Plenoptic Face Reconstruction

Gonçalo António dos Santos Cruz e Carreira Pedro
Instituto Superior Técnico / Universidade de Lisboa, Lisbon, Portugal
goncalo.pedro@tecnico.ulisboa.pt

Abstract

Two dimensional face models allow presentation attacks, based on photographs or displays. Research in the three dimensional face models is necessary to achieve improved security. This work aims to investigate the use of plenoptic cameras in 3D face reconstruction, focusing on the reconstruction of low gradient areas. Plenoptic cameras capture a scene from different viewpoints and store the information in a single image sensor, thus enabling 3D reconstruction.

Current face reconstruction methodologies based on edge-points reveal difficulties within low gradient areas. A preliminary study lead to the conclusion that segmenting low gradient areas into large enough patches yields enough information for reconstruction. The application of light field shearing on patches bordered by, but not including, high gradients, is the basis proposed for a face reconstruction method.

Experiments on synthetic and real data show consistent depth estimation in low gradient areas using the proposed method, providing additional information to edge-based reconstruction. Two alternative reconstruction methodologies have been analyzed in the context of face applications, and the proposed reconstruction method has been found to provide promising comparison results.

I. Introduction

Plenoptic cameras [15] are capable of imaging a scene from different perspectives, unlike conventional cameras. The information of the different perspectives is stored on a single image sensor, which enables 3D reconstruction easily from a single shot [14], but it also limits the field of view and the 3D reconstruction.

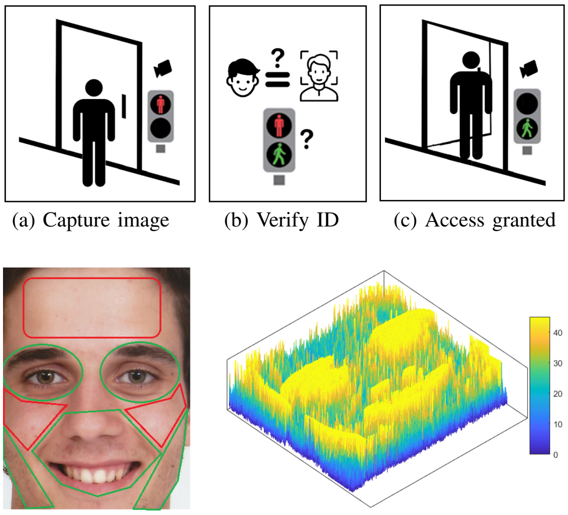
Three-dimensional face models are widely used for several purposes such as: biometric systems, face verification, facial expression recognition or 3D visualization. However, reconstructed face models generated from the optical setups used are quite noisy, due to the lack of texture and thin structures present in the face.

This work is developed in the framework of the research project proposal "Plenoptic Face Imaging and Biometrics in Identity Documents for Security Applications" which focuses on strong authentication combined with robust ID-docs. Our work explores the 3D reconstruction of faces targeting authentication purposes.

A. Facial Biometrics in Security Applications

The research focus of this work is exploring the use of light field imagery, acquired with plenoptic cameras [15], for face detection and recognition.

An application example is using facial recognition to access a secured place, as presented in the storyboard of Fig. 1. A person arrives at the entrance of a secured location and a picture of the face is automatically captured (see Fig. 1(a)). The captured picture is compared to the images stored in the access-control database to validate the person's ID and verify access permissions (see Fig. 1(b)). If there is a match and the access is granted, the person can enter the premises (see Fig. 1(c)). Other examples include ID-doc validation as an aid to law enforcement and digital signature via ID-doc.



(a) Capture image (b) Verify ID (c) Access granted

(d) Face regions (e) Gradient of the image

Fig. 1. Storyboard - using face recognition for access control: (a) image captured at the entrance of secured location; (b) ID and permissions verified in the access control database; (c) Entry allowed only to authorized personnel. Gradient levels in the face: (d) highlight of low gradient regions (red) and high gradient regions (green); (e) plot of gradient information found in a face.

Three-dimensional face models are widely used in biometric systems. The quality of the 3D reconstruction is crucial for these systems, and reconstructed face models are noisy, due to the lack of texture and thin structures present in the face. Nowadays, plenoptic (light field) cameras are point and shoot, portable and have low cost, enabling the development of new Presentation Attack Detection solutions [20].

The ubiquity of smartphones makes them the device of choice for strong authentication. In [12] Mildenhall et al. proposed a light field acquisition setup based smartphones and view synthesis, opening the door for the general public to acquire light field imagery at a low cost.

B. Related Work on Facial Biometrics

Face recognition is a widely accepted biometric in security applications. Facial biometry systems originated in the military in the 1960s and started as manual systems. In the 1990s this area of research was dominated by solutions mapping the input to a lower-dimensional space, like eigenfaces [21], using conventional cameras. Later, model based solutions were developed to overcome sensitivity to scale, pose and facial expression.

Convolutional neural networks (CNN) instigated research on this matter yielding promising results, the best example is Google's *FaceNet* [18]. Other methods arose, such as recovery of a 3D model from a single image, resorting to auto-encoder chains [23]. Application to light fields had only *FaceLFNets* [5] which recovers 3D facial curves.

Alperovich et al. proposed an unsupervised deep encoder-decoder network [1] to extract light field intrinsics, performing disparity estimation, diffuse/specular separation and reconstruction, and being able to estimate depth in highly specular scenes [1]. In the line of handling specularity Johannsen et al. proposed a sparse coding approach to detect if a light field possesses specularity and use the appropriate (one or two-layer) model to estimate disparity.

