2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/ZVHF9195

Towards Temporal Stability in Automatic Video Colourisation

Rory Ward and John Breslin

Data Science Institute,
National University of Ireland, Galway
Galway, Ireland

Abstract

Much research has been carried out into the automatic restoration of archival images. This research ranges from colourisation, to damage restoration, and super-resolution. Conversely, video restoration has remained largely unexplored. Most efforts to date have involved extending a concept from image restoration to video, in a frame-by-frame manner. These methods result in poor temporal consistency between frames. This manifests itself as temporal instability or flicker. The purpose of this work is to improve upon this limitation. This improvement will be achieved by employing a hybrid approach of deep-learning and exemplar based colourisation. Thus, informing current frame colourisation about its neighbouring frame's colourisations and therefore alleviating the inter-frame discrepancy issues. This paper has two main contributions. Firstly, a novel end-to-end automatic video colourisation technique with enhanced flicker reduction capabilities is proposed. Secondly, six automatic exemplar acquisition algorithms are compared. The combination of these algorithms and techniques allow for an 8.5% increase in non-referenced image quality over the previous state of the art.

Keywords: Colourisation, Machine Vision, Deep Learning, Vision for graphics, Video

1 Introduction

Video colourisation is the process of applying colour to monochrome videos [Liu et al., 2021]. It has a broad range of applications, from image and video restoration [Luo et al., 2021], to self-supervised pre-training [Larsson et al., 2016]. Previous to the invention of colour video recorders, black-and-white video recorders were the standard. This has resulted in large amounts of videos which are unavailable in colour. As a result of this, a whole industry has relied on manually colouring these old videos to appeal to a modern audience who expect a more colourful cinematic experience. The process of manually colouring these videos frame-by-frame is very slow and expensive, thus leading to research into automatic video colourisation [Reinhard et al., 2001, Welsh et al., 2002, Hertzmann et al., 2001]. One of the main issues with these types of applications is temporal instability or more commonly referred to as flicker. Reducing this flicker will be the main focus of

Figure 1: Three consecutive frames from Ours (top) compared to DeOldify (bottom). As well as the frames themselves, some of their details are also displayed, these are the frame's saturation and average colour.

 ISBN 978-0-9934207-7-1

this paper. Flicker reduction is the process of removing temporal inconsistency between consecutive frames of a video [Naranjo and Albiol, 2000, Delon, 2006]. This paper has two main contributions. Firstly, a novel end-to-end automatic video colourisation technique with enhanced flicker reduction capabilities. This is achieved by combining the benefits of two standard colourisation methods while minimising their respective limitations. Secondly, six automatic exemplar acquisition algorithms are compared. These algorithms are the connection that allow for the interface between the deep learning and the exemplar based colourisation approaches. The combination of these algorithms and techniques allow for an 8.5% increase in non-referenced image quality over the previous state of the art.

The rest of this journal has the following structure. Section 2 explores the state of the art. Section 3 describes our implementation. Experimental results are shown and discussed in Section 4 and conclusions are drawn in Section 5. This is followed by future work recommendations in Section 6.

2 State of the Art

Three different methods of automatic video colourisation have been developed, these are Scribble [Levin et al., 2004], Exemplar [Zhang et al., 2019], and Deep-Learning [Zhang et al., 2016, Isola et al., 2016, Nazeri and Ng, 2018, Antic, 2019] based approaches, (See Table 1). These methods will be examined in more detail in this section.

Method	User Input Required	Temporal Consistency
Scribble	Yes	No
Exemplar	Yes	Yes
Learning	No	No

Table 1: Comparison of the various automatic video colourisation techniques. The criteria for comparison is whether the technique requires user input and if the technique ensures temporal consistency.

Some of the earliest solutions to this problem involved scribble based methods. These approaches generally worked by the user inputting scribbles into certain areas of an image, and the algorithm then colourised the image based on where it believed these scribbles were applicable to the image. These methods worked reasonably well, but did require user input and their performance depended on how well the user knew which colours were required [Heu et al., 2009].

Exemplar based methods provide the network with the reference image for it to base its colourisation on. This process tends to work well, but is highly dependent on the user's skill in finding a suitable exemplar to match the video being colourised. This generally involves a lot of competent human interaction or a strong database retrieval algorithm with a large selection of images [Liu et al., 2008].

The final method to be examined is that of deep-learning based approaches. These methods work by initially training a model on large datasets of related media. This model is then used to colourise the inputted video. This approach uses deep neural networks to learn complex relationships in media, whether they be temporal or spatial and apply them to new unseen data. These models have achieved state-of-the-art performance in this area. However, one main issue regarding learning based approaches is that of flicker or temporal inconsistency [Lai et al., 2018].

