Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/31/20230137

Designing a dual-camera highway monitoring system based on high spatiotemporal resolution using neural networks

Zhiyi Wang

Glasgow College, University of Electronic Science and Technology of China, Chengdu, 611731, China

2020190503034@std.uestc.edu.cn

Abstract. The criticality of infrastructure to societal development has seen highways evolve into an essential component of this ecosystem. Within this, the camera system has assumed significant importance due to the necessity for monitoring, evidence collection, and danger detection. However, the current standard of using high frame rate and high-resolution (HSR-HFR) cameras presents substantial costs associated with installation and data storage. This project, therefore, proposes a solution in the form of a High Spatiotemporal Resolution process applied to dual-camera videos. After evaluating state-of-the-art methodologies, this project develops a dual-camera system designed to merge frames from a high-resolution, low frame rate (HSR-LFR) camera with a high frame rate, low-resolution (LSR-HFR) camera. The result is a high-resolution, high frame rate video that effectively optimizes costs. The system pre-processes data using frame extraction and a histogram equalization method, followed by video processing with a neural network. Further refinement of the footage is performed via color adjustment and sharpening prior to a specific application, which in this case is license plate recognition. The system employs YOLOv5 in conjunction with LPRNet for license plate recognition. The resulting outputs demonstrate significant improvement in both clarity and accuracy, providing a more cost-effective solution for highway monitoring systems.

Keywords: high spatiotemporal resolution, dual-camera videos, histogram equalization, neural network, sharpen, YOLOv5, LPRNet.

1. Introduction

In the twenty-first century, the advent of the smart highway has become increasingly prominent within modern infrastructure. This evolution aligns with the developmental trajectory of emerging nations, the principles outlined in the Sustainable Development Strategies, and the pervasive wave of economic globalization. A smart highway is designed to collate traffic information, document traffic conditions, and furnish intelligent services. Key aspects of this setup, such as danger detection, fatigue detection, and illegal driving recording, rely heavily on camera systems [1].

The performance of highway cameras is expected to offer high resolution for clear imagery and high frame rates to capture high-speed objects, which in turn escalates the overall cost. A solitary highway camera can cost between $5,000 to $20,000, and installation expenses can range from $10,000 to $50,000. This does not include ancillary costs like maintenance and data storage [2]. To fully equip a standard highway, cameras are needed approximately every 1 to 1.5 kilometers, which can impose a

DOI: 10.54254/2755-2721/31/20230137

hefty financial load on the constructors and operators. Market analysis of highway camera systems has revealed that the total cost of a high spatial resolution, low frame rate camera, and a low spatial resolution, high frame rate camera is approximately half the cost of a high spatial resolution, high frame rate camera. Furthermore, the video file size of a 4K 120Hz video is over three times that of a 4K 30Hz video and a 720p 120Hz video. Therefore, to reduce the overall cost of the camera system, the project proposes to replace the high spatial resolution, high frame rate camera with one high spatial resolution, low frame rate camera, and one low spatial resolution, high frame rate camera. The goal is to generate high spatial resolution, high frame rate video from these two cost-effective cameras.

Deep learning processed optical flow, first proposed at the International Conference on Computer Vision in 2015 (FlowNet), and later improved at the Conference on Computer Vision and Pattern Recognition in 2017 (FlowNet2.0), tracks the pixel movement between two adjacent video frames. This concept forms the foundation of dual camera high spatiotemporal resolution video acquisition. Subsequently, in 2018, CrossNet combined high spatial resolution images with low spatial resolution images. In 2021, Nanjing University introduced AWNet to apply super-resolution to videos, employing FusionNet—an auto-encoder based network—to optimize the merging weight of the upscaled low spatial resolution frame and the high spatial resolution frame [3].

The most recent addition to this series of advancements is HSTR-Net, which manages to increase the Structural Similarity Index (SSIM) to 0.973 on the traffic dataset. HSTR-Net employs ContextNet, a convolution network, to extract features and flows from various resolution frames by gradually convolutional down-sampling, enhancing the accuracy of light flow. Notably, the SSIM and the Peak Signal-to-Noise Ratio (PSNR) are closely matched, with a difference within 5%. Consequently, the choice of network should depend more on the complexity and processing time. The aim of this paper is to construct an effective and efficient system [4]. The dataset is obtained by a parallel dual camera system, and the High Spatiotemporal Resolution network is based on AWNet. Algorithms for histogram equalization and color mapping are used for color balance and transformation as a preprocessing step before the High Spatiotemporal Resolution process. The resulting high spatial resolution, high frame rate video can be utilized for evidence storage, and feature and behavior recognition. The paper also demonstrates an application of this system for license plate recognition, using a network that combines YOLOv5 with LPRNet to recognize the position and characters of the license plate. The accuracy of the recognition serves as an evaluation metric for this system.

