Proceedings of Seminar and Project

Survey of Light Fields and Depth Estimation Methods

SEMESTER

Oliver Wasenmüller and Prof. Didier Stricker
Department Augmented Vision
University of Kaiserslautern and DFKI GmbH
Introduction

The seminar and project TITLE (INF-XX-XX-S-X, INF-XX-XX-L-X) are continuative courses based on and applying the knowledge taught in the lectures 3D Computer Vision (INF-73-51-V-7) and Computer Vision: Object and People Tracking (INF-73-52-V-7). The goal of the project is to research, design, implement and evaluate algorithms and methods for tackling computer vision problems. The seminar is more theoretical. Its educational objective is to train the ability to become acquainted with a specific research topic, review scientific articles and give a comprehensive presentation supported by media.

In the XXX semester XXX, XXX projects addressing XXX were developed. Moreover, XXX seminar works addressed XXX. The results are documented in these proceedings.

Organisers and supervisors

The courses are organised by the Department Augmented Vision (http://ags.cs.uni-kl.de), more specifically by:

Oliver Wasenmüller
Prof. Dr. Didier Stricker

In the XXX semester XXX, the projects were supervised by the following department members:

NAME

MONTH YEAR
Survey of Light Fields and Depth Estimation Methods

Jonathan Wray\textsuperscript{1} and Yuriy Anisimov\textsuperscript{2}

\textsuperscript{1} WrayJonathanM@sau.edu
\textsuperscript{2} Yuriy.Anisimov@dfki.de

Abstract. This paper aims to introduce light fields, plenoptic cameras, 4-D light fields and depth estimation. This paper also evaluates what is being done in the state-of-the-art and how these methods are able to be implemented to solve problems in the industry.

Keywords: light fields, plenoptic cameras, depth estimation, epipolar plane images.

Fig. 1. LEFT: RAW Lightfield image shown on the left, showing how micro lenses capture light information. Image post-processing on the right. [1]. RIGHT: Depth Map Example [2]

1 Introduction

Light fields are on the rise as a potential disruptive technology in the world of visual computing, graphics, and digital imaging. Computer hardware has advanced enough that it's able to handle more information. In 2006, a thesis on Digital Light Field Photography, author Ng stated that digital cameras could have photosensor resolutions already of 100 megapixels, and industry cameras could already reach over 250 megapixels. [3] Utilizing this sensor technology, light fields are aiming towards shedding limitations and problems presented by standard digital visual technologies. This paper’s objective is to give the reader an overview of light fields, light field cameras, and current state-of-the-art depth estimation methods.

2 Survey of Light Fields and Plenoptic Cameras

2.1 Light Fields (The Plenoptic Function) During the Italian Renaissance, the study of optics was taking on radical development and the painter and inventor Leonardo da Vinci supported the pyramidal theory arguments expressed by Alhazen, Witelo, and Pecham, which introduced that the primary mechanism of vision was the reception (intromission) of light rays from objects into the eye (Kemp, 2006, p.114) [4]. Da Vinci become invested in the development of optics and did drawings representing the light field, which he claims was an infinite number of radiant pyramids (or light rays) extending in all directions.

Light visible to the human eyes is identified as the visual spectrum. The visible light to the human eye lies between frequencies ranging from 390 to 700 nm. Violet is the lowest wavelength
the human eye can detect, existing in the spectrum between 380 - 450 nm and Red is the highest wavelength, existing at 620 to 750 nm. With the application of studying surface illumination in mind, Arun Gershun [5] defined the light field concept, claiming it as “the amount of light traveling in every direction through every point in space”. In his 1936 paper, Gershun wrote that the amount of light arriving at points in space typically varies smoothly from place to place and could therefore be characterized using mathematics like calculus and analytic geometry. A method for measurement at the time was impossible however, so instead he attempted to derive the illumination patterns observed on surfaces due to light sources using closed form. Using modern technology, we can now record, manipulate, and display Gershuns light field [6].

