GRAYSCALE MOVIE COLORIZATION USING DEEP LEARNING TECHNIQUES

Mehmet Furkan ŞAHİN, Ali Asaf POLAT, Assoc. Prof. M. Elif KARSLIGİL
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
sahinmehmet980@gmail.com
aliasafpolat@hotmail.com
elif@yildiz.edu.tr

***Özetçe* —Video renklendirmenin amacı, gri bir videoya renk eklemek, renklendirilmiş videoyu algısal olarak anlamlı ve görsel olarak çekici kılmaktır. Bu şekilde, geçmişten bugüne kadar ulaşmış eski görüntüler, farklı şekillerde görselleştirilip insanlara sunulabilir.**

Bu projede, Evrişimsel Sinir Ağları kullanılarak görüntü renklendirme yapan bir modelden faydalanılarak video renklendirilmiştir. Referans tabanlı renklendirme yapılan bu çalışmada, video içerisindeki kareler belirlenen bir referans fotoğrafa göre renklendirilmektedir. Video akışı içerisinde meydana gelen sahne geçişleri tespit edilmiş, her sahnenin içeriğine uygun bir referans fotoğrafını otomatik olarak öneren sistem tasarlanmıştır.

Anahtar Kelimeler—Video renklendirme, derin öğrenme, referans tabanlı video renlendirme, içerik tabanlı görüntü getirme, yapay sinir ağları

***Abstract*—The aim of image colorization is to add colors to a gray image such that the colorized image is perceptually meaningful and visually appealing. In this way, old images that have been reached from past to present can be visualized in different ways and presented to people.**

In this project, video colorization done by using the model that uses Convolutional Neural Networks to colorize images. In this study, which reference-based colorization is used, the frames in the video are colorized according to a reference photo which is determined by system. The scene transitions occurring within the video have been identified and a system has been designed which automatically suggests a suitable reference photo for the content of each scene.

Keywords—Video colorization, deep learning, reference based video colorization, content based image retrieval, neural networks

I. Introduction

The purpose of image colorizing is to add harmonious colors to grayscale image, while preserving its visual and semantic integrity. In this way, it is ensured that the black-and-white images, classical films and scientific works that have artistic value from the past to the present are colored. The difficulty of image colorizing is the semantic separation of the given picture and the transfer of color between these regions. Another difficulty of image coloring is that coloring can be performed in more than one way. For example, an outfit may have more than one color and when evaluating the result in the coloring section, the accuracy of the given result may vary. To solve this problem, color transfer is guided by a selected reference photograph.

In this project, we colorized grayscale videos by using Deep Learning Methods. To do this, we adopted a model [1] which was trained with large data sets. With this model, a system has been created which can colorize grayscale videos. The system is aimed to eliminate the inconsistencies in the colorized video.

II. Related Works

Image colorizing has been applied with different techniques from recent past to present. While these techniques were based on statistical calculations at the beginning, today they are mostly applied with Deep Learning Methods. Deep learning approaches greatly reduces workload in image colorizing.

Reinhard et al. are at the forefront of the work in image coloring[2]. This research is one of the most basic studies of image colorizing. In this study, LAB color space is used. Mean and standard deviation is calculated for each axis of this color space and color characteristics of the image are extracted. Reference photograph is selected to colorize. Color transfer is performed between the matching regions after the color characteristics of the reference and target images extracted. Since the presented solution is based on statictical calculations, unsatisfied results have been obtained for many situations in image colorizing.

The solution presented in another study for image coloring[3] is based on pixel similarities between the reference image and the target image. This similarity is calculated by the luminance and standard deviation of the corresponding pixel relative to its neighbors. The given reference image and target image are divided into subsets. Then, the characteristics of the pixels are obtained in these clusters by using color values and neighborhood information. According to these characteristics, color transfer occurs between the matching regions. One of the difficulties in image transfer is to find a similarity between a three-channel reference photo and a single-channel target photo. In order to solve this problem, the luminance values of the reference photo were used and the color values (chrominance) were transferred to the target photo. Since the presented solution is based on statictical calculations,

unsatisfied results have been obtained for many situations in image colorizing.

In the work of Mingming He et al[1] deep learning methods used to colorize grayscale images. In this project, the network is divided into two main sub-networks. The first one is determined as the similarity subnet and the second one is the colorization subnet. Similarity subnet measures semantic similarity between the reference image and the target image. For this, the VGG-19 model, which was trained with grayscale images, is used. It was aimed at Colorization Subnet to colorize the target image in the best way with the reference image selected.