2. Theoretical framework

2.1. Overall design

The overall design of the system could be divided into three parts: pre-processing, processing, and post-processing as shown in Figure 1.

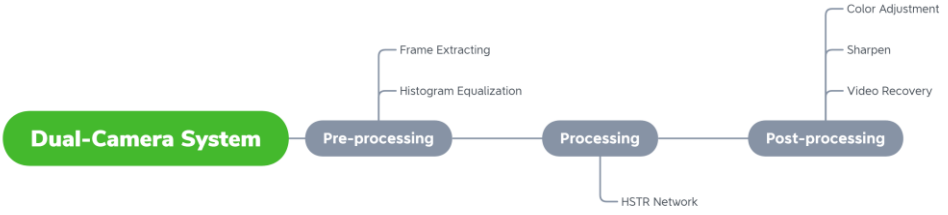

Figure 1. The overall structure of the system (Photo/Picture credit: Original).

Firstly, the preprocessing section includes the frame extraction where the frame is extract from the HSR-LFR and LSR-HFR video as images. The extracted frames are up-sampled by interpolation to match the size of the LSR frame with the HSR frame. The images are then divided to two datasets: reference dataset contains all HSR frames and original frames contains all LSR frames [5]. The color of the two datasets is aligned by histogram equalization and color mapping algorism.

DOI: 10.54254/2755-2721/31/20230137

Then, in the processing section, the frames of the two datasets are grouping by time and combined and by HSTR network. After processing, HSR images with the same resolution of HSR-LFR frames and the same frame number of LFR-HFR frames are generated.

In the post-processing section, the frames could be further sharpening by filters. The frames are combined to generate an HSR-HFR video for further application. In this case, the videos are used for license recognition based on YOLOv5.

2.2. Data collection

The data is collected by a dual-camera system which is set to be parallel on a stand. The performance index of the HSR-LFR camera and LSR-HFR camera are 4K 30Hz and 720p 120 Hz respectively. The focal length, angle of view, and aperture are set to be the same to ensure the graphs in the two cameras are the same [6]. The stand is set on a bridge with a height of 7m which is also the common height of highway camera. The camera system is set up with a downward angle of 15 degree so that the farthest of the frame could reach up to 50m. The system is set as shown in Figure 2.

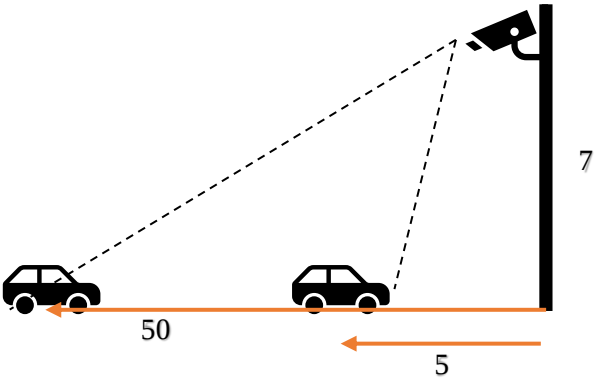

Figure 2. The camera system setting (Photo/Picture credit: Original).

2.3. Preprocessing

2.3.1. Frame extraction and grouping. The High Spatiotemporal Resolution network necessitates inputs of images that are captured simultaneously by the dual-camera system. Therefore, it is crucial to segment the video into individual frames that need to be synchronized by time. Given that Low Spatial Resolution, High Frame Rate videos contain more frames than the High Spatial Resolution, Low Frame Rate videos, the latter should act as a reference frame. All other frames in the original dataset that are closely aligned in time should be referred to this reference frame.