A light field is defined as the collective light rays (or photons) in space that can be parameterized by three coordinates, x, y, and z and two angles and . The radiance along all such rays in a region of three-dimensional space illuminated by an unchanging arrangement of lights is called the plenoptic function (from plenus, meaning complete or full, and optic). The term was first coined in Adelson and Bergen’s[7] 1991 paper prior to the existence of digital light field camera technology. The plenoptic illumination function is an idealized function used in computer vision and computer graphics to express the image of a scene from any possible viewing position at any viewing angle at any point in time.

2.2 4D Light Fields. The camera, designed with inspiration from the human eye, takes in light rays until it hits a sensor, giving us a stopping point, and causing some redundancy from the 5-D plenoptic function because the radiance along a ray remains constant from point to point along its length. Researchers in computer graphics coin terms in two critical papers to the development of the field, each published in 1996 - Levoy calls it the 4D light field [8] and Gortler calls it the Lumigraph [9]. Formally, the 4D light field is defined as radiance along rays in empty space.

2.3 The Plenoptic Camera The 4D light field gained means of photographic measurement in the form of the plenoptic camera, which has existed since inventor Gabriel Lippmann created a technique he called "photographie integrale", which from French usually translated to Integral photography - a suggested integration of a whole image from many different images (or parts); however, complete photography is also a valid translation. Lippmann modeled the invention from the inspiration of an insect eye. The image his technique produced was made of up may smaller images of the object being captured from slightly different angles thanks to the offset of the lenses. When inserted into his viewing apparatus, the observer could move their head to different viewpoints for motion parallax and a sense of depth which is produced through the stereoscopic rendering two different viewpoints create. Lippmann found that twelve lenses were sufficient for his experiments.

Normally when extracting stereo pair images from traditional cameras, they have to calibrated in advance, and there is still risk of images being out of focus, making depth estimation unreliable. A single plenoptic camera is able to take a photo that gives an array of images that can be used for stereo pairs. It also gives the ability to refocus an image that may have originally been blurry. A plenoptic camera gives up an aperture, which allowed for adjustment to the depth of field (DOF), and instead captures all light rays in the field of view (FOV) and DOF. RAW shown in Fig. ??.

The plenoptic camera really was put into the practice when Adelson and Wang released their 1992 paper titled Single Lens Stereo With a Plenoptic Camera.[10] Their paper took inspiration from Lippmann’s integral photography. Adelson and Wang gave birth to what would become known as the Plenoptic 2.0 camera (Fig. ??), Lippmann’s being 1.0. Many took notice of Adelson and Wang’s work and innovation into the realm of the light fields became digital. The appropriately named Digital Light Field Photography [3], a Stanford thesis by Ren Ng done under the mentorship of Marc Levoy would be the breakout for the future CEO of what would be the first consumer-based provider of light field cameras. Ng’s company known as Lytro has produced lightfield hardware technology for photography, cinematography, and Virtual Reality. German competitor Raytrix has developed a high end plenoptic cameras for industrial use. One such being a 3D-Camera with Extended Depth-of-Field proposed in 2012. [11]
2.4 Depth Estimation with Plenoptic Cameras. No longer state-of-the-art, but for the sake of understanding depth estimation, in 2010 Tom Bishop and Paolo Favaro [12] published a method to extract the depth of every pixel from a plenoptic camera-produced image. In their depth map, areas closer to the camera are shaded darker. In addition, a depth map can be combined with the color captured by the camera to represent a fuller image. Bishop and Favaro utilized Lambertian (or ideal matte) objects for their experimentations, leaving the issue of glossy or texture-less objects for later experimenters. The two proposed an iterative procedure to solve the challenges of first, applying an anti-aliasing filtering before constructing the views, and second, the aliasing in plenoptic cameras varies with the depth of the scene. For their solution, they started with a strong anti-aliasing filter corresponding to the smallest usable bandwidth and then refined the iteration as depth is estimated. The method first defines an antialiasing filter onto the microlens array. They project it onto the conjugate image at z (derived along the z axis) and then through each microlens onto the sensor. The scaled filter has a physical cutoff frequency once the disparity map update is negligible. Too much filtering might introduce additional solutions to the energy minimization problem, thus reducing sensitivity. Too little filtering however might remove the correct solution. The algorithm can be summarized as following: (1) Initializing the depth map to the depth at which there would be the most aliasing in the working volume. (2) Estimating the disparity map by minimizing the joint matching error between all combinations of pairs of views. (3) The views are then rearranged as subimages. (4) For each index of microlens pairs, they apply a filter the subimage by the current frequency cutoff, where step 1s depth map is the estimated depth map. (5) Repeat steps 2 - 4 until the disparity map update is negligible.

