Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Spatial Super-Resolution of Light Field Video

by

Balakumaran Palanivel

A Dissertation submitted in fulfilment of
the requirements for the degree
of

Master of Science in Computer Science

Graphics and Vision Technology

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Signed: _____ Date: _____

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Signed: _____ Date: _____

Acknowledgements

I am extremely grateful to my parents without whom, me chasing my dreams would have remained a fantasy.

To Rosemarie Power, Michael Walsh of the IT department who helped me set up a dedicated machine as per my requirements.

To Donal O'Mahony, who understood my requirements and approved my request for a lab machine.

To Michale Manzke for all of his help and guidance throughout the course of the dissertation.

Balakumaran Palanivel

University of Dublin, Trinity College

August 2018

Spatial Super-Resolution of Light Field Video

Balakumaran Palanivel

Master of Science in Computer Science

University of Dublin, Trinity College, 2018

Supervisor: Michael Manzke

This thesis is a comprehensive feasibility study of the light field video technology using acquisition hardware which is portable enough to be mounted on an aerial platform. It studies various issues in capturing light field video using this platform and focuses on solving the problem of low spatial resolution of the video. It analyses in detail an existing state of the art method to produce light field video and proposes a modification to achieve spatial super-resolution thus enhancing the overall video quality. To achieve this, different spatial super-resolution techniques for light field images are weighed against each other and the best fit to be incorporated into the existing video generation pipeline is determined. Two of these techniques are implemented with necessary modification to work with the light field video pipeline and its results are discussed. This thesis also includes detailed background study on the relevance of a light field and light field video. The problems and benefits of having a portable device capable of generating light field video are also discussed from the perspective of aerial photography platforms.

Contents

List of Figures

List of Tables

1 Introduction

The light field is a concept which has been around in the history of scientific research for a few decades now. In this time period, there have been different definitions provided for the light field, each one accurate from the context of respective domains such as mathematics, physics, computer graphics etc. But in all of these definitions, the common aspect is the fact that a light field always represents an entity which is capable of storing and manipulating light information present in a given environment. From the perspective of computer graphics, it is a far simpler way to interactively manipulate light information than the traditional methods of using 3D scene files. Despite these obvious benefits, the applications in which light fields are being employed currently is considerably low. Because the understanding of a light field's properties and capabilities is not on par with a concept that has been around in the research world for a few decades. This is partly due to the lack of computational power during which this concept was theorised which resulted in studies and research of light fields to be temporarily shelved. But with current hardware advancements, several new light field research has commenced and this has resulted in the exploration of using the light field in many applications. However, this culture of light field research is relatively new and there still exists a vast number of domains and applications which remain completely unexplored. One such domain is light field videography. More specifically, light field videography using portable or hand-held light field cameras instead of using existing complex, bulky and sophisticated camera rigs.

1.1 Motivation

The features and content available through light field enables several applications like depth estimation, autostereoscopic displays, etc. The refocusing ability of the light field can be used to compute defocus cues. Hence, the light field has a very wide range of potential applications in domains such as computer vision, virtual and augmented reality (2). However, an important observation is that the fact that the research and content available in light field videography is surprisingly limited. Another potential reason for this, apart from hardware limitations, could be the very limited reception to the commercially marketed light field cameras, which eventually resulted in shutting down of pioneer Lytro (3). The state of the art light field camera such as Lytro Illum had poor image quality due to the trade-off between angular and spatial resolution. In addition, the continuous acquisition mode could only capture videos at 3 frames per second. Since a set of moving images to be conveniently accepted as a video requires at least 24 frames per second, the Lytro cameras resulted in videos with very little temporal information and poor image quality.

There are highly specialised production cameras such as Lytro Cinema which is capable of capturing videos up to 300 frames per second (4). These videos have the same features of light field images such as the ability to refocus, modify the viewpoint etc. But a crucial difference between the Lytro Illum and Lytro Cinema is that the former is a hand-held camera, while the latter requires a huge custom rig for operation. Hence, although the technology to capture light field video is feasible, it has not been integrated into hand-held (smaller size) cameras. There could be a number of applications and advantages to such a camera as they could be mounted on drones and capture videos which could be easily refocused. It can be used for many applications, for instance observing surfaces of huge aircraft to verify its structural integrity after a thunder strike.

But one of the major issues that affect employing a light field video camera in all these applications is the poor image quality. It is caused due to the low spatial resolution which is a side effect of the spatial-angular tradeoff that was already mentioned. Many solutions

have been proposed to improve the spatial resolution by using super-resolution techniques which are a class of techniques that have been designed to enhance the resolution of an imaging system. But the current super-resolution research in the light field is limited to images. This dissertation extends one of these techniques to light field videos.

Therefore the primary motivation of this dissertation is to study the issues of a portable light field video camera and enhance the feasibility of building one by solving one of the fundamental problems of poor video quality using super-resolution techniques.

2 Background and Related Work

2.1 Light Field

Traditionally three-dimensional graphics systems are used to model real or virtual environments for a wide variety of applications. An input to such a system is a scene built with a collection of intricately placed geometrical primitives. And physical texture properties such as reflection, opacity are achieved by applying materials onto the geometrical primitives through shaders. Following this, the scene is usually lit using a pre-configured lighting system. Based on these input specification, the rendering system synthesises an image of this virtual environment using algorithms such as rasterisation, ray tracing, ray casting etc. Advancement in hardware capabilities such as GPUs and highly optimised rendering algorithms have enabled creating new views and interactive modification of virtual scenes to be rendered in real-time with ease. However, a new approach to rendering has recently surfaced called image-based rendering. Such systems are capable of generating different views of the scene from a collection of pre-acquired imagery and without the need for defining a set of inputs generally required for traditional rendering methods. There are a number of advantages to image-based rendering (5).

- Image-based rendering algorithms utilise substantially less computational resources compared to standard rendering methodologies. This makes the system suitable for real-time implementation on personal devices and workstations where the availability of powerful and sophisticated hardware could be limited.

- Interactive manipulation and viewing of the scene are independent of the complexity of the scene.
- The rendering system is independent of the source of input images. The images can be either from the real world, virtual environment or a mixture of both.

Although there are a number of ways to implement the image-based rendering (6, 7), the most robust technique with a wide range of possible views uses the light field function (5). The light field is simply the collection of light rays in 3D space (8). More formally, this can be described as representing the radiance of light as function parameterised by its position and direction in the scene.

The concept of the Light field itself was first described by Lippmann in 1908 (9). Lippmann proposed, using a set of small biconvex lenses to capture light rays in different directions and referred to it as integral imaging. However, the term "light field" itself was first used by Greshun while studying the radiometric properties of light in space (10). But Light field and Light field imaging have become popular recently, especially after a detailed study of the Light field rendering systems (5, 11). These image-based rendering methods defined by (5, 11) was further extended by Aaron Isaksen, Leonard McMillan and Steven J Gortler by adding a number of versatile features. The fixed two-plane parameterisation of the light field was redesigned to be dynamically re-parameterised which resulted in the feasibility of effects such as post-capture refocusing, variable aperture and autostereoscopic light field images (5). These contributions encouraged a focus on the development of devices that are capable of capturing light field efficiently. Ren Ng made a significant contribution to the light field imaging hardware by designing the first hand-held light field plenoptic camera (12). This was necessary since the primary method of acquiring light field imagery was by using complicated camera rigs. This new hand-held camera will later become the forerunner for the commercially marketed light field cameras by Ren Ng's Lytro (3), Raytrix (13).

2.2 Aerial Photography

The type of photography captured from an aircraft or other flying object is called Aerial Photography. The platforms used for this kind of photography include fixed-wing aircraft, helicopters, unmanned aerial vehicles etc. Recently, the popularity of using images acquired using aerial photography platforms has grown to a very large extent. The global market for aerial imagery is projected to reach $3,545 million by 2023, at Compound Annual Growth Rate of 13.4 per cent during 2017-2023 (14). With this enormous market growth, the technology used to capture aerial photography has substantially improved, and the number and variety of platforms available have grown. Early platforms included using primitive methods such as hot-air balloons, kites, pigeons, rockets and fixed-wing aircraft. These platforms often carried unstabilised film cameras which introduced motion-blur and had a very limited resolution. However, technology modernisation has enabled unmanned-fixed wing aircraft, helicopters and multi-rotor unmanned aircraft systems to be used for aerial photography. Quadrotors are inexpensive and accessible consumer product these days. In addition, digital camera systems with very high resolution have become quite standard and gyro-stabilised platforms have addressed the issues of motion blur to great extent.

2.2.1 Applications

The main reason behind aerial photography's substantial economic growth is the versatile applications in which it can be employed. For instance, in the field of exploratory geophysics, aerial photography can be used to determine the physical components of subsurface and its anomalies. Apart from that, there are a few more types of explorations in which aerial photography holds quite an unbeatable importance namely, fossil fuel exploration, hydrocarbon exploration, geothermal exploration and groundwater exploration. To put it simply, aerial photography has enabled any geologist to extract a precise and concentrated picture of the earth and analyse it to determine particular conditions.

Another application of aerial photography that is currently being explored is using drones for the inspection of aircraft surfaces after a flight. This inspection is necessary for testing the integrity of the aircraft after a thunder strike or flying through a highly turbulent weather. This study is usually carried out by a number of engineers visually observing the aircraft surface which requires a platform to be built around it. This takes considerable time, in contrast, a drone can fly around a plane much faster by snapping pictures of the aircraft. These pictures are later analysed by engineers on a screen. The primary requirement of such pictures is to have the adequate visual detail to replace the visual check for routine issues, thus shortening the amount of time a jet is out of service. This system is being widely tested to be incorporated into regular operations of major airlines. (15)

2.2.2 Cameras Used

The cameras used for aerial photography are usually high-resolution digital camera with a traditional setup which contains the main lens that focuses the rays entering the camera at the image sensor. The rays here represent a collection of rays that propagate in the same direction and have the same colour. The camera setup usually also contains some form of a pupil or iris that prevents light rays from reaching the surface of the image sensor. Essentially, what gets recorded in the sensor is the total power of the rays which are hitting a particular location of the sensor. This is shown in Figure 2.1. Hence, a traditional camera will not be able to identify each of the rays that are being summed together to produce the final image.

On the other hand, a light field camera will be able to record the rays themselves within the camera. This is shown in figure 2.2. The popularity of light-field photography has increased substantially in recent years since the development of hand-held plenoptic cameras (12) and due to the benefits of light fields itself.

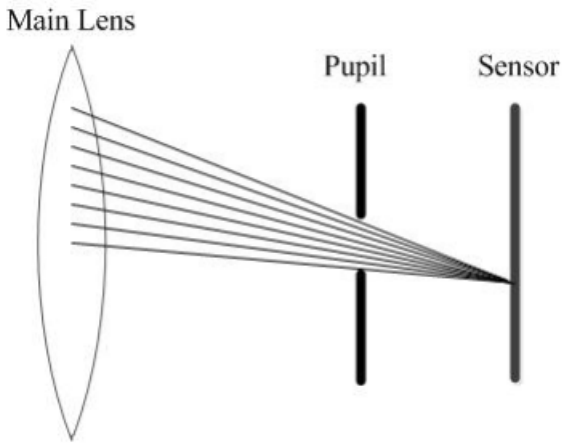

Figure 2.1: This is a model for a traditional camera. A main lens focuses multiple light rays at an image sensor. A pupil restricts which light rays hit the sensor. In a traditional camera, the sensor is able to record only the total of all rays hitting the sensor.