2.3.2. Histogram equalization color adjustment. In a multi-camera system, due to differences in camera resolution, focal length, exposure time, aperture and complicated environments, it is necessary to calibrate the images captured by two cameras. The target is to combine the images of two different cameras. The color histogram is unified to achieve calibration that affects color and brightness. Histogram equalization is a simple and effective image enhancement technology, which changes the gray level of each pixel in the image by changing the histogram of the image, and is mainly used to enhance the contrast of images with a small dynamic range [6]. The original image may be concentrated in a narrow range due to its gray distribution, resulting in an image that is not clear enough. For example, the gray level of an overexposed image is concentrated in the high brightness range, while underexposure will make the gray level of the image concentrated in the low brightness range. Using histogram equalization, the histogram of the original image can be transformed into a uniform distribution (balanced) form, which increases the dynamic range of the gray value difference between pixels, thereby achieving the effect of enhancing the overall contrast of the image. In other words, the basic principle of histogram equalization is to widen the grayscale value with a large number of pixels

DOI: 10.54254/2755-2721/31/20230137

in the image (that is, the grayscale value that plays a major role in the picture), and to widen the grayscale value with a small number of pixels (that is, the grayscale value that plays a major role in the picture). Merge the gray values that do not play a major role in the picture), thereby increasing the contrast, making the image clear, and achieving the purpose of enhancement. The hybrid histogram matching method uses the image histogram of the standard image (Reference) as the standard, and adjusts the color of the target image (Target) to the level of the image histogram of the standard image. That is, by drawing histograms for different colors (RGB), the new distribution of the color in the target image to the standard image is obtained, and the target image is converted into a new image with a similar color distribution to the standard image. Compared with the standard image, the color difference and brightness difference of this kind of image are significantly reduced, and it is greatly optimized for the process of finding optical flow through correlation in HSTR network.

2.4. HSTR network

Super Resolution (SR) is the process of restoring a high resolution (HR) image from a given low resolution (LR) image. SR refers to reconstructing corresponding high-resolution images from observed low-resolution images through software or hardware methods. In this project, high-resolution data can provide more high-definition and high-frame surveillance videos for highway intersections High rate of evidence records, but also improve the correct rate of license plate recognition [7].

At present, there are many super-resolution methods and the state-of-the-art super-resolution methods combines optical flow. Optical flow is to find the correspondence between the previous frame and the current frame by observing the instantaneous speed of pixel motion on the imaging plane, and using the changes of pixels in the image sequence in the time domain and the correlation between adjacent frames, so as to calculate a method of motion information of objects between adjacent frames [4]. After the track of each pixels are obtained, the pixels in the intermediate generated frame could be generated by the interpolation on the track. This method is efficient for images of rigid body moving object and fixed background.

In the practice, or a car moving on a highway with a fixed background, the moving direction of the rigid body car can be obtained through optical flow, so that a high-definition license plate picture can be obtained through the recorded video. As a result, the fixed background should keep the detail from HSR-LFR video and the rigid moving body should keep the sharpness of LSR-HFR video.

The general work flow for the HSTR network is as followed: Up-sampling the LSR image, Obtain the optical flow between the LSF and HSF image, Refer to LSF image to recover the HSF image by optical track, Optimize weights using FusionNet with dynamic filters,Combine the LSF and HSF image to output.

2.5. Post-processing

2.5.1. Image sharpening. Image sharpening is used to be general process before the license recognition. It could sharpen the edge of the figure which could be significantly useful in license recognition by sharpen the character on the license. There are various sharpening method including differential method, Robert gradient operator method, Laplacian method and Gaussian filtering method.

2.5.2. HSI color mapping transformation. HSI(Hue-Saturation-Intensity) is a color frame of reference that could transform a color to another without changing the saturation and intensity. In many cases, the industrial camera used for highway has low color accuracy especially in complicated light conditions. In addition, the model for the license recognition could only suitable for license of specific color so that even small color deviation can lead to low accuracy of recognition. As a result, the videos need color transformation before the process of HSTR.

2.5.3. Video recovery. The result from the network above is separated images so that the images are needed to be combined by sequence to be recovered to a HRF-HSF video. In this project, the output

DOI: 10.54254/2755-2721/31/20230137

video generated from the 4K 30Hz video and 720p 120Hz is 4K 120Hz. Then the HRF-HSF video could be used for clear archive and other uses.

2.6. License recognition

YOLOv5 is one of the most commonly used networks for object detection. By chosen the proper pretrained model, it could detect the position of the license.