3 State of the Art Light Field Depth Estimation Methods

The two types of methods primarily used for most state-of-the-art depth finding methods are Multi-View Stereo (MVS), which takes in multiple images and calculates the disparity between images based on a pixel value, and Epiplolar Plane Image (EPI) solutions which compare two images and measure using a line called the epiploar line cutting through an image, matching pixel values in the other image and calculating disparity / depth based on that (both seen in Fig. 3).

3.1 Globally consistent depth labeling of 4D light fields (2012) In this paper Wanner and Bastian [15] present a reformulation of stereo matching to a constrained labeling problem on epipolar plane images, which they recommend thinking of them as vertical and horizontal 2D cuts through the field. This paper presents a framework for variational light field analysis designed to enable the application of modern continuous optimization methods into usable 4D light field data (specifically depth estimation). This paper also introduces a local data term for depth estimation tailored to the structure of light field data and more robust than what traditional stereo matching methods offer.
to non-Lambertian (matte) objects. This paper also introduces a labeling scheme based on state-of-the-art convex relaxation methods, which allows for the estimation of globally consistent depth maps able to satisfy visibility constraints for all (possibly hundreds of) views simultaneously. This paper draws from previous work done with EPI’s, which have already shown the capability of EPI’s for detecting edges, peaks and troughs with a subsequent line fitting in the EPI to reconstruct 3D structure. Also, previous EPI work shows an iterative extraction procedure for collecting EPI-lines of the same depth, called an EPI-tube. Lines within the same tube are detected by shearing the EPI and analyzing photo-consistency in the vertical direction. In addition, they proposed a procedure to remove specular highlights from already extracted EPI-tubes.

Estimating depth estimates within the EPI first requires the estimation of line direction on the slice. They do this using a structure tensor of the epipolar plane image. They use a Gaussian smoothing operator at an outer scale. The direction of the local level lines is computed using an optimal orientation detection method, from which is derivable the local depth estimate. This method offers a consistent EPI depth labeling, which takes all views into account simultaneously. The optimization yields a piecewise smooth estimation with smooth occlusion boundaries. This optimization is particularly beneficial when considering non-Lambertian surfaces due to their outliers only being within a small subset along the EPI-line. Afterwards, global integration further localizes edges and suppresses noise. This method performed better than stereo methods and Raytrix at the time. It was also able to handle non-Lambertian surfaces.

