DOI: 10.21917/ijivp.2022.0406

AN IMPROVISED VIDEO ENHANCEMENT USING MACHINE LEARNING

D.K. Mohanty[1], R. Rajavignesh[2], V. Elanangai[3] and A. Saranya[4]

[1]*Government B.Ed. Training College, Kalinga, India*
[2]*Department of Computer Science and Engineering, K.S.K College of Engineering and Technology, India*
[3]*Department of Electrical and Electronics Engineering, St. Peter Institute of Higher Education and Research, India*
[4]*Department of Computer Science, Vels Institute of Science Technology and Advanced Studies, India*

Abstract

The visual quality of underexposed videos can be increased with the help of a process known as video enhancement. Within the scope of this research, we present a novel approach to enhancing the exposure of movies that are underexposed. Because it has a low barrier to entry theoretically and can reliably give perceptually pleasant outcomes without adding artifacts, it is ideal for a wide variety of practical applications. This is because it is useful for a wide variety of practical applications. We demonstrate the usefulness of the method by displaying improved films of good quality that were made from a variety of different sorts of underexposed videos. A novel approach to the enhancement and editing of video is presented by our method. Rather than relying on a single heuristic transform function, it makes use of human visual perception to adaptively customize the overall visual appearance of the finished product. We believe that our work has the potential to make a big influence in the field of perception-aware video editing and the applications that are related to it, and that it is an important contribution to the community that works on improving videos. The challenge of video enhancement can be formulated as follows: given a video of low quality as the input, produce a video of high quality as the output, but only for specific uses. Whether it be in terms of the video objective clarity or its more subjective qualities, we are interested in learning how we might improve its overall quality.

Keywords:

Visual Appearance, Video Enhancement, Machine Learning

1. INTRODUCTION

The utilization of digital video is currently everywhere. It is general knowledge that the discipline of computer vision has focused a lot of its attention in recent years on the topic of video improvement because it is one of the hottest topics in the industry. It is necessary to improve the aesthetic quality of the video in order to get it ready for automated video processing, which includes analysis, detection, segmentation, and recognition [1]. This can be accomplished by either providing a better transform representation or by enhancing the video quality. In addition to this, it helps in the analysis of contextual data, which is important for making sense of object behavior. This is accomplished without the need to resort to time-consuming and expensive human visual inspection [2]. Surveillance, general identity verification, traffic, criminal justice systems, civilian or military video processing, and many more situations all make use of captured, processed, and utilized digital video in some capacity.

The challenges that are outlined below [3] are just some of the many difficulties that make it challenging to conduct out video enhancing understanding when working with low quality footage.

Because there is not enough contrast, it is difficult to make out moving things against the dark background. The majority of color-based techniques will be ineffective, however, if the colors of the moving objects and the background are too similar to one another.

- A low signal-to-noise ratio is almost often the effect of using a high ISO (ISO is the number indicating camera sensors sensitivity to light). The details in digital photos may become blurry if the ISO setting is increased too far. A lower ISO value indicates that the camera is less sensitive to light.
- The information is carried by a video signal that is a degraded replica of the original signal, which represented a continuous third dimension. This signal is what conveys the information. The process of physically acquiring the material itself, as well as any later rate or format conversions, are both examples of activities that could be responsible for such deterioration.
- There is a connection between the knowledge that people have about their environments and the way in which they understand the happenings around them.
- The weights that are given to the areas of subsequent images that are occupied by moving objects need to gradually change in order to keep the inter-frame coherence intact.
- The low quality of the video equipment that was being used as well as the incompetence of the operator; the relevance of a single pixel from a low quality image in a region where the local variance is minor, such as the space between the headlights and the taillights of a moving vehicle.

