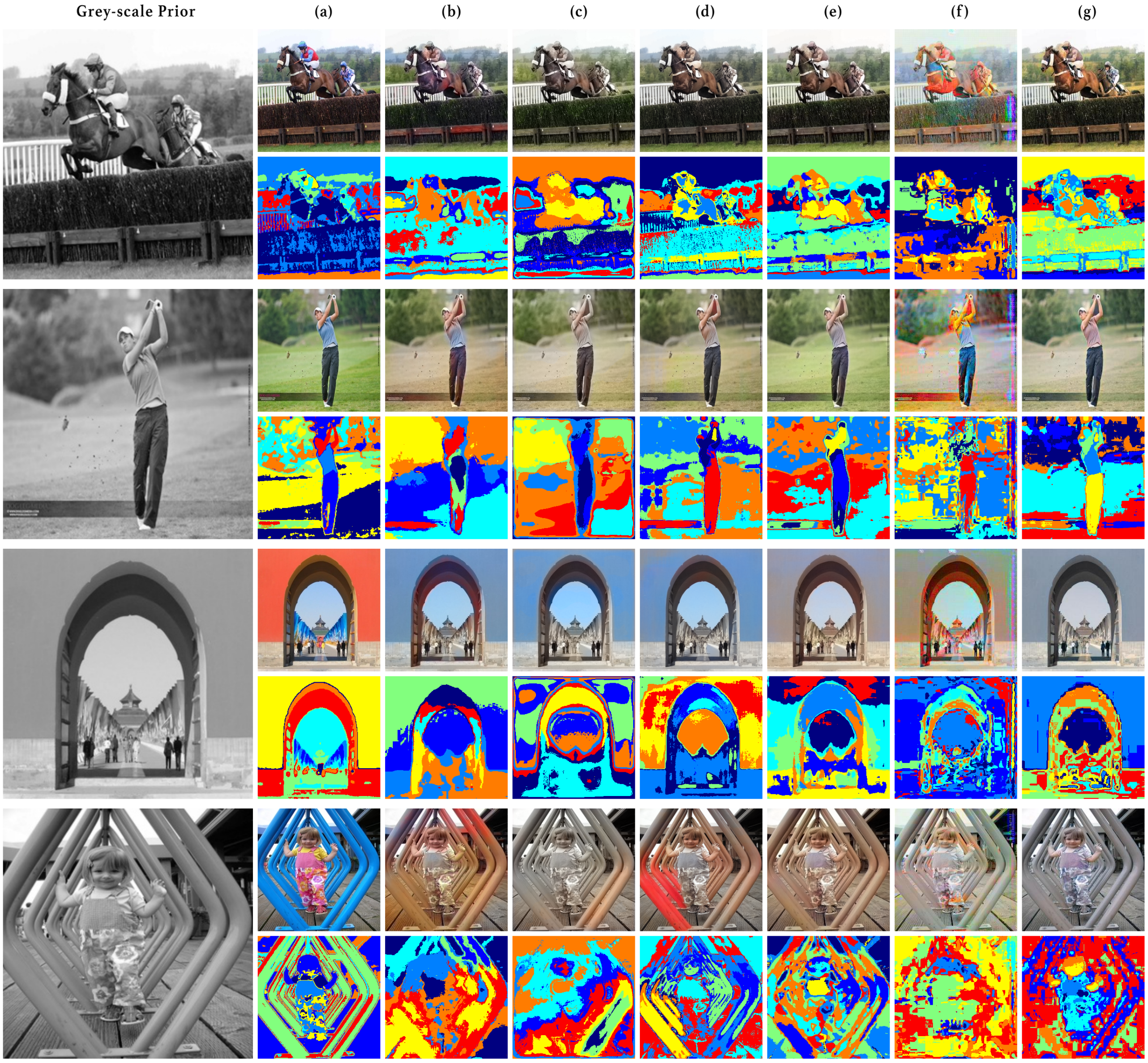

Figure 1: In the left hand column we have samples of the grey-scale priors that are input into each of the state-of-the-art algorithms. Column (a) is the ground-truth colour image from which the greyscale prior was derived. The other columns are as follows. (b) [Zhang et al., 2016] (c) [Larsson et al., 2016], (d) [Zhang et al., 2017], (e) [Iizuka et al., 2016], (f) [Nazeri and Ng, 2018], (g) [Antic, 2019]. The bottom row for each image shows a false colour segmentation based on the K-Means clustering of the **ab** channels for K=8

Figure 1 shows each segmentation of the **ab** channels that are predicted by the various state-of-the-art deep learning colourisation methods. Shown against the ground truth, it is clear from a visual inspection that many of the algorithms produce colourisations that are spatially incoherent. However, we cannot easily say that one algorithm is better/worse than another in this sense as there is not a clear objective measure of this.

5 Conclusion and future direction

There is an obvious problem with spatial coherence in colourisation using deep learning techniques. To improve this, we first need to quantify the problem objectively, which requires the development of an objective measure. In future work we plan to investigate and develop an objective measure for spatial coherence in colourisation.

References

[Antic, 2019] Antic, J. (2019). Deoldify. `https://github.com/jantic/DeOldify`.

[Hariharan et al., 2015] Hariharan, B., Arbeláez, P. A., Girshick, R. B., and Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 447–456.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385v1*, 7(3):171–180.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500.

[Iizuka et al., 2016] Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2016). Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph.*, 35(4):110:1–110:11.

[Isola et al., 2017] Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5967–5976.

[Larsson et al., 2016] Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

[Miyato et al., 2018] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks.

[Nazeri and Ng, 2018] Nazeri, K. and Ng, E. (2018). Image colorization with generative adversarial networks. *CoRR*, abs/1803.05400.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Zhang et al., 2018] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-Attention Generative Adversarial Networks.

[Zhang et al., 2016] Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. *CoRR*, abs/1603.08511.

[Zhang et al., 2017] Zhang, R., Zhu, J., Isola, P., Geng, X., Lin, A. S., Yu, T., and Efros, A. A. (2017). Real-time user-guided image colorization with learned deep priors. *CoRR*, abs/1705.02999.