After the position of the license is detected, the license could be recognized by the LPRNet. LPRNet is a classic and lightweight convolutional network for character recognition which could provide real-time result from images. It could provide a reliable result and be able to be applied on the embedded system on the highway.

3. Detailed analysis of techniques

HSTR network based on optical flow is a state-of-art field but with various choices as shown in Table 1.

Table 1. The HSTR networks based on optical flow.

Method	Date	Author
CrossNet [3]	2018	Haitian Zheng et al
ToFlow-interp [8]	2019	Tianfan Xue et al
SRNTT [9]	2019	Zhifei Zhang et al
RIFE [10]	2021	Zhewei Huang et al
RIFE(2T2R) [10]	2021	Zhewei Huang et al
AWNet [4]	2021	Ming Cheng et al
MASA-SR [11]	2021	Liying Lu et al
HSTR-Net [5]	2022	H. Umut Suluhan et al

To evaluate the performance of the networks, we use the Vimeo90K dataset which is a famous dataset for superposition. The evaluating matrix obtained from the test set is as Figure 3 below.

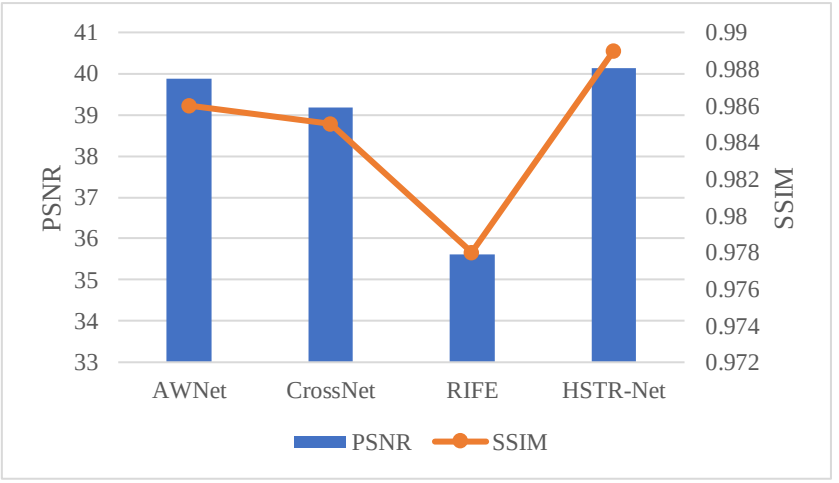

Figure 3. The evaluation among several networks (Photo/Picture credit: Original).

It is shown that the HSTR-Net obtains better result on this dataset and followed by AWNet. In addition, the performance of the networks are relatively close since the performance difference between AWNet and HSTR-Net is less than 1%. However, since the HSTR-Net has more complex structure (more layers and extra convolutional networks), the AWNet is used in this project to reduce cost and for more lightweight design.

DOI: 10.54254/2755-2721/31/20230137

4. Proposed system and implementation

Evaluating all the method under the required condition and circumstance, AWNet is used for HSTR while the YOLOv5 and LPRNet is used for license detection. As for the color adjustment, HSI mapping and histogram equalization algorism are used. The video could be further sharpened by Laplacian method depending on the usage. In short, the flow chart of the whole system is shown in the Figure 4 below.

The implementation of the sections is shown as followed.

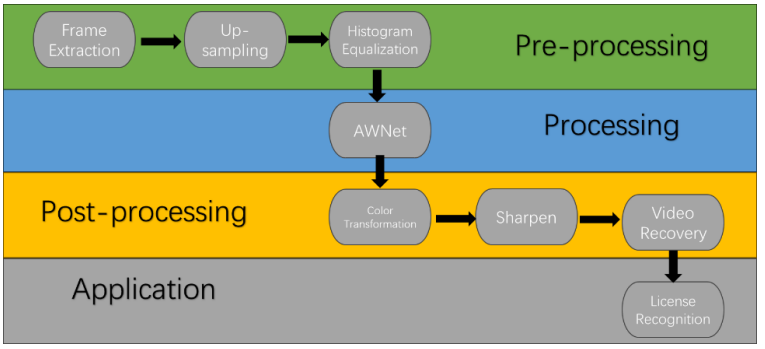

Figure 4. The work flow of the system (Photo/Picture credit: Original).

4.1. Pre-processing

Firstly, the frame images is extracted from the video and saved as PNG files as shown in Figure 5.