The research divides existing video enhancement methods into two categories: self-enhancement and frame-based fusion enhancement. These categories are determined by whether or not the enhanced video has high-quality background information. Improving the quality of the film that was shot originally is the normal procedure for raising the resolution of low-resolution video files. It does not include any credible background information at all. This category includes processes such as enhancing videos with a high dynamic range (HDR), enhancing videos that have been compressed, transforming videos using wavelets, and improving contrast. These techniques are often referred to as self-enhancement of bad video quality, which is a word that describes how they are applied. The video is of such bad quality that it cannot be seen due to the lack of available light. The reason for this is that there are portions of the low-light footage in which there is almost no detail that can be made out at all. There is no way to retrieve any data that has already been lost, regardless of how much the brightness is adjusted. The visual quality of low-resolution films can have their quality increased by utilizing a technique called frame-based fusion. This technique integrates the illumination data of clips that were filmed at various periods. The technique depends on acquiring data about the context that is of a high quality in order to merge videos of lower quality.

2. RELATED WORKS

It is common practice to bring up either the spatial domain or the frequency domain when discussing the processing of a picture. The term spatial-based domain, which refers to the image plane, describes the category of techniques that include the direct manipulation of pixels contained inside an image. The manipulation of the converted spatial frequency spectrum of the image serves as the conceptual foundation for frequency-domain processing approaches. Enhanced procedures that are based on a number of different combinations of methods from these two areas are not uncommon.

In addition, the same enhancement strategy can be implemented in both domains, delivering the same results. There have been many different suggestions made for methods of enhancing videos that use the same image processing. Despite their widespread application, there is not yet a single, overarching standard that can be applied when developing algorithms for video improvement. This is despite the fact that such algorithms are quite common. In a similar vein, there is no single theory that can be applied to improving videos in general. The existing methods of video enhancement serve as the basis for this survey. These methods can be categorized as either belonging to the spatial-based region or the transform-based domain, respectively [6].

When improving a video in the spatial domain, individual pixels in the image are changed directly. Real-time implementations benefit greatly from spatially-based domain methods because these approaches are not only conceptually simple but also have a low level of temporal complexity. However, the requirements for stealth and durability are rarely satisfied by these methods. An overview of spatially-based domain augmentation approaches may be found in the references [7]. Examples of transform-based domain video enhancement approaches include the discrete wavelet transform (DWT), the Fourier transform, and the discrete cosine transform (DCT) [8]. These techniques perform an examination of mathematical functions or signals in terms of frequency, and they act directly on the transform coefficients of the image. By making alterations to the transform coefficients, the purpose of this approach is to elevate the overall picture quality of the movie. The efficiency of the computations that are necessary for transform-based video enhancement, the availability of the image frequency composition, and the simplicity with which the image specific altered domain properties can be applied are only some of the many advantages of this method [9]. The process of improving an image cannot be easily automated, and not all aspects of the image can be improved at the same time to the same degree. The process of improving an image cannot be easily automated.

3. PROPOSED HIGH DYNAMIC VIDEO ENHANCEMENT

High dynamic range imaging, also known as HDRI, is a collection of techniques that allows for a greater range of luminance between an image brightest and darkest parts than is possible with conventional digital imaging or photography. HDRI was developed by Microsoft Research and stands for high dynamic range imaging. HDR photographs, which have a higher dynamic range, are better equipped to capture the whole spectrum of lighting intensities that are present in real-world scenes [10]. These intensities span from the bright sunlight to the dim stars. The dynamic range of real-world environments, on the other hand, is significantly more than what can be recorded by texture maps with 8 bits per channel. The dynamic range of an image can be increased by combining many images of the same scene that were taken with exposures of various, but consistent, lengths. The finished product is a radiance map with floating-point values that accurately represent what may actually be observed in the environment around us. The majority of high dynamic range (HDR) images come from either computer renderings or the combining of multiple still images, a process that is referred to collectively as low dynamic range (LDR) [11].