From our analysis of the existing approaches, we have observed that each one either requires user input or does not allow for temporal consistency. Our approach does not require user input and allows for temporal consistency.

3 Implementation

Our Implementation can be segmented into four main stages that follow on sequentially from each other. These stages are video collection, video pre-processing, flicker reduction and finally performance evaluation. Each of these steps will now be explained in greater detail.

3.1 Video Collection

Initially a suitable video to test our methods on needed to be sourced. Bold Emmet was selected, and then obtained through Trinity College Dublin (TCD). This video was selected because it provided a temporarily inconsistent deep-learning based colourisation which flicker reduction could be applied to. This temporarily inconsistent deep-learning based colourisation is typical of colourisation applied to older and low resolution videos. Therefore, this makes Bold Emmet a good representative of this category of videos.

3.2 Video Pre-Processing

Before flicker reduction could be performed, some pre-processing was required. Firstly the video needed to be segmented into scenes. Our flicker reduction technique works on the principle that colours within a scene are relatively consistent to the point where knowing a good representation of these colours in one frame is beneficial to the colourisation of the whole scene [Welsh et al., 2002]. Once the scene had been obtained it was necessary to colourise it using a state-of-the-art colouriser [Antic, 2019].

3.3 Flicker Reduction

Flicker reduction is a two-step process. The first step is the acquisition of a suitable exemplar image. The next step is to use this exemplar to colourise the monochrome clip. Usually obtaining a "good" exemplar image is difficult, as explained in Section 2. One of the novel parts of this paper is that this is no longer the case. One of the major challenges of this project was understanding how to describe the term "good" in relation to an exemplar image. Two metrics were investigated to determine their correlation to "good". These metrics were average colour per frame and saturation per frame.

3.3.1 Average Colour Per Frame

Flicker can be defined as a sudden change in colour between consecutive frames. As such, the average colour per frame of the video was calculated, (See Figure 2). From this the minimum and maximum values could be obtained, and then the corresponding frames compared, (See Figure 3).

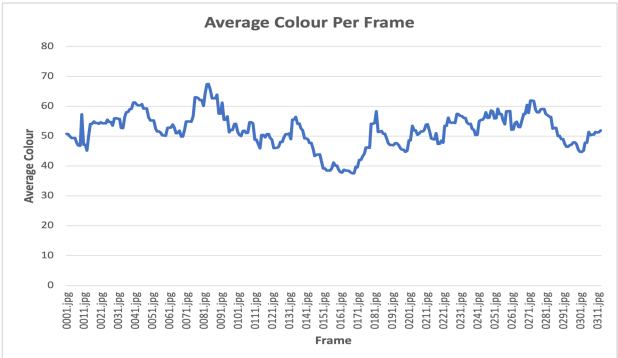

Figure 2: Plotting the average colour per frame of the colourised clip.

Figure 3: Comparing the frames with the minimum (left) and the maximum (right) average colour.

3.3.2 Saturation Per Frame

In addition to average colour per frame, saturation per frame was also calculated, (See Figure 4). Saturation per frame was chosen to be examined because it appeared that some of the colourised frames were being over-saturated during colourisation. The index of the frame with the maximum and minimum values were identified and the corresponding frames compared, (See Figure 5).

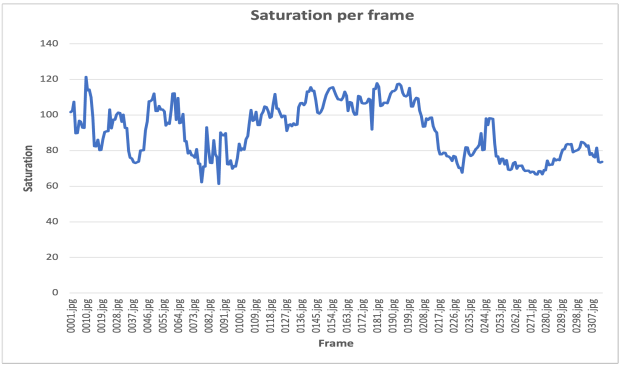

Figure 4: Plotting the saturation per frame of the colourised clip.

Comparing the relative graphs and the relative frames, (See Table 2), it can be seen that saturation has greater discriminatory power in terms of suitability of exemplar than average colour, as it has a larger standard deviation. Through visual inspection it can also be seen that saturation correlates more closely with the trend of the flicker in the video. Moreover, the sharpness of the graph suggests that oversaturation happens regularly and is not just an isolated incidence.

Method	Maximum	Minimum	Average	Std Dev
Average Colour	67.41	37.55	51.89	5.93
Saturation	121.44	61.35	91.15	15.06

Table 2: Comparing the average colour of the clip to the saturation of the clip in terms of their relative maximum value, minimum value, average value and standard deviation.