2.3 Plenoptic Camera

A plenoptic camera is equipped with a microlens array or printed mask that is placed directly over the camera's imaging sensor. This filters the light that passes through the camera. Based on the properties of the main lens and the filtering device, light-field rays can be extracted from measurements on the imaging sensor. Each ray is described by its two-dimensional intersection with the main lens and its two-dimensional intersection with the image sensor (16). These two measurements constitute a four-dimensional ray. A collection of such rays essentially make up the 4D light field. This can be processed using integral imaging to create a 2D image (5). Plenoptic cameras are referred to as light field camera because they capture the light field and for the rest of the paper, they will be referred to as light field cameras.

The interest in the 4D light field is mainly due to the fact that such an image contains more information in the scene than a traditional 2D image. With this additional information, new features can be computationally generated that is normally not possible with a 2D image. For instance, stereo images can be extracted from 4D light field image, which allows them to be used for 3D visualisations and other computer vision applications(17).

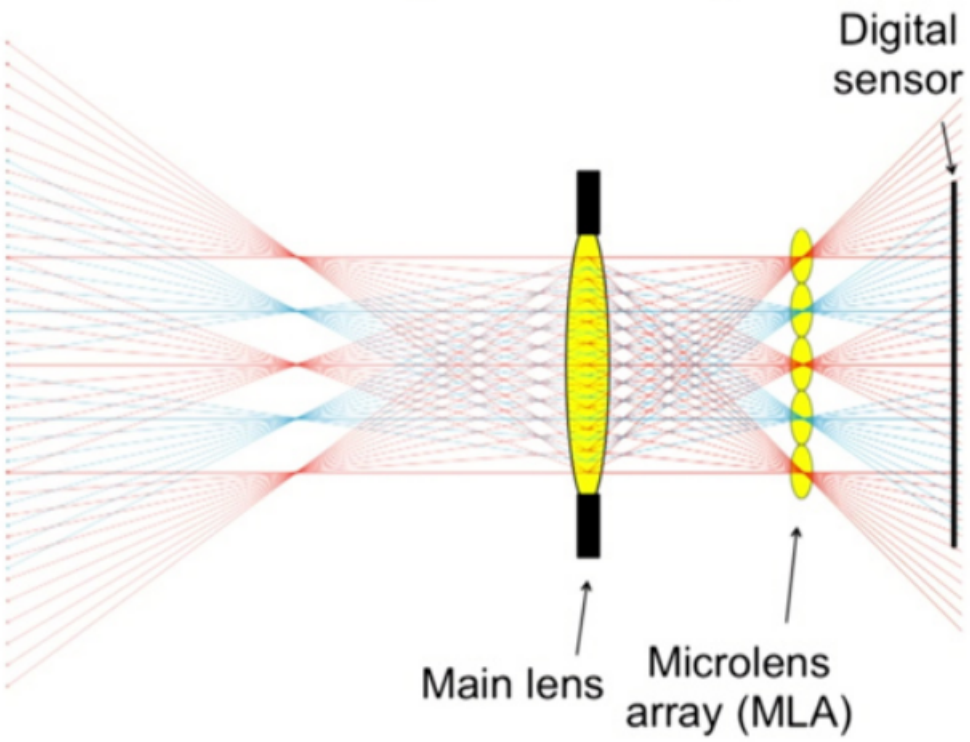

Figure 2.2: In contrast to traditional camera, A light field camera is able to record each of these rays in a 4 dimensional function. Each ray is defined by the location it enters the main lens and the location it hits the image sensor. Two of the dimensions represent the location on the main lens where the ray passes through. The other two dimensions represent the location where the ray hits the image sensor.
https://www.picturecorrect.com/tips/the-future-of-light-field-cameras-and-how-they-work/

The complete 4D light field is not required to create a 2D image. It is possible to make two virtual eyes with a single-lens camera, which allows the user to manipulate the images produced in a manner similar to those produced by a binocular system. Using ray tracing, it is then possible to determine how the light rays within the 4D light field hit each virtual eye. This virtual eyes then produce stereo images (16).
It is possible to digitally refocus a light field image at a different depth of field. This is done by changing the focal plane of the image with a Fourier slice transform (18).

2.3.1 Advatages of Light Field Camera

The aspect of an imaging system to virtually change the focal plane post-capture offers functionality that traditional imaging systems do not. This decouples the need to select

the appropriate focus of an image during capture time, hence there are fewer requirements for real-time focusing and calibration. In most camera systems, the iris of the camera is fixed and based on the brightness of the scene, the exposure time is set. Imaging systems that have focal length set to infinity do not require adaptive focusing and most airborne systems are focused in this way. However, any camera system which is focused at infinity will produce an out-of-focus image when those images are captured from an altitude lower than the focal distance. Focal distance is determined by a number of parameters of the imaging system such as image sensor size, pupil size and focal length of the main lens. Hence for small systems like a DSLR camera, a typical range for focal distance can be between 0 and 25 feet (19). But since a light field camera allows for correction of images that were not properly focused at capture time, there is considerably less dependency on the hardware parameters of the imaging system in determining the focal length. A software application can be built to change the focal depth based on the user's need. Essentially, the focus of the captured image is dictated by the end user and not by the capturing system.

In addition, since the trajectory of the light in the scene is captured in the 4D light field image, it can be used to estimate the depth of the objects within the scene. The light trajectory is interpolated to the origin of the beam with reference to the camera. This is accomplished by taking the gradient of the sampled light field and tracing it along the light rays as they pass through the camera. This information which relates to the path a light ray takes through the camera enables for the interpolation of the origin and direction of the ray. Hence, the depth is estimated by a collection of ray origins and the point at which they intersect with the main lens.

2.4 Light Field camera for Aerial Photography

Traditional light field image acquisition techniques involved sophisticated multiple-camera setups. They require complex rig mountings to hold the cameras which have to be carefully calibrated. Building such a system with adequate structural integrity and wide

viewing angles on an airborne platform are practically infeasible (at least with current technology). Hence a single-lens light field camera is best suited to capture the 4D light field from a drone. This also prevents the additional costs incurred by the traditional systems involving multiple cameras.

But commercially available handheld light field cameras have not been developed incorporating the design elements required for a camera to perform effectively on an airborne platform which is constantly under motion and is relatively unstable. They have very short range image capture and recently has been expanded to be feasible with mid and long ranges of up to 100 m (20). Such range becomes relevant if light field cameras are to be used as potential mono-sensorial range imaging devices in autonomous cars or in mobile robotics. In contrast, it is essential to note that long-range drone aircraft are capable of operating at a range of up to 7000 m. On such a platform the mid-range light field cameras may not be suitable, however, a wide variety of exploratory and investigative applications can comfortably be performed under a range of 70-100 m. For instance, the application of using a drone to observe the surface of an aircraft will not require a camera with a range of more than 100 m. Besides the results of (20) shows that depth measurement accuracy deteriorates with depth. At depths of 30-100 m, which may be considered as typical drone operation range for many applications, depth errors in the order of 3% were obtained from processing small point clusters on an imaged target. Higher errors were also obtained from single point analysis, which stresses the necessity of spatial or spatiotemporal filtering of the light field camera depth measurements. However, despite these obviously large errors, a light field camera may nevertheless be considered a valid option for many applications such as robotics, autonomous driving and unmanned aerial vehicles because of the expansive information available in a 4D light field image, compared to the traditional image (20).

2.5 Aerial Videography

In addition to capturing images, airborne platforms are also very popular to capture videos. There is a growing interest in performing aerial surveillance using video cameras. Compared to traditional images, videos provide the capability to observe ongoing activity within a scene and to automatically control the camera to track the activity (21). It is widely used in several applications such as aerial video inspection of overhead power lines (22), vehicle detection and tracking (23), using aerial video for traffic flow monitoring and management (24) etc. Such cameras require a number of specialised features to ensure the video is of the highest quality. In addition to hardware-based features such as gyroscopic stabilisation, there are several technical challenges that have to be addressed for using the video for effective surveillance. Kumar claims that due to high data rates and relatively small field of view of video cameras, a specialised framework is required which is capable of real-time, automatic exploitation of aerial video for surveillance applications (21).

2.5.1 Aerial Videography using Light Field Camera

The advantages and uses of having a camera mounted on an airborne platform to capture light field images are numerous. But the value proposition of a camera capable of capturing light field video is considerably higher as it can be used for a number of applications. The ability to change the focus of different parts of a video after it has been shot provides great flexibility. It also prevents the effort involved in focus pulling, i.e manually shifting the focus plane of the video to remain focused on a moving object within a shot (1). The concept of light field video is not entirely new. Theoretical and practical studies have been conducted on light field videos since Wilburn designed the first light field video camera (25). However, most such cameras used a collection of miniature digital cameras arranged together to form a camera array. Most of the times the cameras were crammed in a compact package to ensure it can be hand-held, such as the ProFUSION-25c (26). And this hardware limitation often resulted in cameras which had very less angular and

spatial resolution. Hence although the ProFUSION was capable of capturing light field videos at 25fps it had a resolution of only 640 x 480. But modern light field cameras have substantially high spatial and angular resolution compared to this. And for a camera mounted on an airborne platform, this is essential as the object of interest could be far away and higher resolution cameras can capture more information of objects at distant. Despite the low resolution of early light field video cameras, several studies performed using them have shown the advantages of a light field video. Light field video cameras generate multi-view video streams with depth information of the scene. This essentially is a 3D representation of the entire scene which can be effectively used in domains such as 3D television and free viewpoint video (27). Also, the work by Brandon and Li described a novel method to stabilise the videos shot with a hand-held video camera as they have a considerably high amount of camera shake compared to professionally shot videos (28). They claim that video stabilisation is an image rendering problem where new novel images have to be generated given a set of input images captured along a path that is shaky and unsteady. The new output sequence of images has to be along a virtual path different from the original path which was shaky. Solving this problem using a traditional video camera is particularly challenging as it generates a standard video with only single viewpoint available at any given time instant. This introduces parallax and the effect of occlusion between the desired and actual viewpoints. On the other hand, the video generated by a light field video camera provides several viewpoints for any given time instant. Such a video enables interpolation or extrapolation of novel viewpoints using view synthesis methods (28). This remarkably increases the quality of video stabilisation which is possible only by capturing video using a light field video camera. Although there are numerous benefits as mentioned above, the light field video camera poses a set of specialised complications, especially when it is required to be mounted on mobile platforms such as drones.

2.5.2 Existing Challenges

The crucial factor limiting the feasibility of light field videos is the maximum data transfer speed, given a limited bandwidth. The fundamental operating principle of a traditional camera is the recording of the colour of light at each pixel. Hence, the amount of data collected is directly proportional to the resolution of the image. A light field camera, although still operating under the same principle of recording the colour of light, the volume recorded is exponentially higher. Light field cameras have two distinct resolutions, spatial and angular. The angular resolution is the number of different views that are captured in a single exposure. This is proportional to the number of pixels captured in each of the lenslets present in the camera array of a hand-held light field camera. On the other hand, the spatial resolution is the visible resolution of the image produced by each view of the camera. This is proportional to the number of lenslets present in the camera array. In the light field camera built and studied by Andre Ng (12), the light field is captured by an array of 2692 lenslets inside a conventional camera. Each lenslet in this setting corresponds to one viewpoint on the aperture, while different lenslets correspond to different pixels in the final image. The result is an approximately 100-view light field with 90,000 pixels per view. Now, recording a scene in both spatial and angular domains of a light field takes a significant amount of data. For instance, the raw output image of the Lytro ILLUM camera is 5300 x 7600 pixels, which is nearly 20 times the resolution of 1080p videos. Assuming similar bandwidth as a 1080p 60 fps video, one can record light field images only at 3fps. As a matter of fact, this is the rate at which continuous shooting mode is designed in Lytro ILLUM (1). There are modern advanced cameras such as Red camera (2017) or Lytro Cinema which can shoot at a higher frame rate. However, they are extremely expensive and quite bulky which makes it far too complicated to be mounted easily on an airborne platform such as a quadcopter.

