AMME4111 Thesis A Progress Report

Joseph Daniel

October 2019

Contents

Chapter 1

Introduction

Humans have a remarkable ability to perceive visual stimuli emanating from the world around them. Not only do we identify different objects, scene depth, movement, and colour with ease, but we often draw meaning and even enjoyment from the combination of light rays bouncing around space and arriving at our eyes. We are able to automatically compensate for differences in lighting and viewing angle, and we can extract complex information such as emotion from what we process visually. To word it concisely, the human visual perception system is one of the most versatile and robust image processing machines that we know of. Most man-made machines on the other hand are much more easily decomposed into a set of deterministic modules that require explicit, well defined instruction sets. Often, when given an image, there is no straightforward function that directly maps from pixel intensities to the semantic meaning of that image. As researchers tackling problems in computer vision, we are usually interested in developing the algorithms that can help machines not only sense the world, but perceive it in a way that is useful.The objective of this thesis project is to develop a pipeline that simultaneously learns to perform two tasks. The first is visual odometry - the act of estimating camera motion from a sequence of images. The growth of autonomous robotics has meant that there is an every-growing need for accurate navigation and localisation when satellite positioning is unavailable or wheel encoder odometry is unreliable. An autonomous unmanned aerial vehicle navigating an unknown indoor environment is one example where visual odometry is a valuable feedback signal that can be used to precisely control the UAV's position. The second is depth estimation, that is, discerning the distance of objects in a scene from the imaging device. The ability to perform depth estimation from visual cues is desirable in applications such as autonomous vehicles where navigating new environments whilst avoiding collisions is a basic requirement. Being able to draw depth information from a camera sensor will help to enable these capabilities, especially in applications where methods such as Lidar or RGBD sensors are prohibitive - small quad-copters and internal medical imaging devices are examples where this might be the case.

Humans are exceptionally well adapted to these tasks - our two eyes allow us to process the 3D geometry of the scene, while our learned experiences are often able to fill in the gaps where geometric information is insufficient or unavailable. Unfortunately, image processing algorithms are not equipped with this same kind of human intuition, and so the deceptively complex task of estimating the 3D structure of a scene from a sequence of images continues to attract attention from the computer vision community.

The price of camera components is decreasing while image quality continues to improve, not only making cameras an attractive perception module for autonomous robotics, but also spurring the popularity of multiple-view imaging. Embracing this idea, this thesis project seeks to develop a pipeline that performs these tasks by taking advantage of the rich geometric information exposed when using multiple views. Thus, this thesis project adopts the vocabulary and the concepts from the area of light field imaging. The light field describes the intensities and directions of each light ray in space, providing a convenient way of thinking about multi-view geometry which is embraced throughout this thesis.

Another core objective of this thesis is to develop a pipeline that relies principally on unsupervised machine learning models to perform the two tasks side-by-side. This approach is chosen with the intention that a data-driven model can build a robustness to scenarios that are difficult to design for using hand crafted solutions. Furthermore, this pipeline will benefit from being self-supervised - unlike data-driven models that require direct supervision from a human to improve their performance (classifying images of cats and dogs is a common example), this pipeline has no intermediary requirements between the data collection and training routines. Not only does this mean that the expensive process of hand-labelling data is unnecessary for this pipeline to function, but the model will be able to continue training, benefiting from new data that it receives *whilst operational in the field*. An adaptive perception module that learns to operate effectively in new environments is a desirable capability in robotics, especially when considering that effects such as thermal expansion, vibration and shocks can frequently render our equipment calibrations meaningless.

In this progress report we will first establish the conceptual groundwork that this thesis builds upon with a discussion surrounding the seminal texts that have shaped the state-of-the-art literature in visual odometry and depth estimation. In understanding those texts, we demonstrate that existing methods have yielded excellent results, whilst identifying that there is a substantial potential for improvement on state-of-the-art results using multi-view imaging devices.

Chapter 2

Literature Survey

2.1 The Light Field and Multiple View Geometry

Light field imaging has emerged as a powerful tool in computer vision for robotics, offering a rich higher-dimensional representation than what can be captured by conventional optics. The underlying principle used to describe the light field is the plenoptic function, a 7-dimensional mapping that assigns radiance values to the light rays at every position in space, in every orientation, at all wavelengths, throughout all of time [11]. [1] shows that this can be formally expressed as $L(x, y, z, \theta, \phi, \lambda, t)$, and measured in $W/m^2/sr/nm/s$. With the addition of practical constraints however, the plenoptic function can be expressed more concisely as a parameterisation of 4 variables.