In [6] Ferreira et al. propose an automated method for depth estimation using different focal length lenses, yielding results comparable to state of the art methods in less time.

C. Problem Formulation

The European Union currently faces a challenge to protect its citizens' freedom and security without compromising their privacy nor limit their freedom. The use of biometric information to provide stronger authentication is becoming increasingly important for human activities, such as banking. Face biometry is already used nowadays to unlock smartphones and computers. The objective of this work is to provide facial biometric information to help authentication processes, resorting to plenoptic setups three-dimensional information can be retrieved to reconstruct the 3D structure of a face.

D. Report Structure

Section 1 introduces the problem to approach in the work, face reconstruction. In particular presents a short discussion on the state of the art on facial biometrics and reconstruction. Section 2 introduces face modeling, background on plenoptic cameras, and reconstruction methodologies. Section 3 presents a conceptual experiment for reconstruction in low gradient areas. Section 4 presents the proposed method for face reconstruction. Section 5 details the experiments performed with faces, including the creation of the synthetic face models. Section 6 summarizes the developed work and highlights the main achievements. Moreover, this section proposes further work to extend the activities described in this document.

II. Background

Plenoptic cameras enable single shot capture of sufficient information to retrieve 3D information of the world. To comprehend how light fields yield this information it is necessary to model the plenoptic camera. Before proceeding into the camera model we describe the fundamentals of face modeling, required for the creation of synthetic data.

A. Face Modeling

The current pandemic imposed a need to generate synthetic face data, which requires 3D modeling of faces. To generate trustworthy human face models a shift from two dimensional space (picture) to the three dimensional space (face model) is required. Techniques for this purpose include using one or multiple cameras, 3D scanners, and combinations of sophisticated software and hardware. Humans use the face as the main distinguishing feature between people due to its discernible features, such as eye color, shape of sensory organs and wrinkles. An example of the face gradient regions and values is depicted in Figs. 1 (d) and (e).

The face model generation comprises three steps: data acquisition; 3D registration followed by 3D model deformation; and texture generation to cover the 3D model [4]. The data acquisition is the process of capturing the reality through photographs. Feature points are identified and a dot pattern is fitted, which resorting to the deformation of a standard 3D face model yields 3D coordinates for the points. The texture generation consists on gathering color values from the original images and collecting them in the appropriate locations of the texture image, resulting on a textured 3D face model.

B. Plenoptic Imaging

Plenoptic cameras [15] acquire light field images, which represent the light intensity in multiple directions for each point in a plane. They are built based on an array of microlenses in front of a main thin lens, to create an artificial compound eye.

Each microlens is modeled as a pinhole that captures a slightly different perspective from its neighboring lenses. For this reason, standard plenoptic cameras (SPC) can be seen as a camera array [13], where each camera corresponds to a viewpoint that captures a different image. Because of their approximately continuous baseline, SPCs allow the computation of disparities as gradients of Epipolar Plane Images (EPIs).

1) Back-projection Model: Light fields can be defined in the object or image space. In the object space the light field can be described using the two planes parametrization (see Fig. 2). The light field is then indexed by $(s,\ t,\ u,\ v)$, where (s,t) identifies the intersection of light ray and the plane and (u,v) expresses a direction, given by the intersection of the light ray with the plane Ω at unitary distance from Γ.

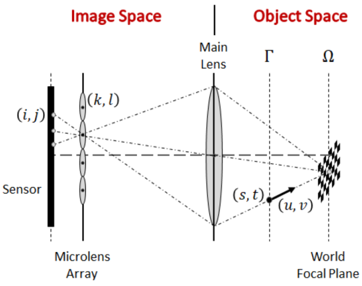


Fig. 2. Geometry of a standard plenoptic camera, with parametrization spaces marked. Extracted from [13]

Light fields can be also represented in the image space (see Fig. 2), where the parametrization is intrinsically related to the camera: a pair (k, l) represents the selection of a microlens and the pair (i, j) indexes the pixel underneath the selected microlens.

The capture occurs in the image space and the metric information is present in the object space, the conversion between spaces is a back-projection, as proposed by Dansereau et al. in [3]. Here we follow the notation of Marto et al.[10] for simplicity, which is:

$$\Psi = \mathbf{H}\Phi \Leftrightarrow \begin{bmatrix} s \\ t \\ u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} h_{si} & 0 & h_{sk} & 0 & h_s \\ 0 & h_{tj} & 0 & h_{tl} & h_t \\ h_{ui} & 0 & h_{sk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vj} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \\ l \\ 1 \end{bmatrix} \tag{1}$$

where the 5 by 5 matrix is the intrinsic parameters matrix $\mathbf{H}$, Ψ denotes the light field in the object space and Φ denotes the light field in the image space.

2) Light Field Reconstruction: After the processing of the acquired images, when a light field is obtained, reconstruction can be considered. In a camera array we can select a line of viewpoints and stack the corresponding images, resulting in a 4D image volume (hypercube). A slice of this hypercube is an Epipolar Plane Image (EPI).

EPIs show the effect of parallax, which is the difference in the apparent position of an object regarding a background when viewed from different positions. Specifically, lines corresponding to closer objects have greater slope than lines from objects further away. This variation in the pixel position of a world feature relative to the variation in the camera considered is called disparity.