2.6 Portable Light Field Video Camera

Therefore, it is obvious to arrive at the conclusion that a light field camera that could be comfortably mounted on an airborne platform should certainly not be greater than the size of a standard hand-held camera. However, the current state-of-the-art hand-held cameras, one of which is Lytro Illum (3), can only capture static images and cannot shoot videos even at 24fps due to the data bandwidth issue described in section 2.5.2. This gap was addressed by the work of Wang who used a learning-based approach to interpolate the low fps video shot by a Lytro Illum camera into a 24fps light field video (1). However, this work uses a hybrid imaging system. A hybrid imaging system is a setup in which two or more independent imaging system, such as, DSLR imaging, PET imaging (Positron emission tomography), CT imaging (Computed Tomography) are fused together to form a new image modality technique. The goal here is to combine the innate advantages of the independent imaging systems to develop usually a new and powerful imaging system. Some existing hybrid imaging systems include PET/CT, SPECT/CT, ultrasound and CT etc. Wang uses a system where a light field and DSLR imaging systems are combined. This system is focused to capture a scene simultaneously and machine learning is employed to combine the outputs of both the cameras to generate the desired light field video. This research essentially performs the interpolation of light field video captured by the Lytro Illum camera. But this video has only three frames per second, which is very less information to perform temporal interpolation comfortably. To bridge this gap, the input from the DSLR video camera is utilised, which is a 30 fps standard high-resolution video. The machine learning-based approach employs two convolutional neural networks (29). The first CNN is designed to propagate the angular information from the light field sequence to the standard video. This enables warping of the input images to the target view. The second CNN combines the warped images to output the final pixels (1). This research successfully enables consumer light field videography and also outperforms current video interpolation methods.

However, this research still has a number of limitations. For instance, since a learning-

based approach is used to estimate the flows between images, the CNN is unable to determine the features in the regions of the image which are occluded. The flow remains undefined in such regions as the information is not available and this produces errors during the interpolation of images. As a result, artefacts are introduced in the occluded regions. There are more limitations which can strongly impact when the solution proposed in the research is employed on an airborne platform.

First, the light field camera employed in this research, Lytro Illum, has a small baseline, hence the object that is shot using this camera cannot be too far away. This aspect directly contradicts the operating principle of aerial photography, where objects can easily be at least 25 m away from the camera. The primary solution for this is a hardware change, that is, using a camera which has a larger baseline at the same time is also smaller and attachable such as phone cameras. Alternatively, Wang also proposes a more fundamental solution to this problem by proposing to integrate the baseline parameters of the camera directly into the learning system. Training the CNNs end to end with information about the baseline can lead the network to learn and infer the differences between the two cameras' hardware configurations (1).

Second, the presence of motion blur in the video causes the interpolation of the flow between two frames of the scene to often fail. This issue was prevented by having an exposure time small enough so that no motion blur occurs (1). Since cameras used for aerial photography also have a wide range of shutter speed, the exposure time can be made small enough to ensure no motion blur occurs (30).

Finally, a crucial hardware based limitation of integrating Wang's research on an airborne platform is the usage of a hybrid imaging system. Using a DSLR and a light field camera in parallel on a drone is challenging. However, there are platforms that allow external hand-held cameras to be mounted upon a drone, but currently, they do not support any commercially available light field cameras. One such platform is GDU Technology's Byrd Premium 2.0 (31). It comes with a 4K camera inbuilt on the aircraft and an ability to mount a GoPro (32) or any other third-party camera.

In addition, the entirety of Wang's work is a focused approach to effectively propagate

all the viewpoints captured using the 3fps continuous shooting mode of the Lytro Illum camera to the 30fps DSLR video, essentially creating a multi-viewpoint light field video. In other words, in this process, only the angular resolution captured by the light field camera is preserved when it is interpolated into a video. The high spatial resolution of the video generated by the DSLR camera is not used and therefore the overall light field video generated is of low spatial resolution. This is mentioned by Wang as one of the potential future works (1). But this is a limitation that highly impacts the overall benefits of a light field video generated when captured from a drone. The high altitude from which the drone captures the scene means the cameras that are mounted on them should be able to produce high-resolution output with detailed information of the scene. A low-resolution video would mean that a lot of information is lost and zooming results in pixelation. This completely defeats the purpose of aerial photography and therefore fixing this is one of the primary factors for the feasibility of light field camera.

But the field of light field photography has always had the problem of generating lower spatial resolution images compared to standard DSLR cameras. This results in lower image quality. According to several online reviews, the lower image quality of Lytro Illum cameras is in fact, one of the main factors behind the fall in sales of Lytro's commercial light field cameras (33). This is due to the fact that in light field photography from earliest work of Lipmann (9), where the same was referred to as integral photography, to more recent approaches of Adelson (34) and Ng (12), known as plenoptic cameras, a common goal was that of increasing the angular resolution of the 4D light field measured. This often comes at the cost of spatial resolution of the 2D images that are rendered from the 4D light field (35). This drawback is very closely coupled to the hardware design of light field cameras and hence remains inevitable. With the increase in the number of lenslets in the camera array the spatial resolution increases, but more lenslets mean, that each individual lenslet is smaller in size, which results in smaller angular resolution. Due to this inverse proportionality, there has always been a trade-off between the spatial and angular resolution during the design and development of light field cameras. Although modern light field technology has enabled development of cameras with good spatial and

angular resolution, from the standpoint of professional photographers, the overall image quality with respect to spatial resolution is still considered to be poor.

2.7 Super Resolution Algorithms

To overcome the disadvantage caused by the trade-off between angular and spatial resolution and to enhance the overall image quality of light field cameras several computational methods have been proposed. These methods are usually called super-resolution algorithms, which as the name implies processes the image and attempts to increase the resolution of the final image programmatically. Essentially, super-resolution algorithm determine additional information from existing data (36). There have been comprehensive studies of both angular and spatial super-resolution. According to Gul and Gunturk there are three different approaches to achieve super-resolution of light fields gul2018spatial.

- First is the approach of directly processing the perspective images of the light-field by applying multi-frame super-resolution techniques. The works of Wanner and Goldluecke (17) and Bishop, Zanetti and Favaro (36) follows this approach.
- Second is using machine learning-based techniques on each individual sub-aperture perspective images of light-field to achieve super-resolution. Gul and Gunturk's research follows this approach (8). They have developed two CNNs, one to enhance spatial resolution and the other to enhance angular resolution. There are other researches which follow this approach, however, Gul and Gunturk's CNNs performs better since angular information is also used to train the network which enhances spatial resolution.
- The final approach is using a Hybrid imaging system where a light field sensor and a standard image sensor, usually some kind of a digital camera, are used in parallel. In these methods, the regular sensors generate high spatial resolution images. The information available in these images are used to enhance the perspective images

generated by the light field sensor. Boominathan's research uses this approach (37). But this system is bulky as it requires additional hardware which also increases the overall cost compared to the first two approaches.

Each of these methods and the research work pertaining to them were analysed in detail to gain insights into the super-resolution technology. The information obtained from this analysis was crucial in shaping the course and structure of the entire project. Hence the super-resolution analysis is documented in the section 3 that focuses on the design of the project, instead of the current section.

2.8 Summary

This dissertation evaluates light field videography based on the work of Wang (1) which already employs a hybrid imaging system. This essentially enables usage of any of the three approaches mentioned above for enhancing the spatial resolution of the light field video. Theoretically, all of the above three approaches can be integrated with the light field video pipeline but the practical aspects of the three methods need to be evaluated and integrated with the system to try and produce a high-resolution light field video.

3 Design

The overall goal of the project is to generate a high spatial resolution light field video. The granular description of the individual goals of the project is listed below. Since this dissertation is a direct continuation of Wang's work, the success and failure of each of the goals are benchmarked against the light field video produced as output by Wang's pipeline.

- The final light field video generated should have a higher spatial resolution than the video generated by Wang (1).
- The angular resolution of the final light field video should be preserved or enhanced compared to the video generated by Wang (1).

The following design elements were crucial to achieving the overall goals of the project.

- Design and development of the component that performs the spatial super-resolution of the light field video.
- Determining the integration point of the component mentioned above with Wang's (1) existing light field video generation pipeline.

Both the aspects listed above are not completely independent. Certain decisions were made because of this interdependency in designing of the overall pipeline.

3.1 Spatial Super Resolution Component

Based on the survey of existing and state of the art methodologies there are three different categories of computational super-resolution. These are listed in Section 2.7 and the comprehensive study carried out in deciding between these three methods is described in this section. In each category of super-resolution, one research publication is considered to determine the feasibility of extending the solution proposed by the aforementioned category to super-resolve light field videos. Each of these research work was selected such that either they are the state of the art in the particular super-resolution category or have comparable results to the state of the art.

3.1.1 Learning Based Super-Resolution

In the work proposed by Mitra and Veeraraghavan (38), a machine learning procedure is used to super-resolve low-resolution light fields based on the disparity between overlapping patches of the light field. A patch here refers to the individual light field views. An offline system learns a Gaussian Mixture Model prior using high-resolution light field training datasets for different discrete values of disparity. This model combined with a MAP estimator is employed to super-resolve each patch in the test dataset with the corresponding disparity. But one of the major drawback of Mitra and Veeraraghavan's work (38) is in the elemental design decision of assuming constant disparity for the light field patches which introduces artefacts in the super-resolved image. In Wang's work (1) on light field video, estimating the disparity in each frame is an essential step to determine the optical flow of the objects in the scene. This is important for interpolation and is also recorded that in occluded regions of the scene where the disparity estimation fails artefacts are introduced in the rendered light field video. It is worth noting that any super-resolution technique has to be applied to each individual light field frame used to generate the final video. Therefore, applying Mitra and Veeraraghavan's technique of spatial super-resolution to multiple light field frames, which has a known issue of

introducing artefacts because of the constant disparity assumption, will result in artefacts in each of the output frames. With artefacts being introduced in every output frame the disparity estimation step will fail and overall error in disparity will get accumulated. This reduces the features present in the output light field video because considerable information is lost in the input frames due to artefacts. Therefore, despite the increase in the resolution, it was decided that this was not a suitable technique to be used because the loss of information is extremely undesirable.