It is possible to encode the contrast or dynamic range of a video more effectively than it is possible to convey it in text. High dynamic range (HDR) photographs and films are constructed to encode a broad range of luminance values, ranging from cd/m2 (the illumination of a moonless night sky) to cd/m2 (the brightness of a brightly lit room) (Illumination of the sun). HDR video can be obtained in a number of ways, including the use of video cameras to capture HDR images, a fixed mask, an adaptive light modulator, and variable exposures for alternate frames, all of which are analogous to the methods that are used to create HDR still images. HDR video can also be obtained through the use of video cameras to capture HDR images.

If you want to render HDR video on a standard display, you will need to apply tone mapping across frames in order to transform the floating points into the equal value in 8 bits. [12] provides a new definition of apparent distortion based on detecting and categorizing visible changes in the image structure, in accordance with a model of the human visual system. This new definition is based on detecting and classifying visible changes in the picture. The accuracy of the metric has been painstakingly calibrated, and its usefulness has been validated through the utilization of perceptual testing.

To record HDR video, you need gear that is capable of recording HDR video as well as the capacity to blend frames that have varying exposures. Upgrades to the hardware are both expensive and difficult to get. Because of this, it is best to choose a method that combines individual frames with varying degrees of exposure. There are a variety of methods available for acquiring the sequences of frames that are essential for the creation of radiance maps; however, each of these methods results in a loss of either spatial or temporal detail. When preparing HDR film for display, temporal tone mapping is used instead of frame-by-frame tone mapping. This is done because fluctuations in luminance values between frames can result in artifacts. The various approaches to temporal tone mapping, such as those that take into consideration the brightness values of the entire frame or the area around it, each have their own set of advantages and disadvantages. Following the presentation of the broad categories, we will then provide a condensed description of various representative approaches.

3.1 RADIANCE MAPS

To produce an HDR image from a collection of LDR photographs, a method that can integrate and correctly map the relative brightness values of the photos is required. The map of

brightness values that was produced as a result of the exposures that were used is referred to as a radiance map, and it is possible for it to have values that span many orders of magnitude.

The non-linear nature of the capturing devices presents the primary challenge when trying to put into practice the process of creating radiance maps. Because the genuine relative luminance values that are caught and saved by the device may not have a linear correlation, even if one location in the scene has twice the brightness value of another place, the luminance value of the brighter pixel may not necessarily be twice that of the darker pixel. This non-linearity adds an additional layer of difficulty to the process of merging images.

Because of the constraint of sensor reciprocity, the solution to this problem is to directly derive the response function and relative radiance values from a sequence of photographs that were taken at varied exposures. This is done so that the problem may be solved. The algorithm has the capability of combining many photos into a single one. The real-world radiance values that are present in a scene are proportionately represented by the pixel values that are used in an HDR radiance map. The brightness of the scene and the global illumination are both measured, and these values are utilized to make modifications to the lighting for newly added elements.

In order to accurately calculate the amount of light in a scene, it must first be segmented into three distinct areas: the backdrop, the foreground, and the artificial elements. The way in which light interacts with each of the three components is modeled using global illumination; however, the light that is reflected from the background scene is ignored in this simulation. The way in which light interacts with the artificial items in the environment is recreated by employing approximations of the scene geometry and the qualities of the materials.

3.2 HDR-BASED CONTEXT ENHANCEMENT

To keep up with the rapid development of display devices, the technology that improves the quality of video images has made remarkable gains in recent years. When it comes to accurately replicating genuine video images taken in the world around us, the intensity representation of display and acquisition equipment, however, has several limitations that make it less than ideal. Many context-based contrast enhancement approaches have been developed and put into use to circumvent the limitations imposed by the intensity representation of display and acquisition devices. However, there are still certain limitations, such as changes to the color information, the loss of data regarding the lighting, and excessive amplification.

The context enhancement strategy that HDR imagery provides is the solution to these problems. A pre-processing step is carried out on an input image using the auto exposure technique that has been provided. This step purpose is to determine whether context augmentation is required. When applied to an input image, the intensity mapping function results in the creation of several new images. It is not essential to register each individual photograph that is utilized in the creation of an HDR image because this step is not part of the process. Using the method, one can expand the dynamic range of an image, which ultimately results in greater contrast.