Following on from these findings, two classes of exemplar selection techniques were employed. The first class of exemplar criteria is based on the saturation of the image. Since flicker is highly correlated with change in saturation, it was used as an exemplar selection mechanism. The minimum (Min_SEA), average (Avg_SEA) and maximum (Max_SEA) saturated frames were used as an exemplar and their resultant colourisation performances were recorded. The second class of exemplar selection techniques is based on non-referenced image quality analysis, specifically the frame with the lowest NIQE (NIQEd) and BRISQUE (BRISQUEd).

Figure 5: Comparing the frame with the maximum saturation (left) to the frame with the minimum saturation (right).

The various saturation criteria was chosen to optimise the clip's saturation and therefore image quality. The blind image quality exemplar selection criteria were chosen to emphasise the best quality frames in the clip. This exemplar then provides the reference to facilitate the colourisation of the rest of the scene. The original monochrome clip was then re-colourised using the process proposed in [Zhang et al., 2019] and the selected exemplars. The results of each of the criteria were analysed and the best clips were combined using a ranking system (Ours).

3.4 Performance Evaluation

Performance can be evaluated qualitatively by visual inspection of the final video. However, performance must also be assessed quantitatively to determine its success in a more objective way. In order to achieve this characterisation of the process, six metrics will be evaluated, four using referenced analysis and two using non-referenced analysis. The four referenced metrics are Power Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) [Horé and Ziou, 2010], Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018] and Fréchet Inception Distance (FID) [Heusel et al., 2017]. The two non-referenced metrics are are NIQE [Mittal et al., 2013], and BRISQUE [Mittal et al., 2012]. In order to compare our methods using the non-referenced techniques, we needed a standard dataset to colourise. The REDS [Nah et al., 2019] dataset was chosen as it would then allow for like-for-like comparisons with other techniques [Wan et al., 2022]. The results of the comparisons will be presented and discussed in the following section.

4 Results and Discussions

Following the methodology described in the implementation section, the evaluation could begin. Temporal consistency needed to be compared for each of the exemplar selection techniques, this would be achieved through the use of frame quality as a proxy metric. This would allow for the comparison of our approach to existing techniques. Furthermore, each of the exemplar acquisition algorithms will also be compared.

The results of our method can primarily be seen qualitatively in the finished colourised video, (See Figure 1). This figure compares two image sequences, the top being an image sequence taken from our clip and the bottom image sequence is the same image sequence but taken from the DeOldify clip. As well as the sequences themselves, some of their associated details are also included.

We will analyse our clip first, it can be seen that the frames are more consistent and natural looking. This is reflected in the associated saturation and average colour details. In contrast, looking at the DeOldify clip, it can be seen that it is less consistent and natural looking. This can particularly be seen in the third frame, where the soldier's hat takes on a very over-saturated yellow colour. This is also reflected in the associated saturation and average colour details, these values have a large spread compared to our clip. These differences can also

 ISBN 978-0-9934207-7-1

be seen in the quantitative analysis, both in the referenced and non-referenced comparisons. The results of this quantitative analysis will now be reported upon.

4.1 Referenced comparisons

Table 3 examines the referenced comparisons. This table was created by initially colourising the validation set of the REDS dataset (as outlined in the "Performance Evaluation" subsection of the "Our Implementation" section) using each of the exemplar acquisition methods. The outputs of this were then analysed using the PSNR, SSIM, LPIPS and FID metrics.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DeOldify	**30.35**	0.89	0.1	86.50
Min_SEA	30.12	0.89	0.1	87.14
Avg_SEA	30.07	0.89	0.1	85.47
Max_SEA	30.11	0.89	0.1	85.24
BRISQUEd	30.06	0.89	0.1	84.15
NIQEd	30.13	0.89	0.1	**77.23**
Ours	30.14	0.89	0.1	79.00

Table 3: Quantitative referenced comparisons on the REDS dataset.

Looking at the table, it is evident that there is no substantial difference in terms of SSIM or LPIPS between the algorithms. Perhaps, these metrics are not able to detect subtle changes in colour. NIQEd achieved the best FID score out of the exemplar selection techniques, followed closely by our technique. PSNR is highest in the DeOldify clip, although just marginally ahead of our model. Summarising the whole table, there appears to be a narrow spectrum of results. The issue with each of these metrics is that they are referenced, they are essentially measuring the distance between the colourised clip and the ground truth. However, this is generally not the main focus in colourisation as it is an ill-posed problem with multiple plausible solutions. Non-referenced metrics solve this problem by considering the naturalness of an image as opposed to its distance from the ground truth.

4.2 Non-referenced comparisons

Table 4 examines the non-referenced comparisons. This table was created by colourising Bold Emmet using each of the exemplar acquisition methods. The outputs of this were then analysed using the BRISQUE and NIQE metrics.