3.1.2 Hybrid Super-Resolution

Since Wang's light field video generation solution works by using the hybrid image acquisition setup, this form of super-resolution is in fact mentioned as future work in the paper itself (1). But since the goal of Wang's work was focused only on the angular resolution, the high-resolution DSLR video was downsampled to match the spatial resolution of the low-resolution light field image frames from the Lytro Illum camera. Hence, the DSLR's high-resolution information is never used, and the work of Boominathan (37) does exactly this - use the spatial resolution information in DSLR of a hybrid imaging system and super-resolve the light field images. Although this dissertation is a direct continuation of Wang's (1) work which used the hybrid camera system for image acquisition, there was no plan to build a hybrid camera rig for experiments. Hence, this dissertation relies on the light field datasets publicly available. However, according to best of the author's knowledge, there are no publicly available datasets where the scene was captured using a DLSR-light field camera hybrid setup except the one published by authors of (1). Any hybrid imaging based super-resolution certainly requires the high-resolution input along with the low-resolution input, which in this case are the DSLR 2D video and frames from the Lytro Illum light field camera respectively. But unfortunately, the dataset of (1) contains only the downsampled DSLR 2D video and the Lytro Illum frames both at a low resolution of 512 x 352. Since there was no other practical way to acquire the required data without building a camera rig, this technique was explored by defining

some workarounds. This is documented in section 4.5, but a feasible solution could not be built with the workarounds, therefore other methods were explored. But this method would have been the most ideal and elegant way to super-resolve the video considering the hardware setup was already available. The features of the high-resolution DSLR video could have been learned by the CNN by letting it train end-to-end with input from both the camera. This would have enabled the network to identify the spatial information and propagate it to the light field video.

3.1.3 Computational Super Resolution

Any super-resolution process which does not employ machine learning, CNNs or any custom hardware such as camera rigs and solely relies only upon logical and mathematical techniques to enhance the spatial resolution of the light field images is considered as computational super-resolution. There are several publications which fall under this category. But this vast literature of super-resolution algorithms can be classified into two major types namely, single-frame and multi-frame techniques. As the name implies in the single-frame technique a single image from the scene is processed individually and independently to increase its resolution. For a light field, the single-frame technique can be applied to every view separately. This eventually increases the overall spatial resolution of the entire light field. However, such an approach will fail to exploit the high correlation that exists between the different views of the light field (39). Also, since each of the views is processed independently, it runs the risk of failing to enforce or maintain the consistency between the light field views.

These can be avoided by using a multi-frame super-resolution algorithm where multiple images of the same scene are used in tandem to enhance the resolution. This approach performs better with light-field images as the design of the algorithm resembles the structure of the light field. In fact, a multi-frame super-resolution technique can also easily be extended for super-resolving videos. Hence we consider the possibility of exploring such a technique as a way to super-resolve the light field video. In particular, the work by

Rossi and Frossard, light field-super resolution using a graph-based (GB) regularisation (39) is used for this purpose. In addition, the state of the art super-resolution methods of this category such as Wanner and Goldluecke (17) and Mitra and Veeraraghavan (38) exhibited lower performance in comparison, both in terms of peak signal to noise ratio (PSNR) and overall visual quality.

A crucial step in super-resolution is the estimation of disparity of light field images input to the system. This is a very challenging because, usually the input is of very low-spatial resolution. And the GB regularisation bypasses this challenge because it did not rely upon accurate disparity estimation like other state of the art methods. Therefore, it was independent of scene geometry and this was another important parameter in deciding to use this method.

3.1.4 Deep Learning Based Super-Resolution

Convolutional Neural Networks have also been used to address and solve the problem of light field super-resolution. In such techniques, a common approach is to build individual CNNs each specialised to perform a single task and operating them in a cascade to achieve the overall result. For instance, the work of Gul and Gunturk (8) two CNNs are built - one to super-resolve the given light fields and the second to generate novel high-resolution views based on the super-resolved input from the previous CNN stage. However, the CNN employs a single-frame super-resolution technique due to which the views are independently enhanced. This is a probable rational behind the fact that the computational super-resolution method of GB regularisation which uses multi-frame super-resolution performs better than the CNN based approach (39). And due to this reason CNN based spatial super-resolution is not explored in this thesis.

3.1.5 Verdict

Based on the detailed practical and theoretical analysis of various super-resolution technique, it was decided to integrate the graph-based regularisation of light field super-

resolution (39) with Wang's work (1) for the spatial super-resolution of light field video. Note that an implementation was attempted with hybrid super-resolution technique, but later pivoted to use this technique instead. More details about this in section 4.5.

3.2 Integeration to Pipeline

Now that the technique to be used for spatially super-resolving light field video has been finalised in section 3.1, the next crucial thing to do is to determine how to integrate this solution with the light field video generation pipeline (1). A fundamental task involved in making this decision is comprehensive study and understanding of the pipeline. However, this section is not a summary of the entire publication of (1). It only highlights the important aspects of the video generation pipeline which were considered for deciding where and how to integrate the GB super-resolution algorithm (39).
Since the following subsections highlight certain components of Wang's (1) work, the block diagram of the entire pipeline is included in the figure 3.1 below for better understanding of the text.

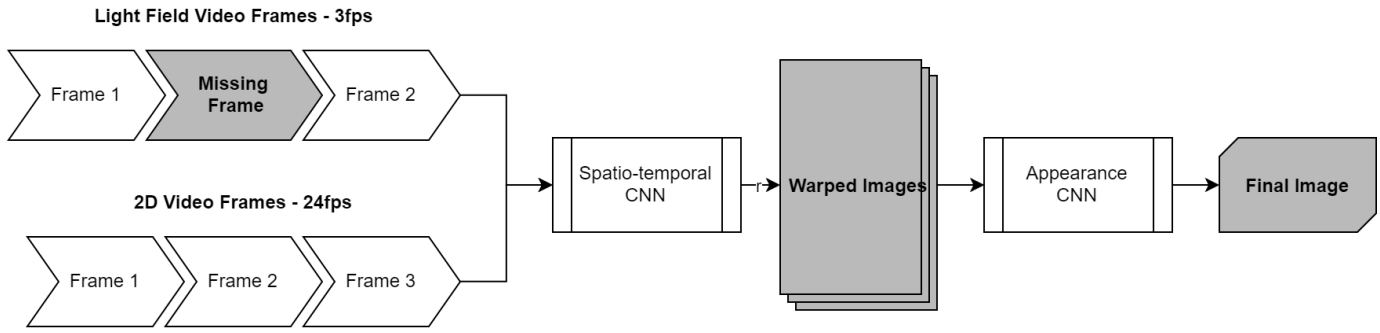

Figure 3.1: Original Light field video pipeline by Wang (1)

3.2.1 Loosely Coupled Integeration

Standard software engineering principles dictate that individual components should be adequately abstracted and decoupled. However, analysing the pipeline revealed that it has not been developed with room for much extensibility and is a closely coupled

pipeline. An attempt to introduce modularity to include the GB regularisation technique into the pipeline proved to be very challenging and also resulted in errors in the final video generated. Hence it was decided best to leave the light field video generation pipeline untouched and integrate the spatial super-resolution algorithm at the beginning or end of the pipeline.

3.2.2 Closely Coupled Integeration

As mentioned in Section 3.2.1, the pipeline is very closely coupled. The two main components of this pipeline are the spatiotemporal flow estimation CNN and the appearance estimation CNN (1). These can be observed in the figure 3.1. The former is used to warp the light field images and the frames input using the 2D video to the various angular view. This part performs the combination of temporal and spatial information in the 2D video with the angular information in the light field frames. The latter is used to accumulate together all the warped images to generate the final image. The appearance CNN is specialised to perform only an aggregating task and does not contribute in any form to the overall quality of the image. Hence, increasing the resolution should be a before this part, that is, the warped images provided as input to the appearance CNN should be the spatially super-resolved images.
There are two distinct ways this can be achieved as listed below.

- Before the start of the video generation pipeline.
- After the Spatiotemporal CNN stage but Before the Appearance CNN stage.

Each of these positions has benefits and pitfalls and have been discussed in dedicated sub-sections.

3.2.3 Adding at the beginning of the Pipeline

In this position, the data from the hybrid image acquisition setup is processed by the GB-regularised super-resolution method before being input to the light field video gener-

ation pipeline. For this to work, the light field images from the Lytro Illum camera have to be spatially super-resolved and then fed to the spatiotemporal CNN. A high spatial resolution light field input will greatly help the disparity estimation step of the spatiotemporal CNN as well. Hence in addition to increased resolution output, better disparity assists in the accurate determination of optical flow and hence the overall video interpolation quality can also enhance. While there are potential theoretical advantages there are some practical challenges to this approach such as implementing the GB-regularisation technique that works with input from the Lytro Illum light field images.

3.2.4 Adding in-between the CNNs

In this position, the warped image output by the spatiotemporal CNN is spatially super-resolved before being input to the appearance estimation CNN. This is a practically simple design since only one set of images are input to the appearance CNN since redesigning the corresponding network to accommodate the super-resolved image dimension is relatively straightforward. The GB-regularisation technique can also be implemented easily because at this stage the data to work on is a set of warped images, which is essentially images of the same scene with slightly different viewpoints. This is very similar to the data used to test the GB-regularised method which was images acquired with a grid of camera which was considered as the light field instead of using a plenoptic camera with a camera array.

3.2.5 Verdict

Based on this study of various options for implementation, it was decided to implement the GB Regularised super-resolution component at the beginning of the pipeline. Because at this position Wang's (1) pipeline is not disrupted at all in any way. This would also mean that, in overall, the pipeline would have three well-defined and self-contained components each abstracted from each other. And these components are interconnected by data processing units which will convert the input and output data of each component

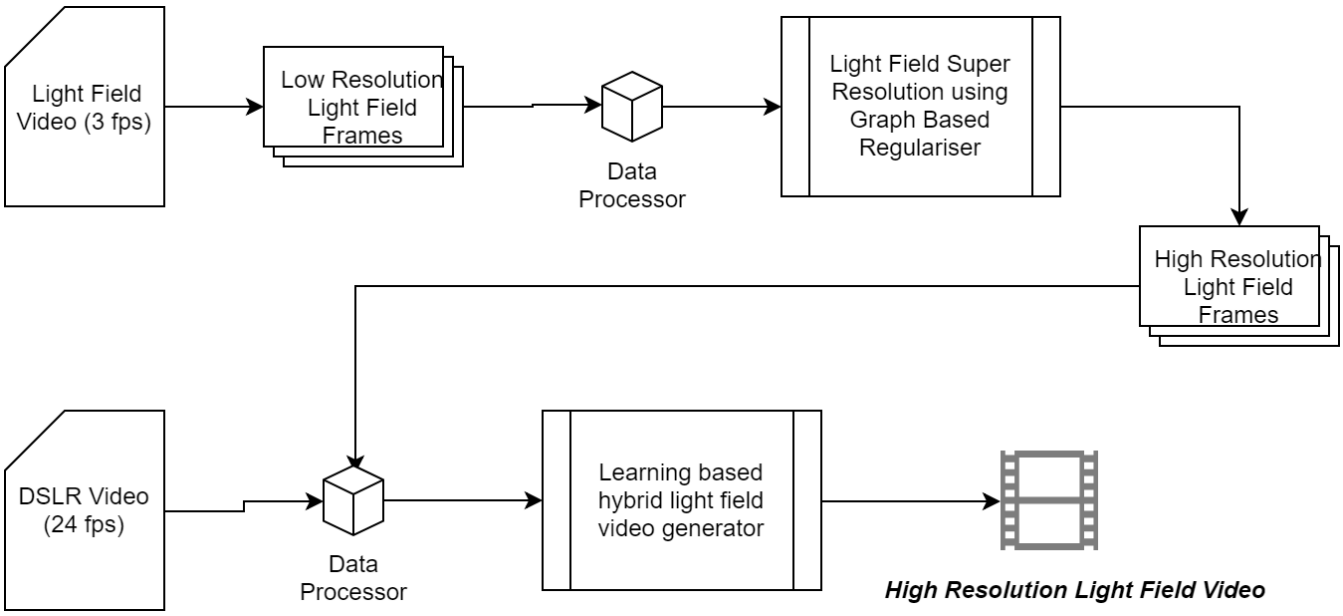

Figure 3.2: Overall Pipeline design diagram

into suitable formats. This is described in the block diagram 3.2.

