A Rainbow From Shades of Gray: Video Colourization

Abhishek Agarwal
Computer Science and Applied Mathematics
IIIT Delhi
New Delhi, India
abhishek16126@iiitd.ac.in

Abhishek Maiti
Computer Science and Engineering
IIIT Delhi
New Delhi, India
abhishek16005@iiitd.ac.in

Surabhi S Nath
Computer Science and Engineering
IIIT Delhi
New Delhi, India
surabhi16271@iiitd.ac.in

I. Problem Statement

Our motivation behind this work is to be able to *"Colour the Past."* We aim to develop an end-to-end framework for meaningful and consistent colourization of black and white videos.

Fig. 1. Colourization of consecutive video frames

II. Introduction

Colour is a characteristic of human visual perception. We are naturally receptive to colour and light intensities. We have all experienced black and white content in some form or the other, either as photographs or movies of the past, but we all agree that black and white representations deny us of a very meaningful, significant and important feature – **colour**. Today, with immense advances in colour photo and videography, black and white colourization is merely a filter or effect applied over the original input.

Restoring colour information from a black and white representation has been a challenge for the deep learning community. Attempts to colourize images have been made for over 15 years with the discovery of newer approaches and optimization techniques which can improve the conversion. Initially, transfer functions, image transformations and methods like chrominance blending, colour film generation, cross-dissovling, correlation minimization and interpolation were used for the task and later machine learning and deep learning based methods have been applied. More recently, Bayesian Schemes and Convolutional Neural Networks have been applied to enhance image and video colourization. Nowadays, colourization is employed as an intermediate task for improving visual understanding. Since a video is made up of image frames, image colourization can be easily extended to videos through frame-wise colourization but since frames are dependent, we want to incorporate temporal information and context to determine colours of successive frames.

III. Dataset

We will be utilizing image and video datasets for our problem. Kaggle provides a grayscale image dataset link1. Alongside, any of the standard image datasets like Pascal VOC, COCO, CIFAR or Imagenet can be easily converted to grayscale. More day to day life image datasets can be found in link2 and link3. For video datasets, we will be using Youtube, the Kinetics dataset containing coloured videos on diverse human focused actions and DAVIS dataset containing 90 video sequences. Since these videos are coloured, we will first translate them to grayscale using frame-wise conversion.

IV. Related Work

Deep learning for image colourization is a well studied topic with the introduction of novel approaches each year [1,2,3,4]. However, video colourization using deep learning appears to be a comparatively lesser developed problem.

An interesting paper titled "Deep Video Color Propagation" was published in 2018 [5]. Here, colour transformations based on spatial and intensity distances for images and colour propagation based on edit distance of two consecutive ground truth images for videos have been employed. But these are computationally very expensive and cannot be used in large scale. Thus this paper has proposed using features from both the local and global images contexts to perform better. Two networks are constructed for local propagation and global transfer:

Short Range Propagation Network: This takes two consecutive gray scale images and tries to estimate a warping function which helps transfer colours of previous frame to the next frame. Spatially adaptive kernels are chosen.

Long Range Propagation Network: For long range color information propagation, the network needs to know image

features not only related to colour, but also to other aspects such as objects present and other semantic information in the image. For this deep features are extracted through reference frame matching. These features are then used to sample colours for the later frames. The dependency between reference (0^{th}) frame and the current (k^{th}) frame is modelled by computing the distance of their respective feature maps.

Fusion and Refinement Network: Using the above two methods, two predictions are obtained. A five layer Convolutional Network is used to combine these and predict the final output. It uses ReLU activation with increasing dilation to increase the receptive fields and a loss function incorporating image loss and warp loss.

The model achieved a PSNR (Peak Signal to Noise Ratio) of 43.64 whereas the previous best model achieved 42.72 over the 1st 10 frames. This level of performance was maintained for 1st 50 frames with PSNR score of 41.23 whereas the next best model only achieved 38.56.

Another paper performing image and video colourization titled "Automatic Image and Video Colourisation using Deep Learning" was published in 2018 [6]. Here, the authors have proposed a CNN architecture for image colourization and video data is stored in an LSTM and is fed frame-wise into the CNN. An LSTM can store long term dependencies hence and since every successive frame is effected by the decision of earlier frames, there applicability of LSTM here is crucial. ReLU activation function is used and Inception model pre-trained on ImageNet was used as the backbone feature extraction network. Data augmentation was performed by flipping and rotating images by a small angle. Also, at every layer, batch normalization is applied to pre-process the data before passing ahead to the next layer. Since these distances model perceptual distances, Euclidean loss was tested which resulted in grayish, de-saturated. Hence the problem was treated as one of multinomial classification and a cross entropy loss scheme was used. The model achieves an accuracy of 68%, which is the best so far for colourization, as compared to previously achieved accuracies of 60.2% by [4].

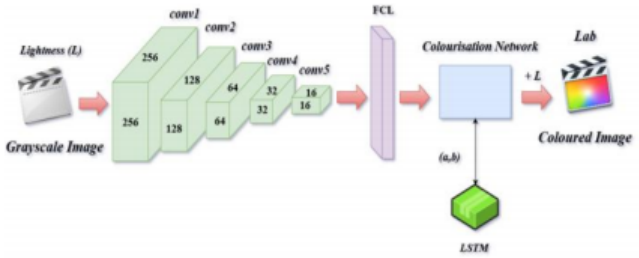

Fig. 2. Network Architecture

V. Implementation Plan

We plan to implement the paper [6]. The following steps will be taken:

- Convert all the frames except the first frame in the input video to grayscale (the first is provided as reference for others)
- Pass image frames through CNN and obtain coloured representation based on loss function
- Implement LSTM to use memory ensure consistency in predictions
- Evaluate performance by comparing with original video colour content

Since this problem involves two similar tasks – image colourization and video colourization, we will also attempt Multi-Task Learning (MTL), with video colourization as our primary task and image colourization as an auxiliary task wherein both the tasks will be learnt jointly. This should lead to an improved learning as both the tasks are related except for video colourization we must have information on how we coloured the previous frames. Hence, we can have one common Linear/LSTM layer for both the tasks. It will help us improve generalization by learning the common features and domain information in the shared layer and thus also help reduce overfitting.

References

[1] Z. Cheng, Q. Yang, and B. Sheng, Deep colorization, in Proceedings of the IEEE International Conference on Computer Vision (2015), pp. 415423.

[2] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In ECCV, 2016

[3] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. ECCV, 2016

[4] Jeff Hwang, You Zhou. Image Colourisation with Deep Convolutional Neural Networks In Stanford University Research Projects, 2016.

[5] Meyer et al. "Deep Video Color Propagation"

[6] Thomas et al."Automatic Image and Video Colourisation using Deep Learning" In 2018 International Conference on Smart City and Emerging Technology (ICSCET)

[7] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang et al., Videoset: A large-scale compressed video quality dataset based on jnd measurement, Journal of Visual Communication and Image Representation, vol. 46, pp. 292302, 2017