A single feature has multiple projections, one for each viewpoint, thus we can use $\mathbf{H}$ to find a constraint that a collection of rays corresponding to the same feature must follow [10]. The relation between space and pixel indexes assumes previous knowledge of z, but assuming a constant position for a feature and taking its derivative, we relate depth and disparity:

$$z = -\frac{h_{si} + h_{sk}\frac{\partial k}{\partial i}}{h_{ui} + h_{uk}\frac{\partial k}{\partial i}} \quad \vee \quad z = -\frac{h_{tj} + h_{tl}\frac{\partial l}{\partial j}}{h_{vj} + h_{vl}\frac{\partial l}{\partial j}} \tag{2}$$

where the disparity is represented by the gradients $\frac{\partial k}{\partial i}$ and $\frac{\partial l}{\partial j}$.

Thus, we obtain 3D points by estimating the disparity, computing the depth z, and use z on the relation between space and pixel indexes to determine x and y [10].

C. Reconstruction Methodologies

We now introduce of a depth estimation algorithm and review the workings of state of the art methods for reconstruction.

1) Gradient Based Depth Reconstruction: In [10] Marto et al. proposed a method to perform depth reconstruction from light field imagery. The method relies on epipolar plane images (EPIs) to perform disparity estimation, extracting the gradient of the image from the EPI using structure tensors.

The structure tensor, $S(k, l)$, has a structure tensor for every viewpoint pixel index, condensing information from horizontal and vertical EPIs. Its eigenvectors allow disparity estimation and its eigenvalues yield a confidence measure for the disparity estimation [2]. Then, low confidence estimates are disregarded, noise is handled and regularization is performed to obtain data pertaining the whole viewpoint area. Lastly, the depth map is obtained from the regularized dense disparity map, using Eq. 2.

2) Reconstruction Fundaments: Light fields have multiple possible representations, but the main representations are: subaperture views (viewpoint images), EPIs, Surface Cameras (SCams) and Focal Stacks.

Mutli-view stereo methods use the subaperture views, relying on patch comparison to find the best correspondence among the images for a set of disparities [8]. Methods that rely on EPIs are also frequent, since 3D points are projected onto lines in the EPIs the depth estimation is reduced to orientation analysis of said lines.

Angular patches or Scams sample the radiance for all corresponding projections of a scene point at the respective depth [24] and can be leveraged to analyze occlusions. The focal stack is composed by a set of refocused images, each at a given depth z, which is no more than integrating over the angular patches at that depth [8].

Depth estimation can be achieved using any of the proposed representations. Generally, light field depth estimation algorithms follow a common pipeline which consists in three main steps (i) information selection, (ii) first reconstruction, and (iii) refinement of the initial estimations.

The first step comprises the choice of the light field representation(s) to be used, as well as the selection of the views to be considered, resulting in a cost volume. The second step performs disparity estimation via global optimization of the cost volume. Common methods are Markov Random Fields (MRF) and graph cut approaches or variations of these with regularization. The refinement stage aims at filling in the missing information of the initial estimation. Usually consists in local filtering (weighted median or bilateral filters) or global regularization of the disparity map.

3) Spinning Parallelogram Operator: The Spinning Parallelogram Operator (SPO) proposed by Zhang et al. in [25] addresses the problems caused by occlusion and noise in light field depth estimation. Using a crosshair of views, the method is used to locate lines in the Epipolar Plane Images (EPI) and to estimate their orientation. Furthermore, local and global confidence measures are calculated to handle occlusion, followed by filter-based in-painting to cover texture-less regions.

The method splits the EPI into two regions - slightly to the left and right of the line in question - and computes histograms for both. The distance between the distributions of color is measured using the χ^2 difference of the histograms. Then, a confidence metric is defined taking the difference between the maximum and average scores, resulting in low confidence for ambiguous and occlusion zones.

The cost volumes for both EPIs are then combined according to the confidence metric, through a weighted summation. The resulting cost volume is regularized for each individual depth label, with the correct information being propagated to similar regions with low texture, using a filter-based method. Lastly, a disparity map is generated using a winner-takes-all strategy.

The line's orientation is determined through the maximization of the distance between the histograms of pixel intensity. Large differences between the histograms indicate the presence of an edge dividing the regions. Thus, the maximum orientation response is taken:

$$\Theta_{y,v}(x,u) = \arg\max_{\theta} \ d_{y,v}(x,u,\theta) \tag{3}$$

where $d_{y,v}(x,u,\theta)$ is the histogram distance measured by the SPO on the EPI $I_{y,v}(x,u)$ (analogous for the vertical EPI). Using θ to define the direction of the lines the corresponding local depth estimations can be obtained by:

$$Z = f\frac{\Delta u}{\Delta v} = \frac{f}{tan\theta} \tag{4}$$

according to [22].

4) Light Field Superpixel Segmentation: A formal definition of light field superpixel (LFSP) is a set of all light rays radiated from a proximate, continuous and similar 3D surface [26]. The superpixel segmentation in light fields aims to simplify their processing by grouping similar pixels among all views in a consistent manner [9].

The method performs robust detection of lines in the central EPIs using directional filters, and line fitting to handle occlusion cases. It enforces view consistency, in an occlusion-aware way, by pairing the lines into regions using depth ordering. This angular segmentation in the EPIs is clustered in the last step, where the estimated disparity is used to regularize the process. Lastly, a propagation step fills unlabeled pixels. The implicit computation of disparity maps allows the recovery the depth information, using Eq. 2.