Such a design will ensure that any errors or issues realised during the implementation process can be isolated to specific components and can be addressed using a compartmentalised approach.

3.3 Light Field Video Viewer

This is another important part of the project that is essential for the easy visual perception of the light field video. Although it is not directly part of the pipeline that generates the light field video itself, a light field video viewer is required to conveniently view the generated video and highlight the aspects of the light field such as post-capture refocusing, post-capture aperture control etc. After an adequate background check, it was verified that there was not an out-of-the-box application which was capable of running light field videos. There was no public standard for light field video formats or methods to highlight the significant aspects of a light field while it was in the video format. This was probably one of the main reasons why Wang (1) decided to develop a special light field video viewer capable of interpreting the generated video and also modifying the camera properties while the video is playing.

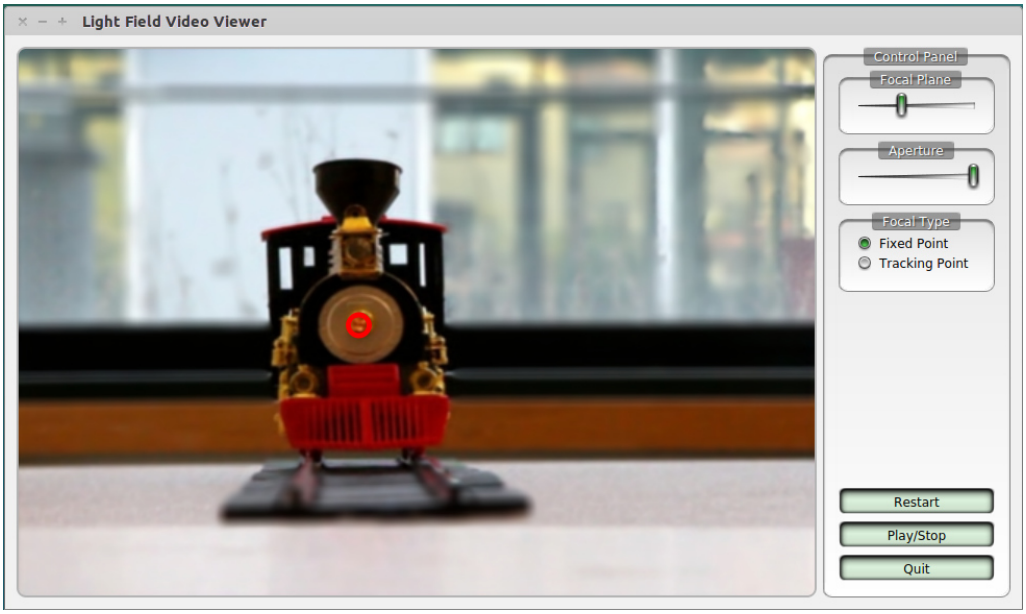

Figure 3.3: Light field video viewer from (1)

3.3.1 Zooming

This viewer, however, did not possess any specific way to highlight the spatial super-resolution of the frames generated by the system. An ideal way to visually observe the effects of super-resolution would be to zoom the super-resolved and non-super resolved video to exactly the same extent and observe for image distortions or pixelations. But there was no feature in the light field video viewer program to achieve the zooming of the video. And it was decided to introduce this feature into the program as part of this project. The details of how this was achieved are not discussed in this thesis as it was relatively straight forward.

4 Implemetation

The previous chapter highlighted many of the crucial design decisions that were made during the course of the project which eventually resulted in the architecture of the entire pipeline. This architecture is picturised in the figure 3.2. This chapter will describe in detail how the different components of the proposed architecture were implemented. Each section of this chapter pertains to a component and explains the tools, techniques and algorithms implemented in it. The various parameters which were considered to optimise the entire pipeline with respect to each component are also discussed, wherever relevant.

4.1 Overview

This project was initially intended to use the hybrid imaging technique for enhancing the spatial resolution of the video. But the unavailability of the appropriate datasets hindered the feasibility of developing such an application. If this approach had been followed the output would have been a single application which was capable of taking input and producing the super-resolved output. But due to the fact, the project pivoted to use a super-resolution using graph-based regulariser the implementation of this project was broken down into components each of which were implemented using different tools and techniques. The aspect that is worth highlighting here is that this pipeline is not a one-click application, that is, the output data of the super-resolved light field from the first stage (light field super-resolution) of the pipeline had to be manually fed to the

second stage of the pipeline (light field video generation). It is acknowledged that it is a poor design, and after exploration, it was determined that automation of this part of the pipeline should be a relatively straightforward task. Hence this was not implemented in this project.

4.2 Input

The input to the system was the dataset published as part of the original light field video pipeline. The images of this dataset were captured using a prototype setup of a high-resolution DLSR and Lytro Illum connected together using a tripod screw adapter. The two cameras were calibrated, synchronised appropriately when made available to the public, hence can be directly used without the need for any pre-processing (1). Unfortunately, this was the only hybrid imaging dataset publicly available and hence this project was not tested with any other datasets.

4.3 Spatial Super Resolution

This section describes the implementation of the light field super-resolution via Graph-Based regularisation technique for a set of light field images acquired from Lytro Illum camera. The entire implementation of this section was done using MATLAB and involved the development of new scripts and modifying existing scripts to suit this project.

The problem of light field spatial super-resolution can be theorized as the process of recovering the high-resolution light field $\boldsymbol{U}$ from its low-resolution counterpart $\boldsymbol{V}$. Where $\boldsymbol{U}$ and $\boldsymbol{V}$ are light fields which are considered as the output of an $\boldsymbol{M \; x \; M}$ array of pinhole cameras, each of which is equipped with an $\boldsymbol{N \; x \; N}$ pixel sensor. And the resolution of $\boldsymbol{V}$ is $\boldsymbol{(N/\alpha) \; x \; (N/\alpha) \; x \; M \; x \; M}$, where α is the super-resolution factor. To solve this problem, it can be formulated as an optimisation problem which involves minimising the

following objective function (39).

$$u^* \in \mathit{argmin}_u F(u) \tag{1}$$

where

$$F(u) \equiv F_1(u) + \lambda_2 F_2(u) + \lambda_3 F_3(u) \tag{2}$$

Each of the terms in the equation 2 captures the constraints enforced on the problem by the structure of the light field. And the constant multipliers λ_1 and λ_2 balance out the effect of various terms in the equation.

The first term $F_1(u)$ represents the consistency between the high and low-resolution views. The relationship between the two is captured using a blurring and sampling matrix and is denoted in (39) as the data fidelity term. It is implemented by calculating the error between the high and low-resolution views and since the entire equation 2 is minimised for optimisation, the difference or error between the output and input is minimised and the fidelity is preserved.

The second term $F_2(u)$ in the equation is used to represent the contribution each neighbouring views make to enhance the resolution of the view that is currently being processed. This term ensures the multi-view structure of the light field is accounted for. It also benefits from the scenario where the current view has some occlusions which the neighbouring views may not. The un-occluded regions of the neighbouring views provide complementary information useful for enhancing occluded regions of the current view.

The third term of $F_3(u)$ is the regularizer. A regularisation term is necessary for the optimisation function used in this scenario because

- Any system of linear equation is ill-posed if the matrix used is fat, that is, the total number of variables to solve for is less than the total number of equation available in the system. The blurring and sampling matrices used in $F_1(u)$ is of this type. Therefore term $F_1(u)$ in equation 2 is ill-posed and hence requires a regularisation function to solve and minimise the objective function.

- To avoid the ill-posed system of equations to solve the equation 2, the term $F_2(u)$ can help. But the warping matrices of $F_2(u)$ are generated at run-time and cannot be relied upon to fix this issue. Hence a regularisation term is essential for the equation 2.

The regulariser used for this purpose was derived from Graph Signal Processing (40). The graph constructed in this process is represented by an adjacency matrix and used in the term $F_3(u)$ of equation 2. The graph is constructed by considering each pixel of the light field image as a node in the graph. Two nodes are connected by an edge if the pixel represented by the first node of one view is the projection of the pixel represented by the second node in a different view. The entire graph is undirected and the edges are weighted. The value of the edge weight is governed by a mapping function which assigns a numerical value to the weight based on the similarity of intensities between two pixels. The term $F_3(u)$ is designed to penalise significant intensity variations along edges that are highly weighted of the super-resolved light field. And the overall minimisation of the optimisation function leads to selective smoothening of the light field. Therefore, a graph constructed in this way promotes the structure of the light field to be incorporated into the optimisation function and thereby have a significant impact on the super-resolved output.

4.3.1 Algorithm

This section briefly describes the set of steps that were followed to construct the individual components of the optimisation function and how it was solved to get the output. The entire algorithm is the iterative execution of the following four steps. The number of iterations was fixed to a value of 200 as suggested in (39).

- Processing the input
- Construction of Regularisation Graph
- Construction of the Warping Matrix

- Minimising optimisation function

Each of the steps is described in detail below. The steps are essentially derived from the super-resolution policy proposed in (39), however, how each of these steps was implemented is different. Only the conceptual approach, certain universal constants and mapping functions were directly used and the rest of the implementation was modified considering the requirement of this project, the dataset and functionalities present in MATLAB. The parts of the algorithm which directly uses entities from (39) are explicitly mentioned (like the value for the number of iterations specified in the above paragraph).

Input Processing

They Lytro Illum (3) images which were available as part of the dataset were first separated into individual views. Each view was stored as an RGB image with dimensions of $height * width * 3$. This was further loaded into two sets of MATLAB matrix of type unsigned integer. One of the matrices was the input matrix, that was to be used for the construction of warping matrices and regularisation graph. And the other matrix was the output matrix. The super-resolved light field image generated after the optimisation step was stored in the output matrix. At the end of each iteration, the values of the output matrix are copied into the input matrix and the algorithm continues as mentioned above.

Regularisation Matrix Construction

According to (39) the effectiveness of the term $F_3(u)$ depends upon the ability of the graph to capture the underlying structure of the light field. First an adjacency matrix for the light field graph G is created and initialised based on the number of pixels present in the input light field matrix. To create the edges between the pixels and add an edge weight to them, we have to identify the projection of each pixel in other views. Considering the current view V, a set of neighbouring views are determined based on the

indices of the input matrix. And this set of neighbouring views is denoted by V^+. A search window is defined in one of the neighbouring views in V^+ and the current view V, whose dimensions are the same. The centre of the search window in both the views is at equivalent positions with respect to the view and the two search window covers the same amount of pixels within their respective view.

Consider the current view as $u(i) = U_{s,t}(x, y)$ and the neighbouring view as $u(j) = U_{s',t'}(x', y')$ and $u(j) \in V^+$. Here the terms denoted by (s, t) is the parameters uniquely identifiying the light field view and (x, y) denotes the pixel co-ordinates within each of the view. This type of (s, t, x, y) parameterisation is one of the fundamental ways of representing light field (41). This is a modification of the original way of representing light fields using the two-plane parameterisation technique.