3.2 Structure and Motion from Scene Registration (2012) Basha et al. [16] propose a method for estimating the 3D structure and the dense 3D motion (scene flow) of a dynamic nonrigid 3D scene, using a camera array. The core idea behind this method is to take a set of camera images create a dense 3D volumetric representation of the 3D space where each voxel (a 3D pixel) holds an estimated intensity value and a confidence measure of this value. This approach sidesteps the usual need to recover and/or handle occlusions and doesn’t require focused reasoning about flow discontinuities, which is typically a requirement that negatively affects scene flow methods. Instead of matching individual light rays, this method matches 3D points where each point aggregates the distribution of light in multiple directions. This feature makes it robust in handling dynamic scenes. This method takes N cameras to capture a dynamic non-rigid scene at two different time steps. Each set of images captured at a single time step, is used to construct a 3D volume, where each cell holds a 2D distribution of the light rays that pass through the the point in the 3D volume. They then approximate the 3D volume to obtain a scalar volume, by applying a nonlinear filter to the captured light rays at each individual scene point. The scalar volume consisting of real scene points as well as points in free space, is a piecewise continuous representation of all three dimensions (x, y
and z). This allows the application of a dense matching of the two scalar volumes, computed at two time steps, prior to 3D structure recovery of the scene. By doing this, the method bypasses the need to address occlusions or sharp discontinuities in both the 3D structure and 3D motion field. Lastly, the computed flow between the two scalar volumes is used for extraction of both the 3D structure and the 3D motion; also allowing for the recovery of the sharp discontinuities in both the depth and the motion field.

3.3 Scene Reconstruction from High Spatio-Angular Resolution Light Fields (2013)
Ray coherence of a dense 3D light field allows this algorithm [17] to operate on individual EPI-pixels instead of having to consider larger pixel-neighborhoods like most stereo approaches use with patches. Due to this, it performs especially well at identifying depth discontinuities and reproduces precise object silhouettes due to the color contrast in these regions. Utilizing this property is key to their fine-to-coarse depth estimation strategy. They first estimate depth at edges in the EPI at the highest resolution, then propagate this information throughout the EPI, and move to the following coarser EPI resolutions. Starting with a full resolution EPI this method, 1) Efficiently identifies regions where the depth estimation is expected to perform well. 2) From here, they introduce a fast edge confidence measure that is computed on the EPI. The algorithm then generates depth estimations for EPI-pixels with a high edge confidence measure. 3) The density estimation is then utilized to improve the initial confidence towards a refined depth confidence, which then provides a good indicator for the reliability of a particular depth estimate. 4) All EPI-pixels with a high reliability are stored as tuples a set and propagated throughout the EPI. 5) Depth estimation and propagation is iterated until all EPI-pixels with a high edge confidence have been processed. 6) To this point, there should be sufficiently detailed regions at the current resolution level of the EPI that have a reliable depth value assigned, while the depth in more homogeneous regions is yet unknown. 7) Using a fine-to-coarse approach then allows down sampling the EPI to a coarser resolution and starts over with the above procedure - computing edge confidence for parts of the EPI that are not yet processed. 8) This procedure is continued until a depth value is assigned to every EPI-pixel, the resulting line segment tuples in the set now are able to reconstruct the complete light field.

3.4 Variational light field analysis for disparity estimation and super-resolution (2014)
Wanner [18] expands on his previous paper [15], making several interesting enhancements. For disparity estimation, they first calculate local slope estimates on epipolar plane images for the two different slice directions using the structure tensor. In 3D space, a point is projected onto a line in whole of y and t space where the slope of the line is related to its depth. The calculation results in two local disparity estimates for each pixel in each view. These disparity estimates can then be merged into a single disparity map in two different ways. Either locally choosing the estimate with the higher reliability, optionally smoothing the result (very fast), or solving a global optimization problem (slow). The experiments show that the fast approach leads to estimates which are slightly more accurate. For disparity ranges 1 (Lytro falls here), the algorithm is very robust, however, the results get worse for larger disparity values when impulse noise is added to the input images.

This method proposes an essential parameter-free super-resolution view synthesis which computes super-resolved textures for a 3D model from multiple views. It does not require a complete 3D geometry reconstruction of a texture atlas, making only use of disparity maps on the input images. The idea behind super-resolution is to define a physical model for how the sub-sampled images can be explained using high-resolution information. The super-resolution is superior to both competing methods compared against. The quality of the disparity map produced with super-resolution is increased significantly, which already had accurate occlusion boundaries and was of high quality. The super-resolution refinement process has high computational cost however. Wanner suggests that theoretically, that within this framework, one could analytically derive weighting factors for the contributions of the input views caused by foreshortening effects due to scene geometry.