Pixels on camera sensors integrate the number of photons arriving at them over a finite period of time removing the temporal dimension, and each colour channel can be thought of as a monochromatic sampling of the light field, removing the spectral dimension. Additionally, and importantly, is the constraint that the radiance of light rays propagating through a vacuum do not change if samples are restricted to the convex hull of the scene, thus reducing the overall dimensionality of the plenoptic function by one parameter [13]. To illustrate this, one could think of the light rays leaving the inside of a bowl sat upright on a table. Many of those rays may only travel a small distance before being blocked by the inside of the bowl itself, meaning those rays will never be registered by any practical measurement device. As shown in figure 2.1 if we consider only the convex hull of the bowl however, we are required only to represent the value of the plenoptic function on the encapsulating surface of the object [9].

Also illustrated is a common convention for describing light rays in this 4 dimensional space called the two plane parameterisation. In this parameterisation, two parallel planes are used to fix both the position and orientation of each ray by fixing their points of intersection with two parallel

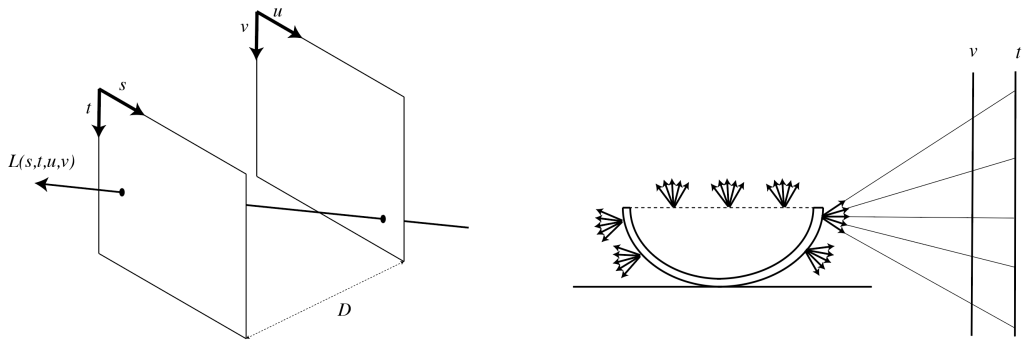

Figure 2.1: Two plane parameterisation (left): the plenoptic function can be described by the radiance along a ray passing through two parallel planes. The free space assumption (right): if we consider only the bundle of rays leaving from the convex hull of the object at a particular instance in time, in a single colour channel, we can parameterise the light rays as a function of 4 variables rather than 7.

planes. By convention, the plane closest to the scene it termed *u, v* and the plane closest to the camera sensor is the *s, t* plane.

This 4D realisation of the light field originated as a model of rendering 3D computer graphics, one which shifted emphasis from notions of texture and geometric primitives to modelling the behaviour of light rays permeating space. Since then however, the conceptual framework of the light field has drawn a following of researchers at an intersecting region of signal processing, computer vision and robotics [3]. This notion of light field imaging finds a foothold in this thesis project through the utilisation of camera arrays, which are devices that sample multiple views of the same scene. Using a camera array is a simple method for acquiring a dense sample of the light field, where the position of the camera determines *(s,t)* while the location of the pixel determines *(u, v)* [22]. The images captured from a camera array are mapped easily to the 4D light field, and identifying corresponding pixels across images exposes a rich tapestry of geometric information about the scene.

One way that this geometric information can be easily visualised is by taking slices of the light field image in the *s,u* or *t, v* axes as shown in figure 2.2. While the idea of taking a 2D slice from the 4D image can seem complex, the task of generating a so called 'epipolar plane image' from a camera array is deceptively simple. Images captured from camera arrays can be stacked to form a solid volume, from which 2D slices can be sampled. Each of these slices yields an image characterised by sheared straight lines, encoding information about the geometry of the scene, including depth and occlusions [2].