It is proposed that a novel method based on histograms be used in order to estimate both the camera response function and the radiance mapping. This technique makes advantage of histogram correspondences to register neighboring frames and to calculate pixel radiances in order to get rid of ghosts. In order to display both low dynamic range (LDR) and high dynamic range (HDR) images, the HDR display device in question must fulfill specific requirements. The issue is investigated from the point of view of a person who employs an HDR display device as opposed to the viewpoint of a person who is responsible for the production of such devices.

3.3 HDR-BASED ILLUMINATION ENHANCEMENT

When it comes to applications that provide aid with driving, HDR video cameras shine due to the vast range of lighting situations that they can manage. These settings include everything from bright sunlight to pitch-black tunnels. However, given that most viewing equipment is only capable of handling a specific range of values, the dynamic range will need to be lowered. In recent years, numerous methods for reducing the dynamic range of still photographs have been put forward as potential solutions. Because of the peculiar characteristics of video data, many of the techniques that can be applied to video do not make it simple to reduce the dynamic range of still photographs.

The goal is to provide a higher visual quality by reducing the dynamic range of the video sequences, so improving their appearance, while simultaneously reducing the number of temporal artifacts. The algorithm in its most effective version, prepared for use in the hardware implementation that is performed by an FPGA. It is reasonable to be concerned about the possibility of developing eye strain as a result of extended exposure to HDR displays.

An additional component that must be taken into consideration is the impact that the surrounding lighting has on the degrees of brightness and contrast that viewers find most appealing. Two different user studies were carried out in order to investigate these problems. In each of the experiments, the participants viewed HDR video on a monitor while the room lighting was altered in various ways. They were also given the opportunity to adjust the display brightness and black level according to their preferences.

3.4 HDR-BASED TEMPORAL PROPERTIES ENHANCEMENT

The temporal features that glare possesses are a potent weapon that can be used to make dazzling light sources appear more realistic and appealing. Utilize a graphics processing unit, more commonly referred to as a GPU, to enable real-time modeling of dynamic glare. This modeling is based on the anatomy of the human eye. This enables a better depiction of HDR images on LDR media, which can be utilized for interactive applications such as video games and feature films, or it can simply be accomplished by adding motion to initially still HDR photographs. We know that this technique improves the sensation of brightness, and that dynamic glare-renderings are often perceived as more appealing than static ones because the results of psychophysical testing showed that they did. In addition, for

videos that are poorly exposed or have a lot of noise, a novel local image statistic is developed to identify impulse noise pixels.

This statistic is then incorporated into the classical bilateral filter to form ASTC, with the intention of decreasing the mixture of the two most common types of noise, Gaussian and impulse noises in both spatial and temporal directions. This is done with the intention of decreasing the mixture of the two most common types of noise, impulse noises and Gaussian noises. After the removal of any undesirable noise, the techniques then make use of the statistical information obtained from the frames' segmentations in order to improve the video contrast by the application of APMF.

3.5 VIDEO ENHANCEMENT USING ML

Finding and combining the pertinent data from a video sequence that was captured with a stationary camera under different illumination conditions is the objective of context-based video enhancement [3]. The enhancement algorithm for videos works by automatically stitching together individual still images of a scene that were captured at different periods. The original video, which was of lower quality, is combined with the context of a background photograph of higher resolution that was taken from the same angle.

This keeps all the important features intact. Because it offers a more exhaustive description of the setting, the merged image that is produced as a result is superior for the visual perception of the scene by both humans and machines. The Fig.4 provides a visual representation of the overarching algorithmic structure of context-based video enhancement.