Based on the intensity of the pixels available within the two search window, a similarity score is calculated using the following formula.

$$Similarity(i, j) = exp(-\frac{||\rho_{s,t}(x, y) - \rho_{s',t'}(x', y')||_F^2}{\sigma^2}) \tag{3}$$

The calculation defined in equation 3 is directly adopted from the graph-based regularisation technique in (39).

- The term $\rho_{s,t}(x, y)$ represents a square shaped patch within the search window centred at the pixel $U_{s,t}(x, y)$.
- The operator $|| \cdot ||_F$ denotes the Frobenius norm.
- The operator σ is a constant value

Based on the similarity index computed between each pixel in the current view and the neighbouring view, the center pixels of the search window pair whose similarity is high will have a higher similarity index value is added to the graph G as an edge. The edge weight is set to the product of the similarity index value and a constant that is determined based on the overall size of the light field. At the end of this step, the entire graph structure is built.

Warping Matrix Construction

The contribution of the warping matrix is accounted for in the term $F_2(u)$. The warping matrix constructed is responsible for propagating the information and other visual cues available in the neighbouring views to the current view. For this purpose, once again, a set of neighbouring views is defined as R_k^+ to the current view U_k. This is similar to the previous section where a set of neighbouring views were defined, namely, V^+. However, the important difference between V^+ and R_k^+ is that while the former is a collection of all the neighbouring views, the latter is only a subset of V^+. The four views U'_k adjacent to current view U_k in the light field is defined as

$$\left\{ U'_k : k' \in R_k^+ \right\} = \left\{ U_{s,t\pm1}, U_{s\pm1,t} \right\} \tag{4}$$

According to the equation 4, the views considered for warping are only the four views - top, bottom, left and right. But there are in total of eight neighbouring views including the diagonal views in the set V^+. It was essential to consider all the eight views in V^+ because it was being used to record the structure of the light field. And skipping views could lead to errors as the light field structure was used for regularisation whose sole purpose was to adaptively smoothen the errors introduced in the terms $F_1(u)$ and $F_2(u)$ of the optimisation function 2. However, to understand why considering only the four views is sufficient for warping, it is essential to understand the process of determining the warping matrices itself. This is detailed in the subsequent paragraphs.

Consider a pixel $u_k(i)$ in the current view U_k. The warping matrix is determined by a convex combination of those pixels around its projection on the neighbouring view $U'_k = U_{s',t'}$. Although the process of convex combination is straightforward the important step is determining the current pixel's projection in the neighbouring view. But it so happens, this is exactly what we had done in the process of determining the regularisation graph of the light field. Hence, the same equation of 3 is used to determine the projection of the current pixel in the neighbouring view. Once the target pixel is determined, an approximate disparity is computed by considering the similarity score obtained and the

following function.

$$\begin{aligned} S(\delta) = & w_{s,t-1}(x, y+\delta) + w_{s,t-1}(x, y+\delta+1) \\ & + w_{s,t+1}(x, y-\delta-1) + w_{s,t+1}(x, y-\delta) \\ & + w_{s-1,t}(x+\delta, y) + w_{s-1,t}(x+\delta+1, y) \\ & + w_{s+1,t}(x-\delta-1, y) + w_{s+1,t}(x-\delta, y) \end{aligned} \tag{5}$$

To put it simply, each line of the equation 5 is a pair of adjacent pixels in one of the neighbouring views. This equation is directly used from (39). The value provided by this equation is the disparity which is used as an offset to determine the coordinates of the those pixles adjacent to the target pixel of neighbouring light field views.

Considering a warping matrix $F_k^{k'}$ that maps from the pixel $u_k(i)$ of the view $U_{s,t}(x, y)$ to one of the neighbouring view. Each row of the said matrix can be filled by computing the convex combination of the two pixels which are closest to the projection of the pixel $u_k(i)$ on the neighbouring views. As already mentioned the projection of the pixel is determined by the equation 3 and the disparity value available from the equation 5 is used for determining the pixels nearest to the projection. And each of the two pixels in the covex combination contributes a value equivalent to a weight that is directly proportional to its similarity to the target pixel of the original view $U_{s,t}(x, y)$.

Now, that the process of determining the warping matrix has been detailed, it can be easily explained why only the four neighbouring views are considered instead of all the eight views. From the process mentioned above, it is evident that the similarity score calculated by the equation 3 plays a major role. The contribution of each pixel is measured by 3 and it was found experimentally that the absolute value of similarity for the diagonally neighbouring views was insignificant compared to the values for the other neighbouring views. The degree of contribution was very low that it had little to no impact on the overall score but the computational cycles involved in calculating the similarity score for every pixel was very high. Considering this condition, the diagonally neighbouring views were not used for evaluation of the warping matrix.

Minimise Optimisation Function

Now all the individual components, namely $F_1(u)$, $F_2(u)$ and $F_3(u)$ is computed, the final step is to solve the optimisation problem in equation 1. This is a straightforward process where the optimisation equation is converted into a quadratic problem and rewritten in the matrix form with co-efficient and variables. This matrix is solved using the iterative approach detailed at the beginning of this chapter. Essentially, the matrix is broken down further to generate linear equations which is solved using the Conjugate Gradient method.

4.3.2 Final Output

The final output after all the iterations of the graph regularisation step is the spatially super-resolved light field images. While feeding the input, these images were loaded into MATLAB matrices for processing. These matrices were operated upon by the algorithm and the output matrices were generated. A simple data processor is used to render these matrices back as images which can be fed to the light field video generation pipeline.

4.4 Light Field Video Generation

The next step of the implementation was to use the super-resolved light field images as input to the light field video generation pipeline designed by Wang(1). The pipeline was directly used without any modification to its components. Hence, no tangible code modification was performed in this part of the project. However, the major task which consumed a good amount of time was understanding the pipeline and essentially rebuilding the same on a local machine. This was a particularly difficult task because although the conceptual details of the pipeline were discussed in depth and published in the research publication, the practical details and instructions were insufficient in the GitHub repository that was made public as part of the publication.

Rebuilding the pipeline involved the tasks which involved setting up the environment and the components of the pipeline. Ubuntu 18.04 LTS being the latest at the time, was first used as the operating system on which the pipeline was to be rebuilt. This was later discovered to be a wrong choice because of the fact that almost all of the dependencies which were related to the project had been updated several versions. There were significant differences in the way each dependency was designed and there were many new protocols placed in the latest Ubuntu (18.04 LTS) which prevented installation of older versions of the dependencies. Hence this entire endeavour was abandoned and it was decided to rebuild the pipeline on Ubuntu 14.04 which was the version on which the original project was developed. Once the platform was finalised individual components were assembled. And the most important parts of the assembly are discussed here.

- **CNN**

  The project used Caffee framework for defining and executing the two CNNs. There were a number of dependencies which had to be set up in the machine to recreate the exact environment. Since the older version of Ubuntu and Caffe were used, individual dependencies such as CUDA, BLAS, cuDNN, Boost libraries etc had to be manually installed. Ensuring version match and compatibility was a tremendous task. This task was important to ensure the existing CNN models could be modified and built successfully. Although eventually, no modification was actually done to the CNN, setting up the CNN build environment was neccessary since the initial approach of using Hybrid super-resolution required the CNN to be modified and re-built to be used.

- **Matlab**

  MATLAB wrappers were required to ensure the CNNs built could be used with the scripts written to control the pipeline. There were a number of issues with the MATLAB versioning, C++ compiler versioning. All of these were systematically

addressed after a considerable amount of investigation into each of the issue.

- **Light Field Video Viewer**

  As mentioned in the section 3.3, a custom tool was available to highlight the features and playing the light field video. The source code of the tool was analysed to implement new features. Setting up the environment for building this tool after the modification was also a particularly difficult task. The important problem was ensuring version compatibility between a number of different dependencies.

- **Data**

  The means of handling the data input to the system and how to process the output from the pipeline for viewing using the light field video viewer was completely glossed over in the original publication. This was another essential step in rebuilding the pipeline which consumed a lot of man-hours.

4.5 Hybrid Super-Resolution

The original approach to achieve super-resolution was to use Hybrid imaging techniques for light field video, which was later abandoned due to unavailability of the required datasets. But before completely abandoning this approach a workaround was attempted. This section details the implementation approach of the failed task.

4.5.1 Approach

The two main requirement for the hybrid super-resolution were high-resolution video and low-resolution light field images. The dataset published by Wang (1) was supposed to have the high-resolution video but unfortunately did not. The video available was a down-sampled version of the original DSLR video which matched the light field image

resolution of 512 x 352. Hence, to bypass this problem, a straightforward solution is to make whatever video available to be of higher resolution than the light field images. Ultimately, this will fulfil the requirement of the hybrid super-resolution technique (i.e) high-resolution video and low-resolution light field image. Note that the overall resolution of the video is still nowhere close to the original DSLR video but is relatively higher compared to the light field images.

4.5.2 Implementation

To simulate a scenario where the input video has higher spatial resolution than the light field images, the light field images were programmatically downsampled. The resolution was decreased by half and this reflected in a mismatch of spatial resolution dimension between the inputs.

This mismatch of dimension meant that the input cannot be directly fed to the pipeline since the kernel size of the original CNN had to be modified. This task was important to ensure that the change in dimension resulted in kernel mismatch errors when the pipeline was run. And this was done after considerable research as understanding the complex network itself was enduring. After due diligence all the essential modifications were made and the pipeline started executing without any errors.

4.5.3 Expected Results

The expectation from this task was that the generated output video will be spatially super-resolved compared to the original light field images. This meant that the high-resolution information in the video is propagated to the light field. Essentially the output would have a resolution close to the input video which is double the resolution of the light field images. The evaluation strategy was to compare and study this video generated to the original video generated by the unchanged pipeline.

4.5.4 Observed Results

After successfully modifying the CNNs with necessary dimensions, the next step would have been to update the learning parameters so that the new CNNs could train themselves to learn how to use the high-resolution information available in the video frames to super-resolve the light field frames. However, this step could never be practically executed since after the dimensionality update, although the pipeline executed, it never ran to completion and produced an output. The learning process never completed and resulted in out-of-memory errors. Neither executing in a machine with better hardware nor modifying the parameters of the CNNs resolved this issue. Even after spending a considerable amount of time and resource these issues remained unresolved and hence the switch was made to use a different approach (using the Graph-based regularised super-resolution (39)). One potential reason for this issue could be attributed to the fact that, in the original paper, Wang (1) downsampled the DSLR video to match the resolution of the light field. This was done to ensure that the CNN structure could be kept simple and modification of the CNN introduced complexities which was not evident and easily comprehensible. However, note that this observation is only a hypothesis and further research is required to isolate the root cause.

5 Evaluation

In this chapter, the project implementation is subjected to a collection of tests to evaluate its performance. The results obtained in these steps are essential in objectively determining the project's success in achieving the goals stated in Section 3. Each section of this chapter first describes the relevance of each test and the parameters used in the tests. Then the rationale behind the values obtained for each parameter is discussed by presenting arguments and inferring decisions on the results.

Some tests performed are not directly relevant to evaluating the quality of spatial resolution of the image. Such tests were done to ensure that the interpolation and angular resolution achieved by (1) were not compromised while enhancing the spatial resolution. These tests were directly derived from the evaluation methodologies used in (1) and the differences are highlighted in appropriate sections. However, it is worth mentioning that all the tests of (1) are not performed and the reason behind skipping few tests are briefly described.