III. Patch Based Reconstruction

This section presents a conceptual experiment and proposes a method to retrieve depth information from low gradient areas. We first introduce the concept of light field shearing, which is the fundamental operation for the proposed method. Then, through a series of experiments we refine the idea of reconstruction using patches of locally planar surfaces. In the end, we summarize the findings of the experience with the proposal of a depth reconstruction algorithm.

A. Light Field Shearing

The process of shearing a light field is equivalent to changing the world plane in focus, as shown by Ren Ng in [17]. As a consequence of this operation, all objects in that plane have zero disparity appearing in the same position on all viewpoint images. In practice, this process consists in a translation of each viewpoint's position by an amount α proportional to its distance to the central viewpoint, as shown in equation 5.

$$\begin{split} L_\alpha(i,\ j,\ k_\alpha,\ l_\alpha) = \\ L(i,\ j,\ k_\alpha + \alpha(i - i_{center}),\ l_\alpha + \alpha(j - j_{center})) \end{split} \tag{5}$$

For features at a given disparity to become in focus, their disparity after shearing must be zero. Thus, the amount α by which the light field is sheared corresponds to the features' disparity. The translation of the viewpoint images, inwards or outwards, causes the appearance of undefined regions in the edges of the viewpoint images. This happens because we are translating the viewpoint to an area that was not captured in the original light field.

B. Light Field of a Locally Planar Surface

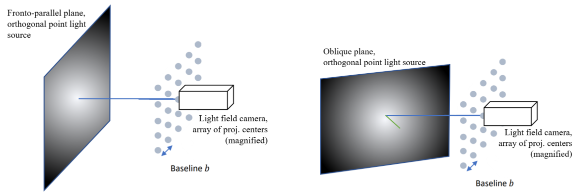
In this section we conduct a study to assess if reliable depth estimation can be achieved on smooth light field areas. The proposed approach uses the shearing operation on a local scale rather than relying on edge points. The study considers only light fields of planar Lambertian objects illuminated by a single point light source, which yield a smooth texture with a small level of gradient in it.

1) Lambertian Surface: When applying shearing to planar Lambertian surfaces, two light fields are considered: (i) the light field of a locally planar Lambertian surface $L(\cdot)$, and (ii) a virtual light field $L_0(\cdot)$, representing a textured plane at the focused distance, which has the same central viewpoint as $L(\cdot)$.

Definition III.1 (Light field of a (virtual) planar object placed at the focused plane (LFFP))**.** The light field of a planar object placed at the focused plane is the central viewpoint replicated at all viewpoints. In other words, it is an imaging texture assumed to be at a constant depth which is the focused one.

Under the assumption that local texture provides enough information for depth estimation, considering a local area for the shearing handles occlusion issues (which appear at depth discontinuities). Depth reconstruction is achieved through the registration of the local texture. Notice that the created virtual object, defining $L_0(\cdot)$, is a convenience for creating an intuitive property and demonstrating it. In practice, a reconstruction algorithm can be built directly on top of $L(\cdot)$.

2) Shearing Study Setup: The two experiment setups, depicted in Fig. 3, contain the same textured plane. The plane's texture emulates point light source illumination, which possesses small but non zero gradient information. The first setup comprises a fronto-parallel plane at a distance of 0.3 units, while the second setup uses an oblique plane.


Fronto-parallel plane, orthogonal point light source
Light field camera, array of proj. centers (magnified)
Baseline b
Oblique plane, orthogonal point light source
Light field camera, array of proj. centers (magnified)
Baseline b

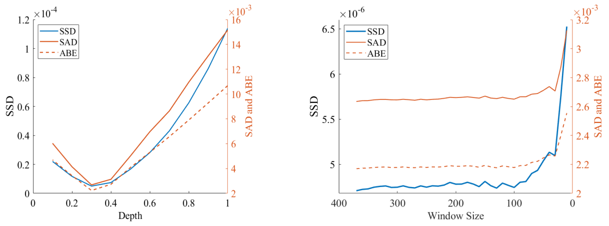
(a) First setup (b) Second Setup

Fig. 3. Proposed experience setups: (a) fronto-parallel plane with texture emulating point light source illumination; (b) oblique plane with the same texture (the plane's normal is highlighted in green).

After the light field acquisition, we apply shearing at different disparities, to the original light field. This results in a series of sheared light fields refocused at different depths.

We expect high viewpoint similarity, therefore we measure it using the difference between the central viewpoint image and every other viewpoint image. Specifically, we compute: Sum of Squared Differences (SSD); Sum of Absolute Differences (SAD); and Average Brightness Error (ABE).

3) Full Light Field Shearing: This experiment uses the first setup, and a set ten shearings equally spaced in a depth ranging from 0.1 to 1 units. We search for the shearing that maximizes viewpoint similarity through the computation of the error metrics, depicted in Fig. 4 (a). The results are consistent, a local minimum on the shearing depth of 0.3 units is found, which corresponds to the plane's real depth.



(a) Error vs Depth

(b) Error vs Window size

Fig. 4. Error metrics plots with SSD (in blue), SAD (in full orange) and ABE (in dashed orange): (a) error metrics against depth used for each shearing; (b) error metrics against the centered window sizes.

4) Local Window Shearing: After achieving a correct depth estimate using the full light field, we investigate if a portion of the image is enough to produce reliable results. On the same setup, only a centered window of the viewpoint image is used when calculating the error metrics.

We selected the shearing corresponding to the highest similarity and obtained the error values depicted in Fig. 4 (b). This figure reveals that the error stays within a limited range until a window size of 100 px, increasing significantly for smaller window sizes.