3.5 Multi-Resolution Approach To Depth Field Estimation In Dense Image Arrays (2015)
Neri et al [19] recognized that substituting data term only local optimization schemes with
global methods, which causes a drastic increase in computational complexity. Recognizing this, the proposed method in this paper adaptively combines local, data term only, and multi-view stereo depth estimates at different spatial resolutions obtained with matching windows locally adopted to the gradient magnitude. Through analysis of their solution, they've found that their proposed method outperforms both the multi-view stereo global and EPI methods at the time of their publication. This paper addresses the case of accurate depth map estimation including image post-processing, scene understanding and coding. Depth methods in large image array problems can be mostly classified as those based on extensions to multicameras of stereo solutions and those based on the analysis of epipolar images (EPI).

3.6 Accurate Depth Map Estimation from a Lenslet Light Field Camera (2015)
Unique to Jeon et al. [2] was their inspiration to build a lenslet lightfield camera, inspired by the Pelican Imaging Camera-Array, or PiCam presented by Venkataraman et al. [20]. PiCam addresses several modifications needed to achieve a strict form factor and base image quality required to make array cameras practical for mobile devices. In PiCams approach, they customized many approaches of their camera array including the lenses, pixels, sensors, and software algorithms all which lead to a performance and form factor that could be used on a mobile phone camera. Jeon et al. used a commercial mirrorless camera (Samsung NX1000) and modified it through removing the cover glass on its CCD sensor and affixing their created lenslet array. Microlenses, or lenslets within the array each had a diameter of 52m, an inter-lenslet distance of 79m, and a focal length of 233m. The main lens had a focal length of 50 mm and an F-number of 4. The camera was calibrated using an open toolbox method provided by Bok et al. [21], which, instead of utilizing sub-aperture images, they directly use the raw images to calibrate micro-lens-based light field cameras. Bok et al. [21] does this by extracting line features by comparing the micro-lens image regions to the raw image. Instead of using post-processed information for initial image processing, the proposed method references Bok et al.’s method [21], in which they extract line features from raw images and utilize them to calibrate the micro-lens-based light field cameras.

This proposed method was developed using a cost-volume-based stereo. The adjustments took into consideration the narrow baseline between the sub-aperture images, thus the pipeline was tailored with three important differences for a three-fold method to accuracy. 1) Rather than passing through the local patches to calculate the cost volume, the sub-aperture images were directly shifted using a phase shift theorem and the per-pixel cost volume was then computed. 2) In order to effectively aggregate the gradient costs computed adaptively from dozens of sub-aperture image pairs, a weight term that considers the horizontal/vertical deviation in the st coordinates between the sub-aperture image pairs is defined. 3) Due to the small viewpoint changes of the sub-aperture images, feature correspondences between the sub-aperture images are used as an additional constraint. With the cost volume, a multi-label optimization propagates and corrects the depth map at texture regions identified as being below satisfactory. From here, the method iteratively refine a local depth map by fitting local quadratic function to estimate a new non-discrete depth map. This proposed method collects matching costs using robust clipping functions, allowing it to tolerate significant outliers. Also, the calculation of the exact sub-pixel shift using the phase shift theorem improves the matching quality. Jeon et al. came up with an effective solution to the significant challenge of addressing the sub-pixel shift for the frequency domain. It does this by implementing a sub-pixel-wise disparity estimation. At the time of publication, this method performed better than both the Lytro and Raytrix depth-estimation methods available for their hardware and outperformed three other existing advanced methods.