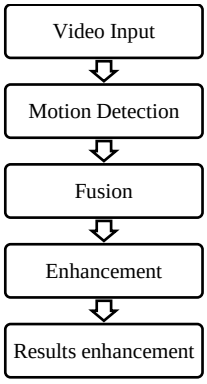

Fig.1. Visual Representation

A process known as image fusion is used to combine numerous photographs of the same location into a single picture that is both more detailed and more useable. The purpose of picture fusion is to produce a composite image that provides a better description of the scene than any of the original photographs by merging the information that is complimentary to and redundant with the information that is contained in each of the original images.

The [6] offer an excellent illustration of a survey of modern fusion techniques. There are algorithms for fusing on the low-, medium-, and high-levels. These distinctions are described in a variety of written works using the words pixel, feature, and symbolic levels respectively. The majority of techniques that are used at the pixel level are able to function in either the spatial domain or the transform domain.

A common procedure for feature-based algorithms is to first segment images into sections and then combine those sections based on the traits that they have in common with one another. High-level fusion algorithms require that the descriptions of the images be consistent with one another so that they can compare and combine the images. The act of fusion itself can take place on any one of a variety of different levels of information representation simultaneously. A recent study looked at both the theory and practice of fusing pictures.

Popular fusion methods for video enhancement include the gradient pyramid, shift invariant discrete wavelet transform [8,9], weighted combination, optimization approach, and biologically based methods such as neural networks and contourlet transform, bio-inspired weight average image. These methods all work at the pixel level to improve the quality of the video.

We find a common fusion equation of video enhancement, which can be obtained by computing the weight average image at each pixel, and we use it to assess existing strategies for context-based video enhancement. This equation can be generated by computing the weight average image at each pixel.

$$F(x, y) = N(x, y) * w(x, y) + D(x, y)(1-w(x, y)) \quad (1)$$

where

$F(x, y)$ - fusion image,

$N(x, y)$ - low quality video,

$D(x, y)$ - high quality background.

$w(x, y)$ - weight, value is set in the range[0, 1].

The procedure that determines w(x, y) importance weights is dependent on the situation in which it is used.

The software enhances film obtained at night by contrasting it with footage captured during the day and applying the lighting ratio that is generated as a result. The following formula can be used to determine how much brighter the improved nighttime footage is: L_{eng} = LDB(x, y) In this particular instance, the formula is:

$$L_{eng} = \text{LDB}(x, y)\ \text{LNB}(x, y)\ \text{LN}(x, y)$$

in which LDB(x, y) and LNB(x, y) stand in for the lighting components of the daytime and nighttime backdrop images, respectively. The illumination of the input nighttime video is represented by the LN(x, y). The static light has been removed from the upgraded version, which can be found here.

Because the illumination ratios of the daytime background images and the nighttime background images can sometimes be significantly lower than 1, the static lighting of the night-time video tends to transition back to the illumination of the day-time video. This is because in some areas, the daytime background images can be significantly brighter than the nighttime background images.

Uses the information from the fusion image more efficiently than it would be used in a single midnight image in order to improve the contrast and signal-to-noise ratio of the combined image. The merged picture has more information packed into it, making it suitable for researching more complicated behaviors at a deeper level. Image segmentation and the removal of objects from images are powerful tools for enhancement, but only if the segmentation is done correctly. However, if there are any errors in the way that they did things, the image that they produce can have an unusual combination of components.

4. PERFORMANCE ANALYSIS

Our C++ code that fixes underexposure issues in films runs on a desktop computer with an Intel Core 2 Duo processor running at 2.4GHz. In this section, we begin by validating our methodology on several videos with insufficient exposure and then compare it to previously developed methods [4]-[5] in terms of both the way the output looks and how long it takes to process. After this, we carry out user research to demonstrate how our strategy truly functions in practice. The inadequacies of our methodology are the subject of our final and most important discussion.