In this study we adopted a model, which was trained by He et al[1], to colorize grayscale videos. Firstly, grayscale video is divided into frames. Then scene transitions were detected in the video which was divided into frames. To do this the algorithm calculates the difference between sequential frames. After the scene transitions are detected, a proper reference image will be recommended to the content of each scene automatically. Following this step, the colorizing process will be done, which is the last operation. This will be done with the work of Mingming He et al.[1].

III. Method

A. Scene Transition Detection

Detecting the scene transitions is important for meaningful colorizing of the video. The reference photo of a scene may not correspond to the content of another scene. The scene transitions are detected and an appropriate reference photo is recommended for the content of the relevant scene. Thus, the reference image is updated whenever there is a scene transition.

This algorithm computes the difference between sequential frames with the absolute difference method given in Equation 1 and determines the scene transitions. It is not possible for the difference between sequential frames to be more than the determined threshold value unless there is a scene transition. If the result of the calculation is above the threshold value, it means that there may be a scene transition.

First, differences between all sequential two frames are calculated by Equation 1. Then 30 sequential frames are determined and the average of these frames is calculated by Equation 2. The maximum difference value is selected among 30 frames which is calculated according to Equation 1. Check if the difference of the selected frame is more than 3 times the average, which is a threshold value, calculated according to Equation 2. If more, this frame is a possible scene transition (P_i). In this way, all the frames of the video are scanned and possible scene transitions are detected.

Consequently, for each possible scene transition, the average of the previous 10 frames, and subsequent 10 frames, a total of 20 frames are calculated by Equation 3 to confirm possible scene transitions. If the difference of a possible scene calculated according to Equation 1 is more than twice the average of these 20 frames, this frame is a scene transition. An example scene transition is shown in Figure 1.

$$d(x,y) = \sum_{i=0}^{n}\sum_{j=0}^{m}\frac{|x_{ij}-y_{ij}|}{n\times m} \tag{1}$$

$$avg_{30}(x_i) = \sum_{j=i}^{i+29}\frac{d(x_{j+1},x_j)}{30} \tag{2}$$

$$avg_{20}(P_i) = \sum_{j=i-10}^{i+10}\frac{d(P_{j+1},P_j)}{20} \tag{3}$$

Figure 1 Scene Transition

B. Automatic Reference Recommendation

In order to make reference-based image colorization, it is necessary to choose a reference image that is semantically similar to the target image. To find the semantic correspondence of the images, we use feature vectors that were extracted from the Resnet-50 model.

ResNet-50 is a convolutional neural network that is 50 layers deep. This network can classify over 1000 different classes. The feature vector is extracted from the last layer of this model before classification. The feature vector is compared with the feature vectors of the images in a large dataset and the closest image is selected. The similarity of the vectors is found by the Euclidean distance given in Equality 4. In the reference photo recommendation section, the size of the dataset is important for recommending a reference photo which is closer to the target photo.

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \tag{4}$$

1) Automatic Reference Recommendation For Videos: One of the difficulties of video colorizing is the selection of the reference photo. Accepting the video to be colorized as a whole and colorizing with the same reference photo causes problems. A reference photo that is suitable for a scene may not be suitable after the scene has changed. For example, the reference photo used for a scene with fish in a documentary video, will be used to colorize the scene where the lions are shown after the scene changes. In this way, there might be inconsistencies in a video colorizing.

In our method, the reference photo is automatically updated as shown in Figure 2. For this, first of all, scene transitions are detected in the video. After the scene transitions are detected, when the colorizing sequence comes to the first frame of each scene, the reference photo is updated and the same reference photo is used for colorizing throughout the scene. While updating the reference photo, the feature vector is extracted from the

last layer of the Resnet-50 model before classification. This vector is compared with the feature vectors of other photos in the dataset, and the closest photo is updated as the new reference photo according to the euclidean distance given in Equality 4. Thus, the incompatibility between the reference photo and the target photo is resolved during video colorizing.

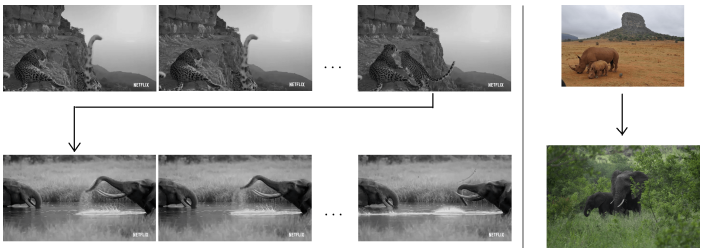

Figure 2 Scene Transition and Updating the Reference Photo

IV. Experimental Results

In our method, first of all, the scenes are detected in the video and at the beginning of the new scene, our system recommends a new reference photo. The first frame of the scene is colored according to the reference photo selected. At the point of selecting the reference photo for the rest of the scene, it is intended to give the previous frame, which was colorized first, as the reference photo to the next frame. Thus, it is predicted that more consistent results will be obtained. However, the results were not as expected and consistent results could not be obtained. This is because the error in a colorized frame increases cumulatively.