5) Locally Planar Surfaces: The experiment is now extended to study the effect in locally planar patches, namely the method's performance and the impact of patch comparison. We now use setup number two, which due to its inclination presents a foreshortening effect.

The procedure is the same as before. However, the global shearing corresponding to the lowest depth yields the smallest error, since objects closer to the camera occupy larger areas of the image (foreshortening effect). Since the right side of the image is mostly in focus it contributes with small errors, conversely the left side yields high errors. The minimum is attained at the best proportion of close to focus/out of focus areas, and occurs for a window size of 150 px.

So, instead of comparing a centered window, the viewpoint images were divided in a grid and the corresponding patches were compared. Areas closer to the camera present a significantly lower SSD value, while areas further away present higher SSD value. We conclude that the position in which the error is calculated influences the result. Moreover, the grid size also influences the values attained, with a grid of 30 by 30 yielding a minimum error 39% smaller than with the 3 by 3 grid.

The Lambertian assumption combined with a texture of some gradient is sufficient for our purposes, so we drop the single point light source hypothesis.

C. Depth Reconstruction Algorithm

We now propose a Naïve algorithm for depth estimation in smooth areas of the light field, to be used as a complement to edge based reconstruction methods. The algorithm searches for the disparity that locally maximizes the viewpoint similarity, which should correspond to the real depth value.

The first step is the division of the input light field in patches, assumed to have constant depth, following a grid scheme. The second step is the application of shearing to the light field patches, resulting in a set of sheared light fields, one for each sheared disparity. The third step determines the viewpoint similarity for the sheared light field patches through the sum of squared differences. The final step is depth assignment, performed via minimization of the sum of squared differences error.

The assumption of regular patches having similar depth values can quickly fall in real scenarios. In such cases the first step should be adjusted to perform a more adequate segmentation, namely with a free shape instead of squares, that yields locally planar patches. Furthermore, the independence between patches should be conditioned in a regularization step to prevent outliers.

IV. Face Reconstruction

In this section, we analyze the previously discussed reconstruction methods. Then, we propose an improved version of method proposed in the last section, targeting face application. The novelty comes from a more sophisticated approach to patch segmentation, based on level sets of the reconstruction confidence. This metric gets lower as one strays from edge points, however useful information can still be extracted from these regions. The remainder of this section details the calculation of the reconstruction confidence from the structure tensor followed by the detection and segmentation methods.

A. Reconstruction Methods

Here we highlight the strengths and weaknesses of the method proposed in III-C and compare it against the spinning parallelogram operator and the superpixel segmentation estimation. Then, we introduce refinement on patch segmentation which will be further explored in the remaining of the section.

1) Square Patch Based Reconstruction: Under the assumption that a human face can have locally planar patches, we consider the application of the algorithm proposed in section III-C. The application of this method to faces consists in a Naïve approach, designed to prove that it is possible to estimate depth where edge based methods struggle. Furthermore, it bears the advantage of being simple to understand and apply. The assumption of constant depth in grid squares is unrealistic, however it suffices for extraction of a dominant depth.

2) Spinning Parallelogram Operator: The SPO estimates the orientation of epipolar lines by comparing the regions on either sides of the proposed lines. Since this method outputs a depth labeling for each pixel, an additional step is required: the conversion from depth labels to metric depth values.

This method has proven very robust due to the comparison of small regions with weighting. Despite not explicitly modeled, the method still handles occlusion boundaries robustly because of the maximization of histogram distance. Furthermore, it performs very well in fine structure thinning and fattening, and has a strong performance in general discontinuities [8].

Conversely, in the matters of surface reconstruction it performs below average. Specifically, it struggles in areas of low texture (small gradients) despite having the best tradeoff on discontinuities and fine structures [8]. Nonetheless, the SPO is very robust to noise, artifacts and occlusion [25].

3) Light Field Superpixel Segmentation: The LFSP is defined in the 4D space and aims at light field segmentation [9], but since it implicitly creates a disparity map depth estimation can be attained. This method's clustering step does not explicitly handle occlusion, with only a mild influence by a high disparity weight in the regularization. Furthermore, when pixels are occluded from both sets of views the labels are propagated without occlusion awareness nor spatial smoothing. If the background and foreground share similar textures, it is difficult to segment them using existing cues.

4) Level Sets on Reconstruction Confidence: Since the method proposed in section III-C consists of a Naïve approach, we now propose an improvement to it. The difference lies in the patch segmentation step, which rather than blindly splitting the light field in squares it adapts the size and shape of the patches. To do so, the method relies on the reconstruction confidence obtained via structure tensor, thus enabling the detection of regions with lower confidence and focusing on those areas. This method will be further explained in the remaining of this chapter.

B. Level Sets on Reconstruction Confidence

To detail the working of the proposed method we start by describing the confidence measure and its computation, followed by how we use that information to segment the light field.