The geometric information encoded in a light field sampling can alternatively be visualised by processing the image into a 'focal stack'. Focal stacks closely resemble images with shallow depth of field such as those that can be captured from a commercial DSLR camera. Light field focal stacks

Figure 2.2: Epipolar Plane Images (left): Shown as a slice of a volume, the images formed by dissecting the image in the s,u and t,v planes are characterised by sheared straight lines, with the grade of the slope encoding the amount of parallax experienced by a pixel at that u or v coordinate. Synthetic aperture focusing (right): taking the average of every image from the camera array yields an image where different parts are in focus depending on the alignments of the images.

differ from focus in the optical sense however in that they are synthetic and can be recomputed after the image has been taken, effectively allowing control over the depth of field and focal depth in post-processing. Focal stacks can be computed from camera array images by layering images over one another and taking the average value for each pixel. The result is that parts of the scene that closely overlap appear in focus while areas with poor overlap create a 'bokeh' effect. More formally, if the relative pose of each camera is known, a specific focal stack for any desired depth can be computed by projecting each image onto the desired focal plane, and computing their average [21].

These representations of the light field will play an important role in this thesis project as we experiment with different methods for feeding light field images to the machine learning pipeline. An important consideration in any machine learning algorithm is the feature space - based on what particular inputs will the algorithm be making its decision? Raw images contain millions of measurements and thus represent an incredibly high-dimensional feature space for neural networks to process. Light field images are several times larger, and thus it is important that some form of dimensionality reduction is used to ease the training process. With the goal of investigating effective methods of feeding light fields to neural networks, this thesis will explore the use of three different light field formats as the entry point to the machine learning pipeline. The first two will be the focal stack, and epipolar plane image described above, interpreting the images as a 3 dimensional volume created by stacking 2 dimensional images on top of one another. The third will treat the light field as a 4 dimensional volume, requiring a 3D signal processing pipeline to fully take advantage of the dimensionality.

2.2 Machine Learning for Image Processing

An oft-quoted anecdote in the computer vision community tells of MIT researcher Seymour Papert, who in 1966 assigned a summer project that sounded simple enough, namely to construct a 'visual system' that could describe what objects it saw by name [17]. While the regimes of computer vision have evolved substantially since 1966, many of the ideas, and challenges have persisted. This is embodied in the popularity of projects such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [20], drawing researchers from institutions around the world. It was at the ILSVRC annual challenge where in 2012, a convolutional neural network achieved a top-5 error rate of 15.3%, outperforming all previous submissions by 10.8% [12].

While deep neural networks for computer vision have gained massive popularity since the success of 'AlexNet', the history of neural architecture models begins much earlier, with the perceptron as described by Frank Rosenblatt in [18]. The fundamental building block of neural networks, the 'perceptron' is a module that accepts several inputs, and produces a single output computed as a weighted sum of each input - allowing complex functions to be approximated when several perceptrons are layered together as a 'multi-layer perceptron' [15]. The process of finding the optimal set of weights that produce the desired output given a set of inputs is referred to as training, and in practice is usually found by optimising some cost function using the backpropagation algorithm [19]. The result is that multi-layer perceptrons are often able to approximate complex functions, which is incredibly useful in both regression and classification tasks.

Convolutional neural networks (CNN's) are similar to multilayer perceptrons, but introduce a spatial awareness that makes them particularly well suited for extracting high-level features from image data.

2.3 Depth Estimation and Visual Odometry

Since the very beginning of photography, cameras have captured 2D representations of a 3D world, meaning information about physical structure is lost. With the addition of a second viewpoint however we can learn a little more about the shape of the scene as governed by epipolar geometry [10]. Meanwhile, a video is a sequence of images, taken in very quick succession, and so one could think of video footage as a multi-view camera where each image is separated not only spatially but also temporally. Identifying the amount of motion between two temporal frames of a camera is the goal of visual odometry, but doing so monocularly is problematic because with a single camera there is no way of concretely discerning the actual magnitude of the movement based on pixel data alone,meaning some kind of scale factor needs to be estimated based on characteristics of the image [5, 16, 23, 24]. In fact, this scale ambiguity is often exploited by film makers - what appears as a sweeping shot of a vast landscape on the big screen is often modeled as a miniature film set in the

studio. Because the image is monocular, there is no way to ground our measurements of scale in real world units, and so we resort to our imaginations and learned experiences to fill in the gaps. What *is* preserved in these monocular setups however is the overall structure of the scene and motion of the camera - we may not know how large the object is or how far the camera has moved, but we *can* compute the shape of the object as well as the direction of camera motion.