Method	NIQE ↓	BRISQUE ↓
DeOldify	16.23	65.43
Min_SEA	16.12	66.31
Avg_SEA	16.15	66.32
Max_SEA	16.16	66.33
BRISQUEd	16.14	66.33
NIQEd	16.15	66.33
Ours	**15.01**	**60.02**

Table 4: Quantitative non-referenced comparisons on real old films.

Looking at the table, our method outperforms DeOldify by about 8% in terms of the NIQE score and 9% in terms of the BRISQUE score. To summarise, out method achieves an 8.5% performance gain on the previous state of the art in terms of non-referenced image quality analysis.

 ISBN 978-0-9934207-7-1

5 Conclusion

We have seen that by using exemplar based colourisation after deep-learning based colourisation we can reduce flicker in a video. An appropriate exemplar must be chosen from the deep-learning based colourised video. Saturation and blind image quality evaluation were found to be useful indicators of suitability of an image to be the exemplar. We have seen that different criteria are applicable to different scenes. We learned that an optimal combination of deep-learning and exemplar colourisation techniques with automatic exemplar selection can outperform any singular colourisation technique. A hybrid approach is better than any individual exemplar, scribble or deep-learning based approach. We have contributed a novel end-to-end automatic video colourisation technique with enhanced flicker reduction capabilities. Six automatic exemplar acquisition algorithms were also compared. The combination of these algorithms and techniques allowed for an 8.5% increase in non-referenced image quality over the previous state of the art.

6 Future Work

An area of possible future work would be to expand the criteria for exemplar selection and assess their usefulness. Contrast could be investigated and assessed to evaluate its suitability as an exemplar selection criteria. This could result in a further increase in performance. Another interesting area to investigate would be using optical flow. Optical flow works on the idea of calculating motion vectors between successive frames. This could allow for more precise colourisation of particular objects within a video sequence as they travel through time.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No.18/CRT/6223, and also by Grant Nos.16/RC/3918, 12/RC/2289_P2 and 16/RC/3835. We would like to thank the Irish Centre for High End Computing (ICHEC) for their computing resources. We would also like to thank Trinity College Dublin (TCD) for allowing us to use their videos.

References

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

[Delon, 2006] Delon, J. (2006). Movie and video scale-time equalization application to flicker reduction. *IEEE Transactions on Image Processing*, 15(1):241–248.

[Hertzmann et al., 2001] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 327–340, New York, NY, USA. Association for Computing Machinery.

[Heu et al., 2009] Heu, J.-H., Hyun, D.-Y., Kim, C.-S., and Lee, S.-U. (2009). Image and video colorization based on prioritized source propagation. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 465–468.

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

[Horé and Ziou, 2010] Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

[Isola et al., 2016] Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004.

[Lai et al., 2018] Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., and Yang, M.-H. (2018). Learning blind video temporal consistency.

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

[Levin et al., 2004] Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. *ACM Transactions on Graphics*, 23.

[Liu et al., 2008] Liu, X., Wan, L., Qu, Y., Wong, T.-T., Lin, S., Leung, C.-S., and Heng, P.-A. (2008). Intrinsic colorization. *ACM Trans. Graph.*, 27(5).

[Liu et al., 2021] Liu, Y., Zhao, H., Chan, K. C. K., Wang, X., Loy, C. C., Qiao, Y., and Dong, C. (2021). Temporally consistent video colorization with deep feature propagation and self-regularization learning. *CoRR*, abs/2110.04562.

[Luo et al., 2021] Luo, X., Zhang, X. C., Yoo, P., Martin-Brualla, R., Lawrence, J., and Seitz, S. M. (2021). Time-travel rephotography. *ACM Transactions on Graphics*, 40(6):1–12.

[Mittal et al., 2012] Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.

[Mittal et al., 2013] Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.

[Nah et al., 2019] Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., and Lee, K. M. (2019). Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*.

[Naranjo and Albiol, 2000] Naranjo, V. and Albiol, A. (2000). Flicker reduction in old films. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, volume 2, pages 657–659 vol.2.

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

[Reinhard et al., 2001] Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41.

[Wan et al., 2022] Wan, Z., Zhang, B., Chen, D., and Liao, J. (2022). Bringing old films back to life. *CVPR*.

[Welsh et al., 2002] Welsh, T., Ashikhmin, M., and Mueller, K. (2002). Transferring color to greyscale images. *ACM Trans. Graph.*, 21(3):277–280.

[Zhang et al., 2019] Zhang, B., He, M., Liao, J., Sander, P. V., Yuan, L., Bermak, A., and Chen, D. (2019). Deep exemplar-based video colorization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8044–8053.

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

[Zhang et al., 2018] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.