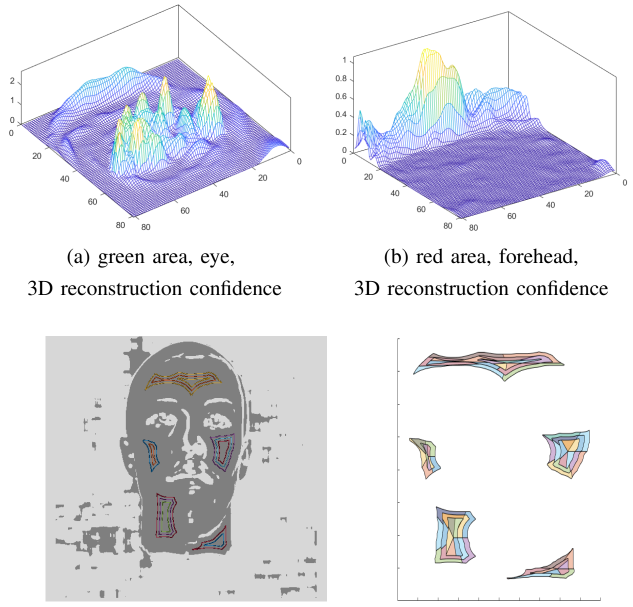
1) Confidence Measure: In section II-C1 we describe how the structure tensor can be computed and the information that can be extracted from it. Specifically, a confidence metric for the reconstruction is provided through the eigenvalues of the structure tensor λ_{max} and λ_{min}. The eigenvector corresponding to the greatest eigenvalue, λ_{max}, indicates the dominant gradient direction. Orthogonally we have the eigenvector corresponding to the smallest eigenvalue, λ_{min}, thus the more uniform the gradient directions the smaller its value.

Since the structure tensor is considered in the epipolar plane images, the focus is to provide a confidence level for the detected edges. Thus, the difference between the eigenvalues should suffice as a confidence metric for any given point. Specifically, the confidence c in each point is defined as:

$$c = \lambda_{max} - \lambda_{min} \tag{6}$$

as proposed by [2]. Once the structure tensor and the corresponding eigenvalues are calculated we possess a confidence map pertaining the whole image, as shown in Fig. 5(c).

2) Patches defined by Level Sets: In section II-A we described the types of gradient regions found in the face, see Fig. 1 (d), and in section IV-B1 we relate the gradient in the image with the reconstruction confidence metric extracted from the structure tensor.



(a) green area, eye, 3D reconstruction confidence

(b) red area, forehead, 3D reconstruction confidence

(c) Level curves over confidence map

(d) Level set segmentation

Fig. 5. Mesh plots of the confidence information: (a) around the eye region, green in Fig. 1 (d); (b) forehead, red in Fig. 1 (d). Stages of patch segmentation using reconstruction confidence: (c) confidence map with level curves overlaid, dark gray represents zones below the threshold and light gray zones above; (d) patches obtained via radial segmentation of level sets.

An example of the confidence in smooth regions is presented in Fig. 5 (b), where the frontier between hair and forehead (at the top) displays medium/high confidence and the forehead presents low confidence. Conversely, the eye region in Fig. 5 (a) comprises mostly medium to high confidence due to the high gradients present.

We use the confidence measure provided by the structure tensor, to create a binary confidence map via threshold. The segmentation is performed on top of the confidence map, resorting to the distance transform which enables the extraction of level curves. An intermediate step filters the level curves by size to choose only interesting regions, not too big nor too small.

Once we possess a set of light field patches, we can apply the depth estimation to each one. We apply shearing for a set of disparities, then we compute photo-similarity across the viewpoints of each sheared light field. The resulting array of errors for each patch, caused by the differences in the viewpoints, enables depth assignment through the minimization of the average error value.

Optimization can be applied on the reconstruction phase. We propose a scheme that targets clusters of level sets for independent optimization, yielding local consistency within the smooth areas. The optimization problem, computed for each zone, is of the form:

$$\min_{z} \quad C_S(L, z) + \omega_1 * C_C(z) + \omega_2 * C_R(z)$$

where z is an array with the depths for each patch in the zone and L the light field. The function C_S computes the SSD from viewpoint differences, while C_R and C_R yield the difference between neighboring patches on the same level and adjacent levels, respectively. This strategy searches for the shearing depth that minimizes the error, while simultaneously enforcing similarity between adjacent patches, yielding a consistent depth estimation for each smooth area.

C. Summary of the Proposed Methodology

We propose an algorithm to complement edge-based depth estimations in low gradient areas. The algorithm comprises two main phases: extraction of patches from the reconstruction confidence and 3D reconstruction from patches.

The first phase computes the reconstruction confidence and performs the light field segmentation. It receives a light field as input, computes the structure tensors, which yield the reconstruction confidence, and then extracts the level sets. Each level set is radially split onto pieces and the light field is segmented accordingly, resulting in a set of light field patches.

The second block performs the reconstruction from the light field patches. Each light field patch is sheared for a discrete set of disparities, then for each sheared light field the photo-similarity is computed. Lastly, we assign to the patch the depth associated with the smallest error (greater photo-similarity).

V. Face Reconstruction Experiments

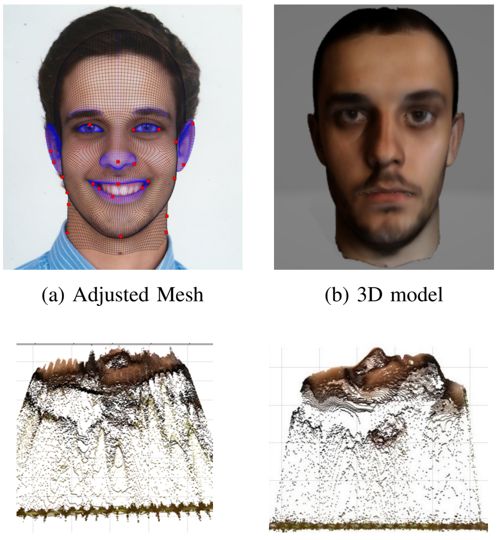
In this section, we test the methods on synthetic and real light field data. First, we describe the creation of the synthetic face model and compare the two methods used for acquisition purposes. The remainder of the section is dedicated to the four methods, first we assess the performance of the SPO [25] and LFSP [9] on faces. Then, we present the experiments that lead to the creation of our method followed by the results attained with it.