5.1 Subjective Evaluation

Subjective evaluation is the technique where multiple subjects, usually individual volunteer or focus groups, are asked to rate (a numerical value called opinion score) the images enhanced by the super-resolution algorithm under study. Based on every individual's score the mean opinion score is calculated which acts as the indicator of the overall image quality. Since the ultimate sensors of any visual signal are the human eye, subjective

evaluation provides reliable data for testing super-resolution algorithms (42).

But conducting such experiments prove to be very expensive and time-consuming as well. Also, such evaluation methods cannot directly be incorporated into the process of designing and optimising super-resolution and interpolation algorithms. Due to these practical difficulties, only a nominal amount of subjective evaluation was done in this project. This is recorded in the following paragraph.

The two sets of images represented in figures 5.1 and 5.2 is from the dataset published

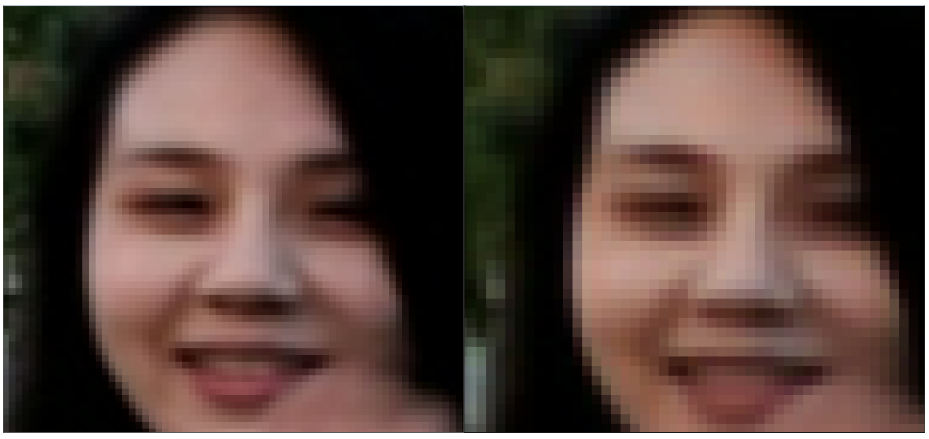

Figure 5.1: Girl Dancing - Superresolved (left), Original (right)

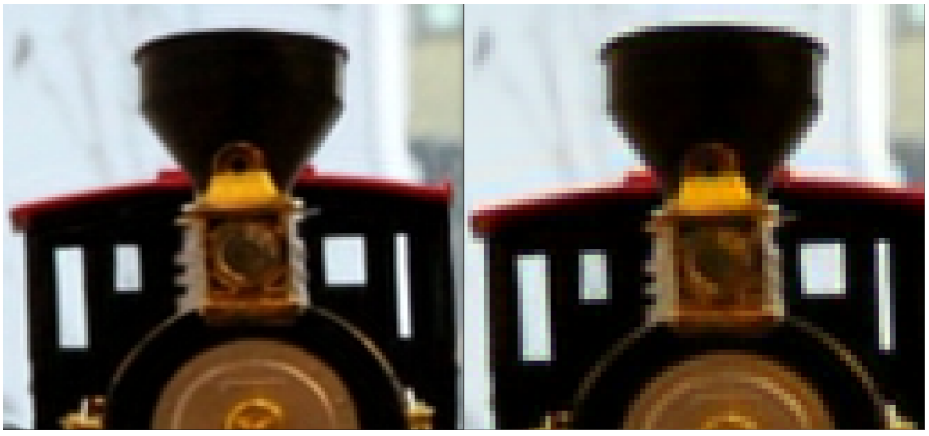

Figure 5.2: Train - Superresolved (left), Original (right)

as part of (1). The results generated after super-resolution were not evident and hence the images were zoomed to a level where individual pixels can be observed to facilitate subjective evaluation. Observing the figures 5.1 and 5.2 it is evident that there is a tangible amount of enhancement in spatial resolution. The borders between sections of different colour are sharper. This can be very easily observed in figure 5.1 where the

sharpness of the border between the girl's hair and face is enhanced. While a certain degree of sharpness is enhanced in figure 5.2, it is not as comparable to the amount in figure 5.1. This could be due to the fact that the edges are more profound in the dancing girl's image compared to the train's image.

5.2 Qualitative Evaluation

Quality of the super-resolution algorithm can be appraised by common image quality assessment methods such as peak-signal-to-noise-ratio (PSNR) and the structural similarity (SSIM) index. This is specifically to evaluate the quality of spatial super-resolution by comparing the frames of super-resolved light field video generated in this project with the frames of original light field video generated by Wang in (1). The results prove that the super-resolved light field video is of higher quality than the original video. Since the original video performed better than the state of the art video interpolation techniques (1) and the scope of this project is only super-resolution, the comparison to other video interpolation techniques is not performed in this project.

5.2.1 PSNR

Historically, the most preferred and suitable metric to evaluate the quality of super-resolved images are Peak Signal-to-Noise Ratio (PSNR) (43). This is the most commonly used measure to determine the difference between the source and super-resolved image and is given by the following formula.

$$PSNR = 10\ log_{10}\frac{MAX^2}{MSE} \tag{1}$$

Here MAX is the maximum possible pixel value and MSE is mean square error. MSE is defined as the sum of the square of intensity differences between a noise-free $\boldsymbol{m\ x\ n}$ image I and its noisy approximation K. The same is given by the following equation.

$$MSE = \frac{1}{m\,n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \tag{2}$$

Table 5.1: The effect of super-resolution compared using PSNR

Method	**LF Super-Resolved**	**LF Original**
PSNR	34.4412	29.22
PSNR Single Frame	21.1717	19.23

The first entry of table 5.1 is computed by determining the average PSNR values across all of the in-between frames and the four corner views as specified in the original paper. The second entry is the PSNR value of all the views of a single frame. Based on both the results the newly generated video performed better compared to the original light field video.

5.2.2 SSIM

Although PSNR can be argued to be a good approximation of the overall image quality, research studies have proved that it fails to correlate well with human perceptual visual quality (44). Hence, a new evaluation metric called Structural Similarity (SSIM) index is considered which is capable of approximating the way the human visual system processes structural information by assigning a numerical value. This numerical value is computed by using the variance and mean pixel intensity between the two images under comparison and the covariance between the images respectively. SSIM approximates the contrast, luminance and the structure of the image. The mean structural similarity is obtained by computing the local similarity of small windows across the images and averaging them (43).

The table 5.2 lists the SSIM value index for the original and super-resolved light field video. The way the input images were collected for evaluation is same as the methods used for determining PSNR in Section 5.2.1. In the results published in (1) the SSIM values were already very close to a perfect similarity value of 1. And the images compared

Table 5.2: SSIM between the original and super-resolved light field frame

Method	**LF Super-Resolved**
SSIM	0.9237
SSIM Single Frame	0.9754

between the frames of the video with the ground truth. In this tests, the SSIM value is compared between the original video and the super-resolved one. From the results captured in Table 5.2, we can infer that the super-resolution preserved the high similarity index and no substantial loss was introduced.

5.3 No-Reference Evaluation

Computing PSNR or SSIM requires the original high-resolution reference image to determine the extent to which the super-resolved image corresponds with it. And the fundamental problem is that a perfect quality high-resolution image is never available for such an comparison. Although human observers can easily spot the degradation or enhancement without the reference image, it is a subjective process and cannot be integrated into the super-resolution pipeline. Therefore estimation of image quality using an assessment that co-relates well with human perception but without the use of any reference image is necessary. Therefore this evaluation is also done in this project using Matlab's inbuilt models.

For this purpose, a No-Reference Quality Assessment Model provided by Matlab is used. The model itself is called the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) and uses a training based approach to compute a quality score. The training is done using predictable statistical features called Natural Scene Statistics (NSS). The BRISQUE model uses images along with subjective quality scores for training and thus has the highest co-relation with human perception of quality compared to other models provided by Matlab (45). The default BRISQUE model is used for evaluation. But the most suitable model for this project would be a custom model trained by a collection of

light field images each with a subjective evaluation image quality assessment score. Since such a dataset is not available, the default model is used. The BRISQUE score obtained by using this model is captured in Table 5.3.

According to MATLAB documentation the image that has the better perceptual visual

Table 5.3: The effect of super-resolution compared using BRISQUE

Method	**LF Super-Resolved**	**LF Original**
BRISQUE	20.6586	24.5643
BRISQUE Single Frame	23.1123	22.6723

quality will have a lower BRISQUE score (46). And from Table 5.3 it is evident that the super-resolved video frames have a better perceptual visual quality than the original light field video frames. Also comparing to the data observed in Table 5.2 which also evaluates the overall perceptual quality, the BRISQUE score value follows a similar pattern.

5.4 Depth From Video

In this test, the video is analysed to determine how accurate is the depth information of objects in the scene is captured. Usually, light field images with better angular resolution will have rich depth information as well. But this project is focused on spatial resolution and this test is performed only to ensure that the spatial super-resolution did not compromise the angular resolution of the original light field video. This test is performed in (1) and Wang has reported the results compared to a state of the art depth estimation method called depthTransfer. However, in this project, the frames of super-resolved video is compared with the frames of the original light field video and analysed for any loss of depth information. This is done by extracting the middle view of both the light field video and comparing it with the ground truth image. The data used for this purpose is part of the dataset used for training the CNNs in (1) and made available publicly. The middle view of the generated light field video is compared because its orientation is the closest to the orientation of the ground truth image captured by the light field camera.

Figure 5.3: Original

Figure 5.4: SuperResolved

The figures 5.3 is the inverted result of the absolute difference between the frame generated by the original light field video and the ground truth. On the other hand, the figure 5.4 is the absolute difference between the ground truth (figure 5.6) and super-resolved light field video and the ground truth (figure 5.6). We can observe from the image that both of the differences look quite similar. This means that the super-resolution has not compromised any of the information generated during interpolation. The same argument can be extended to other views of the light field to prove that the views generated by the super-resolved pipeline are consistent with the views generated by the original pipeline. This implies that the depth information that can be acquired from the light field is not affected in any manner.

To further strengthen the argument on the similarity between the figures 5.3 and 5.4,

Figure 5.5: SSIM between Fig 5.3 and Fig 5.4

Figure 5.6: GroundTruth

the structural similarity index (SSIM) (Described in 5.2.2) is calculated and is shown in the figure 5.5 and the value itself was **0.837**, which indicates very high similarity between the images.

5.5 Quantitative Evaluation

Wang performs the quantitative evaluation in (1) to measure the performance of interpolating the light field and 2D video frames into a light field video. Hence, the system is compared with different video interpolation techniques by varying the quantity of the input. The quality of the video is determined by calculating the PSNR and SSIM values of individual frames. But this has already been evaluated in detail in previous sections. Therefore another parameter valid for quantitative evaluation - the overall time taken to generate the super-resolved interpolated video is considered in this section. The time taken for each dataset is described in the table 5.4. It can be easily observed that the

Table 5.4: Time taken for each datasets.

Dataset	**Time (Minutes)**
Train	195
Cat	240
Girl	234
Cat on Train	292

time taken is in order of several hours. And based on analysis of the total time taken it was detected that over 97% of the time is incurred by the GB regularised spatial super-resolution that has been integrated into the light field video generation pipeline. The main reason is due to the optimisation process which bypasses the requirement of high-resolution light field images as input for the super-resolution process. Every iteration of computation of the inter-view graph and warping matrices takes considerable processing time. This is reported in the original publication as well (39). In addition, this process had to be applied for each light field frame produced by the camera at 3 frames per second. This exponentially increased the overall time. Since the goal of the project was

only achieving super-resolution of light field video, the optimisation of the overall time taken is not carried out.