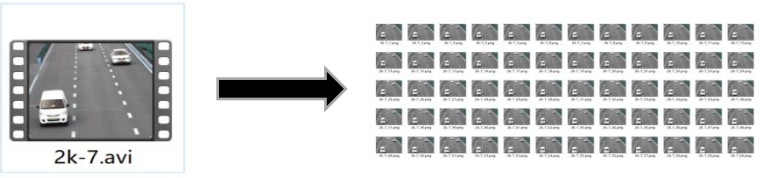

Figure 5. The frame extraction of a video (Photo/Picture credit: Original).

Secondly, it is found that the color of the frames of the two cameras are different as shown in Figure 6 so the brighter one (HSR-LFR) is set as the reference and the darker one is adjusted by histogram equalization method to match the color. As a result, the color is adjusted to be the same as shown in Figure 7.

LSR-HFR (b) HSR-LFR

Figure 6. The color difference between the frames from two cameras (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/31/20230137

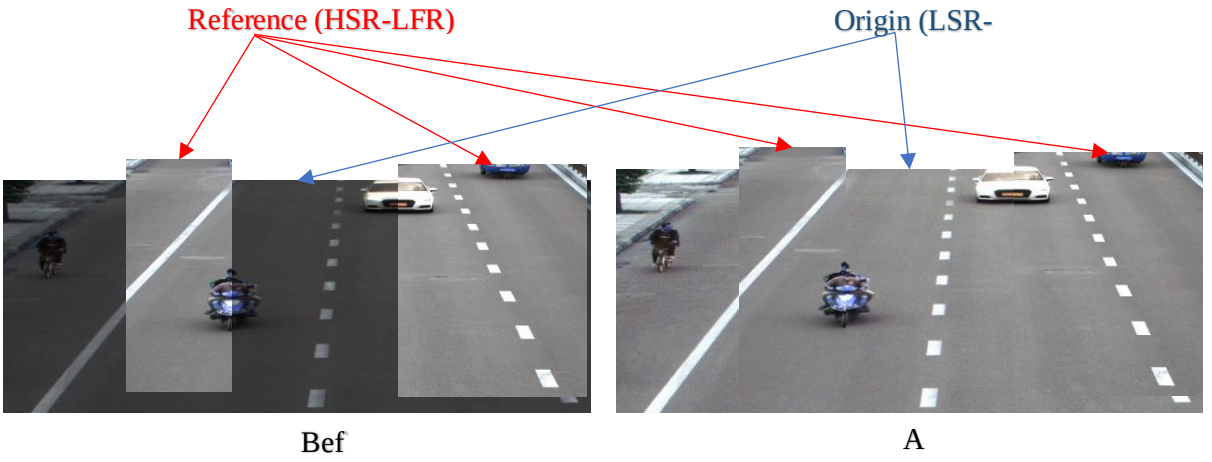

Figure 7. The comparison of color before and after histogram equalization process (Photo/Picture credit: Original).

It is shown that the color difference of the two images are eliminated after the histogram equalization process. Since the FlowNet will calculated the correlation between pixels, the matched color between frames are important. As a result, it is believed that this process will greatly benefit the following process on optical flow.

4.2. Processing

The pre-processed images are then sent to the AWNet pair by pair, and the output result is shown in Figure 8.