A. Synthetic Data

We start by creating a three dimensional model of a face, then we compose a scene resorting to virtual reality, and acquire the light field.

1) 3D Face Model Generation: The 3D face modelling was performed in *Blender* with the *FaceBuilder* add-on. This solution requires only a normal camera, like a webcam, to capture images that will shape and texture the model.

The creation of the model encompasses four main steps. The first step is the addition of a standard blank model to be deformed. The second step is to acquire reference photos of a face, which will provide texture to the model. The third step is to shape the model, for which a mesh of a face is adjusted over the face in the reference pictures, as shown in Fig. 6 (a). A small number of adjustment points, displayed in red, suffice for a correct fit. The last step is to deform the standard model and generate the texture, based on the previous step, yielding a 3D textured model (see Fig. 6(b)).



(a) Adjusted Mesh (b) 3D model

(c) VRML Reconstruction (d) Blender Reconstruction

Fig. 6. 3D face model creation. (a) Mesh with adjustment points (in red) over reference picture. (b) Example of a final 3D model. Reconstruction using [11] in profile view. (c) VRML data reconstructed. (d) Blender data reconstructed.

2) Dataset acquisition: Scenes were composed, to include the 3D face model in a realistic environment, and acquired using two approaches. The first approach resorts to VRML for scene composition and to Matlab's Virtual Reality toolbox for light field acquisition. With this approach the intrinsic matrix $\mathbf{H}$ has to be estimated and the ground truth information is harder to obtain. The second approach comprises both scene composition and light field acquisition in Blender, using a light field add-on [7] which yields the light field, ground truth information, camera parameters and metadata. This scene was composed in Blender units, where the focal distance is 8 units. For reference, the distance between ear tips is 1.5 units and the background is 3 units away from the tip of the nose.

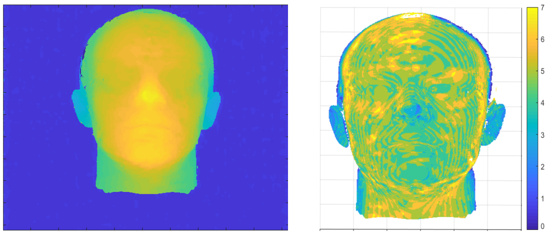
Reconstruction was performed using both datasets to evaluate data quality. A profile view of the results, shown in Fig. 6, was chosen for insight on the fine structure details. The VRML approach yields a larger but a much noisier reconstruction (see Fig. 6 (c)) than its Blender counterpart, which distinguishes small features such as the gap between the lips (see Fig. 6 (d)). Thus, the Blender approach [7] is a clear choice for data generation.

B. Spinning Parallelogram Operator

In this section we evaluate the performance of the spinning parallelogram operator (SPO) on both synthetic and real data.

1) Synthetic Data: Using the light field acquired in Blender we perform reconstruction using the SPO. Results, shown in Fig. 7 (a), reveal high similarity, with good depth discrimination.

The ground truth information was labeled to be comparable, yielding the computation of a labeling error. We only consider the error in the face region (see Fig. 7 (b)). The error goes up to 7 labels of difference in some points, however an offset of 4 labels was found. This offset can be explained by inaccuracies in the disparity range provided to the method.



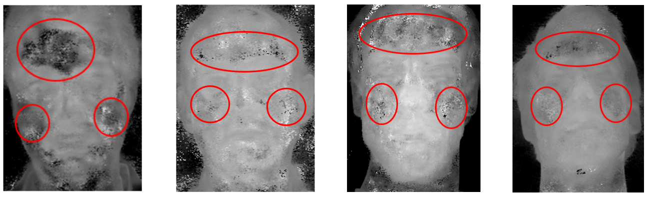
(a) SPO estimation (b) Filtered error

Fig. 7. SPO reconstruction results: (a) estimated depth labeling; (b) labeling error, pertaining the face region.

Despite the aforementioned offset in the labels, a clear background/foreground separation is attained, with similar contours to the ground truth image, for instance the gap between the lips was well captured.

To have a metric comparison we estimate an affine transform from the depth labels to the metric values provided in the ground truth image. The resulting mean absolute error (MAE) is 0.02 units for the face region and 0.03 units (approximately 4 mm) for the whole image. Due to the recovery of metric values using an affine transform of the ground truth, the error is small.

2) Real Data: We now evaluate the results obtained for real data from the IST-EURECOM face database [19]. The light fields were reduced to a bounding box of the face, provided in the database.



(a) Sub. 001 (b) Sub. 002 (c) Pose 1 (d) Pose 2

Fig. 8. Real data depth labeling with bad results highlighted. IST-EURECOM [19] data: (a) subject 001; (b) subject 002. Our data: (c) pose 1; (d) pose 2. [Credit for the light field acquisition: Miguel Rodrigues.]

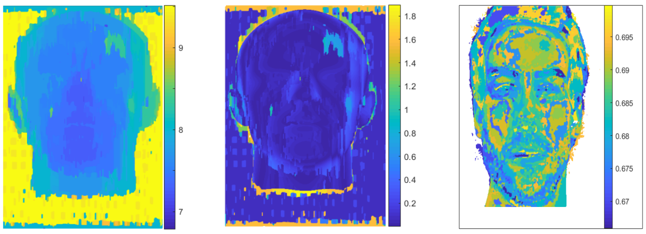
The application of the SPO in this cropped light field yields results significantly faster (than the full light field) and improves the quality of the results, specifically in the level of detail. The reconstructions, presented in Figs. 8 (a) and (b), show good background/foreground separation. However, the method struggles with low gradient regions and yielding estimation errors, as shown by the highlighted regions.