Figure 3 Cumulative Error

It is clearly seen that the error is gradually increasing in Figure 3. When the first frame, which is colorized erroneously, given as a reference to next frame, error rates increases cumulatively.

In the proposed method, no cumulative errors occur. This method ensures that the reference photo is updated whenever there is a scene transition. All the frames in the scene are colored according to this updated reference photo. Thus, the cumulative error is eliminated. Results are shown in Figure 5,6,7,8,9,10.

Looking at Figure 4, it is seen that the errors occurred during colorizing of a frame are not transferred to the next frame.

Figure 4 No Cumulative Error

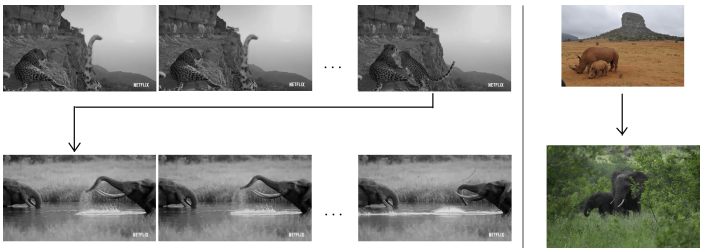

Figure 5 Grayscale Input 1 With References

Figure 6 Colorized Output 1

Figure 7 Grayscale Input 2 With Reference

Figure 8 Colorized Output 2

Figure 9 Grayscale Input 3 With Reference

Figure 10 Colorized Output 3

1) Success Measurement: The success of colorizing according to the reference photo which is proposed by the system will be measured with different metrics. These are PSNR and SSIM metrics. Success measurement will be done on videos in different categories. These categories are videos of flowers, animals, landscapes and people. Colorful videos will be converted to grayscale video and colorized and compared with the original version of the video.

Table 1 PSNR and SSIM Results

Video	PSNR	SSIM
Hayvan	29.59 dB	0.904
İnsan	30.91 dB	0.928
Çiçek	29.13 dB	0.872
Manzara	29.52 dB	0.909

2) Performance Analysis: There are factors that affect the operating time of the system. These are the size of the image to be colorized, the size of the dataset created to suggest an auto reference photo, and the size of the photos in the dataset. The results of the analysis made considering these factors are shown in Table 4.2 and Figure 5.3.

It can be seen that at Figure 11 and 12 when the input photo size decreases, the coloring time decreases. When the number of photos in the data set and the size of the photos increase, the time to find a reference photo increases.