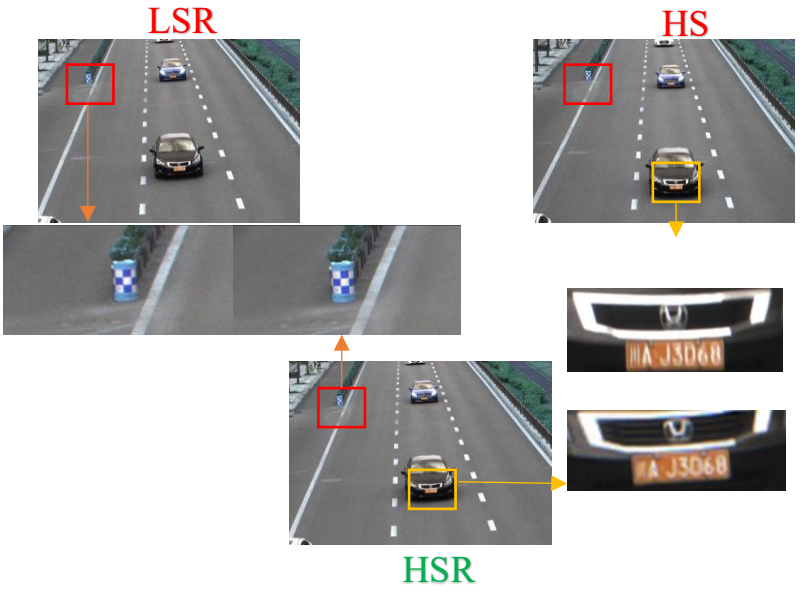

Figure 8. The detailed comparison between unprocessed images and result (Photo/Picture credit: Original).

It is shown that the LSR-HFR images has blurry stable objects since the resolution of the camera is limited and the HSR-LFR video contains blurry figure on moving objects such as cars. These problems are adverse to evidence and certificate storage and license plate recognition.

DOI: 10.54254/2755-2721/31/20230137

After processed by the AWNet, both the details from the fixed background and moving object are saved in the HSR-HFR images. For example, the figure of the buffer in the background are sharpen than the original LSR-HFR image. Moreover, the driver and license plate is clearer and more recognizable the HSR-LFR video. As a result, the process video could increase the accuracy of the license recognition in the following section.

4.3. Post-processing

Since the final target of the project is to increase the accuracy of license recognition, the HSI mapping transformation and Laplacian sharpen are posed on the video.

It is shown that because of the camera color difference, the color of the license (blue background with white characters) is presenting wrong colors (yellow background with white characters) that will disturb the recognition since the license dataset in China is based on true colors. To address this problem, the color is mapped to HSI reference and the yellow is transformed to blue without changing other parts and brightness. At last, the Laplacian sharpen method is also applied and the result is as shown in Figure 9.

(a) Before (b) After

Figure 9. The license before and after histogram equalization and sharpen process (Photo/Picture credit: Original).

It is shown that the result processed is clearer and sharper than the original images. The color of the license is practical for the LPRNet to recognize so that this process will definitely increase the accuracy of recognition.

4.4. License recognition

Input the images into the YOLOv5 and LPRNet for license recognition, the recognition result is shown as Figure 10.

（a）LSR-HFR (b)HSR-HFR (c)Advanced HSR-HFR

Figure 10. The comparison of results of the system (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/31/20230137

It is shown that the advanced HSR-HFR image (after post-processing) is the only image whose license could be recognized. Both LSR-HFR and unprocessed HSR-HFR video failed to be recognized. Even though the accuracy is not satisfying, it still shows significant improvements. The recognition result is better when the vehicle moves closer to the camera as shown in Figure 11.

Figure 11. The recognition result of closer vehicles (Photo/Picture credit: Original).

4.5. Video recovery

The separated frames could further be recovered to a HSR-HFR video by combining the frames with the given frame rate which is 120Hz in this project. The video could also be stored separately which will save storage cost as shown in Figure 12.

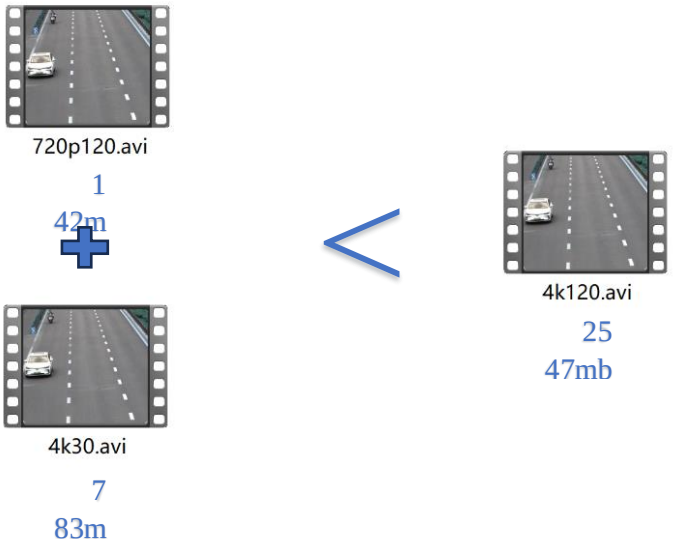

Figure 12. Store LSR-HFR and HSR-LFR rather than HSR-HFR videos (Photo/Picture credit: Original).