6 Conclusion

The final product of this thesis is a new pipeline which is capable of producing the light field video of good spatial resolution. And as shown in the section 5 the overall quality of the video is enhanced without any compromise to the other features of light field video. The spatial super-resolution has been enhanced while preserving the integrity of the light field video such as angular resolution and other post-capture processing effects. The principle demonstrated in this dissertation fixes one of the major problems in the development of light-weight, hand-held and low-cost light field video cameras so that it can be used on an airborne platform. Many other issues have been identified in the development of such a camera and documented as part of this dissertation. This provides scope for a lot of research that can be performed. In fact, there was very limited work being done one light field videos. One of the major reason being the poor overall image quality due to the spatial-angular resolution tradeoff resulting in a meagre return on investments. This dissertation by successfully enhancing video quality using super-resolution tries to restore the commercial viability of light field video research. Light field video has many applications such as dynamically refocusing during video runtime, tracking the focus of an object throughout the scene etc. Although these features are not developed in this research lack of spatial resolution was rendering these features effectively useless. By enabling usage of such features, this dissertation extends the many benefits of current light field technology. Detailed comparative study of various spatial super-resolution algorithm has been performed. This study has been included in this report which can serve as a survey material for future works.

6.1 Future Work

Almost all of the issues that can be addressed in Light Field videography that is documented in this dissertation can be explored as a future work.

6.1.1 Real-Time Light Field Video

Another important aspect that can be improved relates to the Quantitative evaluation mentioned in the section 5.5. From the time taken to generate each video, it is evident that the entire system takes in the order of several hours. But many systems can benefit from the video generation being real-time in order to use the features of light field video effectively. For instance, computer systems on driverless cars have only a few milliseconds to interpret the current situation from the data produced by the camera and make a driving decision. Therefore, real-time light field video generation can be pursued as a potential future work.

6.1.2 Evaluation with other Spatial-SuperResolution

The evaluation results that have been recorded in section 5 of this dissertation only proves that the new video generated has certainly better spatial-resolution than the original light field video generated by Wang (1). But it is worth highlighting that the work by Wang (1) never intended to achieve good spatial resolution. Hence the evaluation results do not prove that the method proposed in this work is the best way to spatially super-resolve the light field video. Although a detailed qualitative analysis was done to compare different spatial light field super-resolution techniques before finalising on using Graph-Based Regularisation (39), a more thorough study would involve integrating other techniques into the pipeline and evaluate the different light field video produced against each other and the ground truth.

Bibliography

[1] Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A Efros, and Ravi Ramamoorthi. Light field video capture using a learning-based hybrid imaging system. *ACM Transactions on Graphics (TOG)*, 36(4):133, 2017.

[2] Abhilash Sunder Raj, Michael Lowney, and Raj Shah. Light-field database creation and depth estimation. *Dept. Computer Science, Stanford University, Technical Report*, 2016.

[3] Lytro Inc. *Lytro Cameras*, (accessed June 3, 2018). URL `https://support.lytro.com/hc/en-us`.

[4] M. Schubin. More, faster, higher, wider: A brief history of increases in perceptible characteristics of motion imaging. *SMPTE Motion Imaging Journal*, 125(6):32–40, Aug 2016. ISSN 1545-0279. doi: 10.5594/JMI.2016.2579138.

[5] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 31–42, New York, NY, USA, 1996. ACM. ISBN 0-89791-746-4. doi: 10.1145/237170.237199. URL `http://doi.acm.org/10.1145/237170.237199`.

[6] James F. Blinn and Martin E. Newell. Texture and reflection in computer generated images. *Commun. ACM*, 19(10):542–547, October 1976. ISSN 0001-0782. doi: 10.1145/360349.360353. URL `http://doi.acm.org/10.1145/360349.360353`.

[7] N. Greene. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications*, 6(11):21–29, Nov 1986. ISSN 0272-1716. doi: 10.1109/MCG.1986.276658.

[8] M Shahzeb Khan Gul and Bahadir K Gunturk. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Transactions on Image Processing*, 27(5):2146–2159, 2018.

[9] G Lippmann. Epreuves reversibles, photographies integrales. *J. Academie des sciences*, 2(11):446–451, 1908.

[10] Gershun A. The light field. *Journal of Mathematics and Physics*, 18(1-4):51–151. doi: 10.1002/sapm193918151. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm193918151.

[11] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 43–54, New York, NY, USA, 1996. ACM. ISBN 0-89791-746-4. doi: 10.1145/237170.237200. URL http://doi.acm.org/10.1145/237170.237200.

[12] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005.

[13] Raytrix gmbh. *Raytrix Cameras*, (accessed June 3, 2018). URL https://www.raytrix.de.

[14] P&S Market Research. Aerial imagery market by type (low oblique, vertical, high oblique), by application (surveillance & monitoring, geospatial, energy & resource management, conservation & research, exhibition & live entertainment, disaster management, construction & development), by industry (government, energy & mining, defense, agriculture & forestry, civil engineering & archaeology, media & entertainment), by geography (u.s., canada, u.k., germany, france, russia, china, japan, india, u.a.e., south africa, mexico, brazil) - global market size, share, development, growth and demand forecast, 2013-2023. https://www.researchandmarkets.com/research/fppwgm, 2017.

[15] FlightGlobal. *Easyjet Drone Insepction*, (accessed June 8, 2018). URL https://www.flightglobal.com/news/articles/easyjet-to-roll-out-drone-inspections-from-2018-441652/.

[16] Michael Dominick Yocius. Development of airborne light field photography, 2015.

[17] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, March 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.147.

[18] Ren Ng. Fourier slice photography. *ACM Trans. Graph.*, 24(3):735–744, July 2005. ISSN 0730-0301. doi: 10.1145/1073204.1073256. URL http://doi.acm.org/10.1145/1073204.1073256.

[19] Canon Inc. *EOS REBEL T1i/EOS 500D Instruction Manual*, (accessed June 20, 2018). URL http://gdlp01.c-wss.com/gds/7/0300002157/01/eosrti-eos500d-im-en.pdf.

[20] Hannes Sardemann and Hans-Gerd Maas. On the accuracy potential of focused plenoptic camera range determination in long distance operation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:1 – 9, 2016. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2016.01.012. URL `http://www.sciencedirect.com/science/article/pii/S0924271616000277`.

[21] Rakesh Kumar, Harpreet Sawhney, Supun Samarasekera, Steve Hsu, Hai Tao, Yanlin Guo, Keith Hanna, Arthur Pope, Richard Wildes, David Hirvonen, et al. Aerial video surveillance and exploitation. *Proceedings of the IEEE*, 89(10):1518–1539, 2001.

[22] CC Whitworth, AWG Duller, DI Jones, and GK Earp. Aerial video inspection of overhead power lines. *Power Engineering Journal*, 15(1):25–32, 2001.

[23] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng Han. Vehicle detection and tracking in wide field-of-view aerial video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 679–684. IEEE, 2010.

[24] Alejandro Angel, Mark Hickman, Dinesh Chandnani, and Pitu Mirchandani. Application of aerial video for traffic flow monitoring and management. In *Applications of Advanced Technologies in Transportation (2002)*, pages 346–353. 2002.

[25] Bennett S Wilburn, Michal Smulski, Hsiao-Heng Kelin Lee, and Mark A Horowitz. Light field video camera. In *Media Processors 2002*, volume 4674, pages 29–37. International Society for Optics and Photonics, 2001.

[26] ViewPLUS Inc. *Datasheet of ProFUSION 25c 5x5 camera array system*, (accessed June 5, 2018). URL `https://www.ptgrey.com/Content/Images/uploaded/KB-Data/ProFUSION_25_datasheet.pdf`.

[27] P. Merkle, A. Smolic, K. Muller, and T. Wiegand. Multi-view video plus depth representation and coding. In *2007 IEEE International Conference on Image Processing*, volume 1, pages I – 201–I – 204, Sept 2007. doi: 10.1109/ICIP.2007.4378926.

[28] Brandon M Smith, Li Zhang, Hailin Jin, and Aseem Agarwala. Light field video stabilization. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 341–348. IEEE, 2009.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

[30] Penn State College of Earth and Mineral Science. *Geospatial Applications of Unmanned Aerial Systems*, (accessed June 5, 2018). URL `https://www.e-education.psu.edu/geog892/node/549`.

[31] GDU Technology. *Byrd Advanced Drone, capable of carrying additional camera*, (accessed June 5, 2018). URL `https://store.gdu-tech.com/products/byrd-advanced`.

[32] GoPro. *GoPro Cameras*, (accessed June 5, 2018). URL `https://shop.gopro.com/EMEA/cameras/`.

[33] Peta Pixel. *Review of Lytro Illum*, 2016 (accessed May 25, 2018). URL `https://petapixel.com/2017/01/12/look-lytro-illum-camera-future-failed/`.

[34] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence*, 14(2):99–106, 1992.

[35] Todor Georgiev, Ke Colin Zheng, Brian Curless, David Salesin, Shree K Nayar, and Chintan Intwala. Spatio-angular resolution tradeoffs in integral photography. *Rendering Techniques*, 2006:263–272, 2006.

[36] T. E. Bishop, S. Zanetti, and P. Favaro. Light field superresolution. In *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9, April 2009. doi: 10.1109/ICCPHOT.2009.5559010.

[37] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, pages 1–10. IEEE, 2014.

[38] K. Mitra and A. Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 22–28, June 2012. doi: 10.1109/CVPRW.2012.6239346.

[39] Mattia Rossi and Pascal Frossard. Light field super-resolution via graph-based regularization. *CoRR*, abs/1701.02141, 2017. URL `http://arxiv.org/abs/1701.02141`.

[40] A. Elmoataz, O. Lezoray, and S. Bougleux. Nonlocal discrete regularization on weighted graphs: A framework for image and manifold processing. *IEEE Transactions on Image Processing*, 17(7):1047–1060, July 2008. ISSN 1057-7149. doi: 10.1109/TIP.2008.924284.

[41] Aaron Isaksen, Leonard McMillan, and Steven J. Gortler. Dynamically reparameterized light fields. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 297–306, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. ISBN 1-58113-208-5. doi: 10.1145/344779.344929. URL `http://dx.doi.org/10.1145/344779.344929`.

[42] H. Yeganeh, M. Rostami, and Z. Wang. Objective quality assessment for image super-resolution: A natural scene statistics approach. In *2012 19th IEEE International Conference on Image Processing*, pages 1481–1484, Sept 2012. doi: 10.1109/ICIP.2012.6467151.

[43] H. M. Keshk, M. M. Abdel-Aziem, A. S. Ali, and M. A. Assal. Performance evaluation of quality measurement for super-resolution satellite images. In *2014 Science and Information Conference*, pages 364–371, Aug 2014. doi: 10.1109/SAI.2014.6918212.

[44] Z. Wang, A. C. Bovik, and L. Lu. Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313–IV–3316, May 2002. doi: 10.1109/ICASSP.2002.5745362.

[45] MATLAB. *Matlab No-Reference Quality Assessment Model*, (accessed July 22, 2018). URL `https://uk.mathworks.com/help/images/train-and-use-a-no-reference-quality-assessment-model.html`.

[46] MATLAB. *Matlab BRISQUE Quality Assessment Model*, (accessed August 8, 2018). URL `https://uk.mathworks.com/help/images/ref/brisque.html`.