3.7 What Sparse Light Field Coding Reveals about Scene Structure (2016)
Johannsen et al. [22] present an EPI method for depth estimation in light fields which employs a specifically designed sparse decomposition to leverage the depth-orientation relationship on its epipolar plane images. The proposed method learns the structure of the central view and utilizes the data to construct a light field dictionary for which groups of atoms correspond to unique disparities. This
dictionary is then utilized to code a sparse representation of the light field. Through analysis of coefficients in this representation with attention to the disparities of their corresponding atoms, there is a yield of accurate and robust depth estimation. Statistical analysis of the coefficients can be employed on multi-layered light fields to infer the respective depth of the superimposed layers in either reflective or transparent object surfaces. This method employs what they call twin peaks, or two disparity layers. The dictionary has trouble explaining regions that lack texture. In this work, for the first time, the authors unify the idea of orientation-based depth reconstruction with sparse light field coding based on generation of a depth-based dictionary. Using the generated light field atoms, they employ the Lasso as mentioned by [23] in order to compute sparse coding coefficients. Experiments demonstrate that this method far surpasses previous work for multi-layered disparity estimation (transparency and reflective surfaces) in robustness and accuracy. For strictly Lambertian scenes this method still performs on par with earlier methods which are not occlusion-aware.

3.8 Depth from Gradients in Dense Light Fields for Object Reconstruction (2016)
Ycer et al. [24] are interested in reconstructing an entire object in 3D for computer graphics and virtual reality. Despite the significant research efforts up to the point when this paper was produced, objects with with thin features and fine details still pose a problem to multi-view stereo (MVS) due to numerous reasons shared in the paper. The first being the fact that these features only occupy a small number of pixels in the views they are visible in, making location detection difficult. Also, fine features are typically only visible in a few of the views within the multi-view stereo sample. Matching thus becomes difficult. Most patch-based MVS techniques aren’t able to detect thin features, since they typically need the patches of objects to be several pixels wide. In addition, graph-cut based reconstruction techniques like those in Hornung and Kobbelt 2006 paper, Hierarchical volumetric multiview stereo reconstruction of manifold surfaces based on dual graph embedding, and Vogiatzis et al’s 2007 paper Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency face difficulties with texture-less thin features due to it being hard for such methods to localize the features using photoconsistency values inside a volumetric discretization, which often results in the elimination of these features in the reconstructions.

Progress has been made with light field depth maps with fine detail detection by both Kim et al. [17] and Wanner et al.’s paper [15], both of which have already been discussed. The results of those papers proposed methods are per-view depth maps, which are not always globally consistent and require an additional surface reconstruction step to generate more usable representations for something like triangle meshes. In addition, such techniques often specialize in regularly sampled light fields (such as regular 2D grids) and do not easily adaptable to more casual, hand-held capture scenarios. A more recent method proposed by Ycer and his team in June 2016, in their paper Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction, they were able to generate pixel-accurate segmentations for casually captured light fields using pixel-level approaches. However, segmentations could only be used to describe the objects in form of visual hulls, where concavities are missing.

3.9 Robust and Dense Depth Estimation for Light Field Images (2017)
Navarro et al. [25] give a detailed survey of different state-of-the-art depth estimation methods to the point of publication. This paper estimates the disparity between pairs of specific views by a two-view stereo method and then combine these results to obtain a robust disparity estimation. They utilize the state of the art two-view stereo method for computation of the disparity between these two fixed views. This algorithm also includes a multi-scale and a multi-window process, which permits the correct estimation of slanted surfaces and depth discontinuities. Thanks to the validation criteria, only correct matches are kept thanks to validation criteria. This paper introduces a new interpolation method based on optical flow formulation in order to fill in invalidated matches. They make a modification of the optical flow state-of-the-art, introducing a new fidelity term imposing the estimated flow to be near the computed disparity at the validated points. This method amplifies the precision of the matches computed by which have an a priori fixed and finite precision.
4 Conclusion

This paper has been an overview of light fields and current depth estimation methods being used in the industry today. There is great potential in the world of light fields moving forward in terms of post-production of new ways of acquiring 3D models.

References