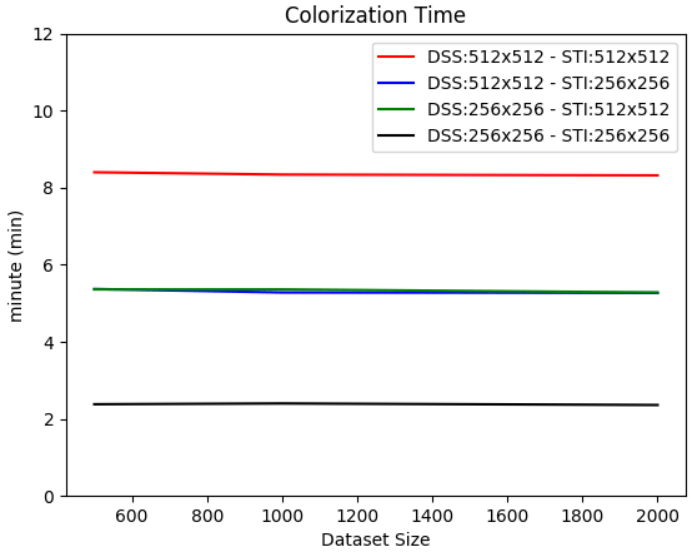

Figure 11 Performance Analysis Graph 1

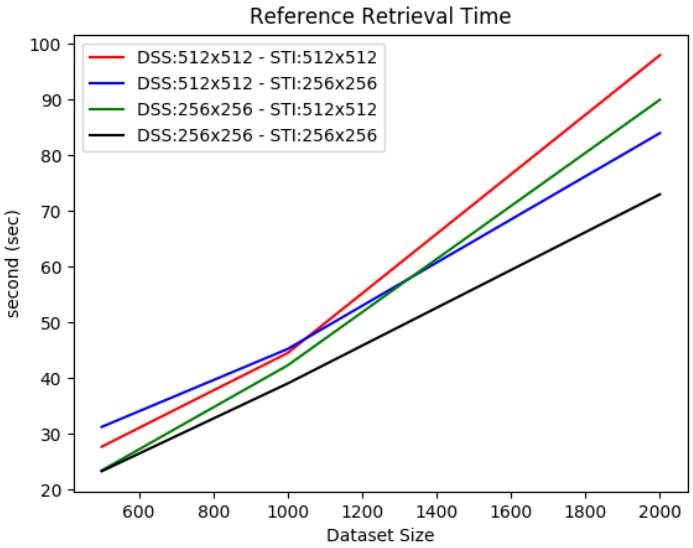

Figure 12 Performance Analysis Graph 2

Table 2 Performance Analysis Table

VKFA	VKFB	GFB	RBS	RS
500	512x512	512x512	≈7.65 sec	≈8.40 min
500	512x512	256x256	≈31.21 sec	≈5.37 min
500	256x256	512x512	≈23.38 sec	≈5.36 min
500	256x256	256x256	≈23.25 sec	≈2.38 min
1000	512x512	512x512	≈44.50 sec	≈8.34 min
1000	512x512	256x256	≈45.22 sec	≈5.28 min
1000	256x256	512x512	≈42.28 sec	≈5.36 min
1000	256x256	256x256	≈39.02 sec	≈2.40 min
2000	512x512	512x512	≈1.38 min	≈8.19 min
2000	512x512	256x256	≈1.24 min	≈5.27 min
2000	256x256	512x512	≈1.30 min	≈5.28 min
2000	256x256	256x256	≈1.13 min	≈2.36 min

DSS : Data Set Size, SPDS : Size of Photos in Data Set, STI : Size of Target Image, RRT : Reference Retrieval Time, CT : Colorization Time,

V. Discussions

Our project suffers from some limitations. First, small changes in successive frames in the video cause sudden changes in the color of the object. This causes inconsistency in the video. Looking at Figure 13, it is seen that there is a sudden change in color on the jacket as a result of coloring. Secondly, the colors of objects can be ambiguous. The object can be colored with a color which is different from the previous scene due to the update of the reference photo when the scene transition occurs. Figure 14 shows the scene transition and recommended reference photos in the flower video. In Figure 15, it is seen that the same flower is painted in a different color after updating the reference photo. Another limitation is that the scene transition cannot be detected. The reason for this case is that the scene changes with a slow transition. Since the difference between successive frames is smaller than the threshold value determined by the scene transition detection

algorithm, the scene transition cannot be detected and the reference photo cannot be updated. Looking at Figure 16, it is seen that the successive frames are not very different from each other and the scene transition cannot be detected. Since the scene transition cannot be detected, the image cannot be colored consistently.

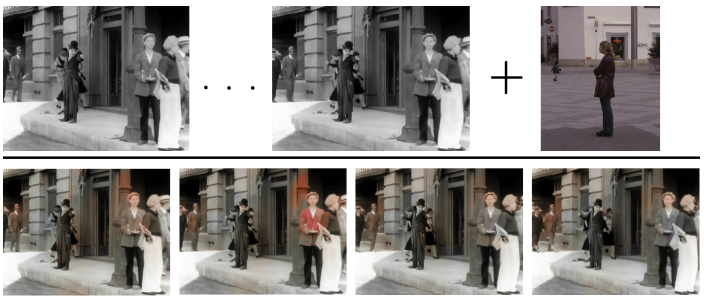

Figure 13 Sudden Color Change

Figure 14 Scene Transition in Flower Video

Figure 15 Colorization Result After Scene Transition

Figure 16 Slow Transition of Scene

VI. Conclusion

In this project, a reference-based video colorization system has been developed. For this, a study, which makes image colorization with deep learning methods[1] is adapted for video colorization. While colorizing the video with the proposed method, the scene transitions in the video were detected. When there is a scene transition, the selected reference photo may not be suitable for the next scene. Therefore, the reference photo needs to be updated. For this purpose, a system has been developed to automatically update the reference photo, which is suitable for the content of the relevant scene. When there is a scene transition, a suitable reference photo is selected for the scene from the our dataset with the content-based image retrieval model.

Even if there are different scenes in the grayscale video that we choose to colorize, each scene is colorized with a suitable reference photo in our video colorization work. In this way, the coloring results are visually satisfying.

Consecutive frames in the same scene will not be very different from each other. With this in mind, in our study, all the frames of the relevant scene are colorized with the same reference by updating the reference photo when there is a scene transition. Thus, a reference photo is selected for each scene instead of choosing a reference photo for each frame of the video.

References

[1] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," 2018.

[2] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[3] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, pp. 277–280, 07 2002.