One intuitive approach to performing visual odometry is to observe the direction of movement of each pixel between the two images. However one would find that unless every point in the scene laid on a single plane parallel to the camera sensor, each pixel would move a different amount according to its distance to the camera. Not only does the relative motion of each pixel depend on the the distance travelled by the light ray from the object to the camera, but the colour and intensity of those pixels may change drastically if the scene contains shiny objects. These challenges in visual odometry mean that the task of simultaneous depth estimation and pose prediction have continued to attract attention from the computer vision and robotics communities.

One recent approach that has driven a large body of research is the use machine learning to perform both of these tasks, utilising convolutional neural networks to learn a non-linear mapping directly from a pair of images to their depth maps, as well as the relative pose between them [4, 6, 8, 14, 25]. Combining the spatial awareness of the convolutional down sampling operation with a neural networks ability to learn accurate approximations for complex, non linear functions, these approaches have found success in both supervised [14, 4], and unsupervised settings [6, 8, 25]. In the supervised family of algorithms, [4] and [14] take advantage of datasets such as KITTI [7], containing ground truth depth maps collected using state-of-the-art depth sensors and poses measured from inertial sensors.

Unsupervised experiments such as [6, 8, 25] on the other hand exploit the constraints imposed by epipolar geometry to learn depth either by using known camera poses or by estimating pose in addition to depth. An important consideration in optimising a model with a neural architecture, is the differentiability of each operation that goes into producing the outputs. Any operation that is non differentiable would not support the backpropagation of errors through the network, thus disallowing optimisation algorithms which depend on partial derivatives.

Chapter 3

Progress Report

Chapter 4

Updated Research Proposal

Bibliography

[1] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

[2] Robert C. Bolles and H. Harlyn Baker. Readings in computer vision: Issues, problems, principles, and paradigms. chapter Epipolar-plane Image Analysis: A Technique for Analyzing Motion Sequences, pages 26–36. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[3] Donald Dansereau. *Plenoptic Signal Processing for Robust Vision in Field Robotics*. PhD thesis, Australian Centre for Field Robotics, 2014.

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014.

[5] Paul Verlaine Gakne and Kyle O'Keefe. Tackling the scale factor issue in a monocular visual odometry using a 3D city model. 2018.

[6] Ravi Garg, Vijay Kumar B. G, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. *CoRR*, abs/1603.04992, 2016.

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016.

[9] 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.

[10] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[11] Ivo Ihrke, John Restrepo, and Loïs Mignard-Debise. Principles of light field imaging: Briefly revisiting 25 years of research. *IEEE Signal Processing Magazine*, 33:59–69, 09 2016.

[12] 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.

[13] Mark Levoy and Pat Hanrahan. Light field rendering. In *Proc. ACM SIGGRAPH*, 1995.

[14] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *CoRR*, abs/1502.07411, 2015.

[15] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969.

[16] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–652–I–659 Vol.1, Los Alamitos, CA, USA, jul 2004. IEEE Computer Society.

[17] Seymour Papert. The summer vision project. 10 1966.

[18] Frank Rosenblatt. *Principles of Neurodynamics*. Spartan Books, 1959.

[19] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[21] Vaibhav Vaish, B. Wilburn, N. Joshi, and M. Levoy. Using plane + parallax for calibrating dense camera arrays. volume 1, pages I–2, 01 2004.

[22] Li Yao, Yunjian Liu, and Weixin Xu. Real-time virtual view synthesis using light field. *EURASIP Journal on Image and Video Processing*, 2016(1):25, Sep 2016.

[23] Dingfu Zhou, Y. Dai, and Hongdong Li. Reliable scale estimation and correction for monocular visual odometry. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 490–495, June 2016.

[24] Dingfu Zhou, Yuchao Dai, and Hongdong Li. Ground plane based absolute scale estimation for monocular visual odometry. *CoRR*, abs/1903.00912, 2019.

[25] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.