5. Challenges and future work

The system is a practical and efficient processing system for highway dual-camera system. It is beneficial to various applications such as license recognition. However, there is also several challenges for this system.

Though the simplest network is chosen for this system, the network is still relatively complicated. This shortcoming could led to hardware and time cost for the HSTR process. Additionally, the accuracy of license recognition is still inadequate for fully automatic process.

To further optimize the system, a lightweight HSTR network could be designed by sacrificing partial clarity since the improvement of extra layers or networks is not significant comparing to the processing speed. Furthermore, the license recognition accuracy could be improved by reduce the distance to the vehicle and pre segment the target area by other networks.

DOI: 10.54254/2755-2721/31/20230137

6. Conclusion

This project successfully achieves its goal of obtaining High Spatiotemporal Resolution through a dual-camera system, which has significantly enhanced the efficiency of license plate recognition applications. The overall architecture of the system has been analyzed from multiple perspectives and dimensions. Taking into account factors such as efficiency and accuracy, the system design is broken down into three parts: preprocessing, processing, and postprocessing.

The preprocessing part includes frame extraction and the process of histogram equalization. Various High Spatiotemporal Resolution networks were explored, and AWNet was chosen due to its high accuracy and simplified network structure. In the postprocessing section, the application of Laplacian sharpening and color adjustment is utilized to restore the frame. Once the frames have been reassembled into a High Spatial Resolution, High Frame Rate video, the license plate is recognized using a system that combines YOLOv5 with LPRNet. This combination provides a popular and lightweight solution for the task of license plate recognition. Findings indicate that the histogram equalization process can be beneficial to the High Spatiotemporal Resolution process. The AWNet was utilized as the High Spatiotemporal Resolution network and the results aligned with expectations and met the required standards. The experiment showed that details from the fixed background and moving objects are preserved by combining the High Spatial Resolution, Low Frame Rate video and the Low Spatial Resolution, High Frame Rate video. The application of sharpening and color adjustment processes improved the clarity and color accuracy by eliminating blurry edges and restoring areas with color differences to their proper color. In the license recognition results, it was shown that the combination of YOLOv5 with LPRNet successfully recognized the license plate, while the unprocessed videos could not, thereby demonstrating that the system successfully increased the accuracy and efficiency of various tasks.

References

[1] C. Liu et al, "New Generation of Smart Highway: Framework and Insights," Journal of Advanced Transportation, vol. 2021, pp. 1-12, 2021.

[2] A. Dosovitskiy et al, "FlowNet: Learning optical flow with convolutional networks," in 2015, DOI: 10.1109/ICCV.2015.316.

[3] H. Zheng et al, "CrossNet: An end-to-end reference-based super resolution network using cross-scale warping," in Computer Vision – ECCV 2018Anonymous Cham: Springer International Publishing, 2018, pp. 87-104.

[4] M. Cheng et al, "A Dual Camera System for High Spatiotemporal Resolution Video Acquisition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, (10), pp. 3275-3291, 2021.

[5] W. Song et al, "Heterogeneous spatio-temporal relation learning network for facial action unit detection," Pattern Recognition Letters, vol. 164, pp. 268-275, 2022.

[6] W. Burger, M. J. Burge and SpringerLink (Online service), Digital Image Processing: An Algorithmic Introduction. (3rd 2022. ed.) 2022. DOI: 10.1007/978-3-031-05744-1.

[7] S. Luo and J. Liu, "Research on Car License Plate Recognition Based on Improved YOLOv5m and LPRNet," IEEE Access, vol. 10, pp. 1-1, 2022.

[8] T. Xue, Jiajun Wu, Donglai Wei, and William T Freeman, "Video enhancement with taskoriented flow," International Journal of Computer Vision (IJCV), vol. 127, no. 8, pp. 1106–1125, 2019.

[9] Z. Zhang, Zhaowen Wang, and Hairong Qi, "Image super-resolution by neural texture transfer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7982–7991.

[10] Z. Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou, "Rife: Real-time intermediate flow estimation for video frame interpolation," arXiv preprint arXiv:2011.06294, 2021.

DOI: 10.54254/2755-2721/31/20230137

[11] L. Lu, Wenbo Li, Xin Tao, Lu Jiangbo, and Jiaya Jia, “Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.