Table.1. Comparative analysis of 960 × 544 × 78 videos

Video Frames	Accuracy	Precision	Recall
10	91.25	92.88	93.12
20	92.62	93.81	94.62
30	92.99	94.25	95.81
40	93.65	95.62	96.66
50	94.58	96.61	97.84

There are a total of three stages, with stages two and three being the ones that require the maximum amount of time. The second stage is the perception-driven progressive fusion, and the third stage is the spatio-temporal filtering that maintains the texture of the image. Despite the fact that our present C++ implementation is not optimized, our method is significantly faster as in Table.1. In the majority of situations, our perception-driven progressive fusion can handle a video streams in approximately 15 minutes. The most time-consuming component of our technique is the spatio-temporal filtering that we use to preserve the texture. The smoothing process for a video of the aforementioned size normally takes about 30 minutes. Nevertheless, we are of the opinion that a parallel GPU implementation or the application of efficient sampling and hashing methods can significantly boost the runtime performance of our methodology.

5. CONCLUSION

In this article, we will discuss an approach to the improvement of context-free videos that is motivated by human perception. The primary objective of this work is to piece together visually interesting video segments that have been created through the application of a variety of tone mapping curves. The majority of what we do can be broken down into these three stages. To begin, a multi-exposure image sequence is constructed by remapping each video frame with a set of tone mapping curves in order to include several exposures of each frame. This is done so that the series can contain multiple exposures of each frame. We employ several metrics of visual perception quality to help us adaptively discover the locally best exposed sections within the multi-exposure image sequences. After that, we apply the suggested progressive fusion technique to integrate all of these regions into a single movie that has the appropriate exposure. Last but not least, we apply a texture-preserving spatio-temporal filtering technique to the properly exposed movie in order to cut down on the noise interference even further.

REFERENCES

[1] M. Bhende and V. Saravanan, Deep Learning-Based Real-Time Discriminate Correlation Analysis for Breast Cancer Detection, *BioMed Research International*, Vol. 2022, pp. 1-9, 2022.

[2] K.C. Chan and C.C. Loy, BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5972-5981, 2022.

[3] C. Nithya and V. Saravanan, A Study of Machine Learning Techniques in Data Mining, *International Journal of Scientific Research*, Vol. 1, 31-38, 2018.

[4] D. Wu and R. Wang, Edge-Cloud Collaboration Enabled Video Service Enhancement: A Hybrid Human-Artificial Intelligence Scheme, *IEEE Transactions on Multimedia*, Vol. 23, pp. 2208-2221, 2021.

[5] S. Tulyakov, J. Erbach and D. Scaramuzza, Time Lens: Event-Based Video Frame Interpolation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16155-16164, 2021.

[6] K. Asha and M. Rizvana, Human Vision System Region of Interest Based Video Coding, *Compusoft*, Vol. 2, No. 5, pp. 127-138, 2013.

[7] I. Dave, M.N. Rizve and M. Shah, TCLR: Temporal Contrastive Learning for Video Representation, *Computer Vision and Image Understanding*, Vol. 219, pp. 103406-103418, 2022.

[8] P. Karthika and P. Vidhya Saraswathi, IoT using Machine Learning Security Enhancement in Video Steganography Allocation for Raspberry Pi, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 6, pp. 5835-5844, 2021.

[9] S. Niklaus, Revisiting Adaptive Convolutions for Video Frame Interpolation, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1099-1109, 2021.

[10] S. Bertoni and A. Facoetti, Action Video Games Enhance Attentional Control and Phonological Decoding in Children with Developmental Dyslexia, *Brain Sciences*, Vol. 11, No. 2, pp. 171-181, 2021.

[11] S. Yang, Y. Shan and W. Liu, Crossover Learning for Fast Online Video Instance Segmentation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8043-8052, 2011.

[12] S. Selvi and V. Saravanan,. Mapping and Classification of Soil Properties from Text Dataset using Recurrent Convolutional Neural Network, *ICTACT Journal on Soft Computing*, Vol. 11, No. 4, pp. 2438-2443, 2021.