Lastly, we evaluate the method on light fields acquired by us, with a Lytro Illum camera in "selfie" position, yielding a depth range between 0.6 and 2 meters. Again we consider only the cropped light field for the reconstruction. Results, shown in Figs. 8 (c) and (d), reveal good background/foreground separation. Furthermore, facial features were captured, even if in a coarse way. However, we see a struggle with smooth regions (highlighted) where inconsistent depth assignment is found. We conclude that the SPO could benefit of a complementary method for estimation in low gradient areas.

C. Light Field Superpixel

In this section we evaluate the performance of the light field superpixel (LFSP) in depth estimation, on both synthetic and real data.

1) Synthetic Data: The light field acquired with Blender is used for the reconstruction. The reconstruction, shown in Fig. 9 (a), yielded a good background/foreground separation and a correct depth range. However, a significant lack of detail in the minor structures is noticeable.



(a) Synthetic data (b) Absolute error (c) Real data

Fig. 9. Light Field Superpixel reconstruction results: (a) synthetic data reconstruction; (b) synthetic data absolute error; (c) real data face reconstruction.

The error map, shown in Fig. 9 (b), reveals high error in the silhouette of the face due to errors in the background/foreground separation. The mean error is 0.17 units (approximately 2.3 cm), this value is explained by the error in the object frontiers, each background/foreground mistake accumulates around 2 units in error (approximately 27 cm). Furthermore, several smooth regions of the face stand out for having above average error (see lighter blue in Fig. 9 (b)). This method yields a correct depth range and general depths, however fails to provide correct depth for finer details.

2) Real Data: To further evaluate this method's performance we use the same real data acquired by us as in the SPO. We perform segmentation and use the intrinsic matrix **H** obtained via camera calibration to obtain a metric reconstruction. The superpixel size of 50 yielded an interesting segmentation for both light fields, with good clustering of smooth regions. The depth estimation presents a large error in the background but provides an acceptable depth range for the face.

The isolated face reconstruction, depicted in Fig. 9 (c), confirms a reasonable depth range and exposes inconsistencies in depth assignment related to the superpixel segmentation. This is particularly clear in the cheeks and forehead where a heterogeneous and inconsistent depth assignment is found, resulting in "walls" or "trenches" between smooth regions. Thus, we obtained compelling evidence that this method struggles in smooth areas.

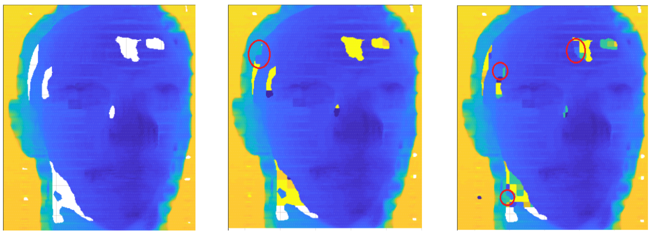
D. Square Patches

We now illustrate the preliminary experiments performed in face (VRML generated scene), with the method proposed in section III-C. The depth range for the scene is estimated to be between 0.05 and 0.1 units with the foreground/background frontier at around 0.07 units.

1) Error Metrics Evaluation: We select a global shearing, according to the depth range of the foreground, to assess the impact of the grid size in the error. Using grids with size ranging from 3 by 3 (100 px patches) up to 30 by 30 (10 px patches), we compute the error metrics.

The result for grids of size inferior to 10 by 10 is a coarse segmentation between foreground and background areas, with the latter yielding higher error values than the former. For a grid of 10 by 10 interesting results emerge, with smooth regions presenting a lower error value. High gradient regions present medium/high error values since they are not in focus in the global shearing. For a grid of 30 by 30 a better definition of smooth regions is attained while still displaying a low SSD value.

2) Reconstruction Results: The impact of the grid size, N, on depth estimations is now assessed, with the reconstruction algorithm applied to the dataset. In this analysis we complement the estimation obtained with [11], (see Fig. 10 (a)).



(a) 1st reconstruction (b) N = 20 grid (c) N = 30 grid

Fig. 10. Our estimations complementing a first reconstruction: (a) first reconstruction obtained with [11]. Reconstruction complemented with: (b) grid size 20 estimation; (c) grid size 30 estimation.

For values of N lower than 20 the results present insufficient level of detail resulting in erroneous depth estimation. For N equal to 20 we start to see accurate estimates, highlighted in Fig. 10 (b), however most of the estimation is still not accurate. For N equal to 30 we attain an increase in correct estimations, highlighted in red in Fig. 10 (c), corresponding to the best results obtained throughout the experiment. It was also observed that each grid size yields different correct regions. This results point to the need of adapting the size and shape of the region considered in the local shearing operation.

E. Level Sets on Reconstruction Confidence

This section is dedicated the method proposed in section 4, which aims to complement the depth estimation in low gradient regions. To do so, it relies on the reconstruction confidence obtained via structure tensor to target areas for reconstruction.

a) Synthetic Data: The method was first applied to the Blender light field, to assess if this method can complement estimations in a first reconstruction. We show the obtained confidence map as well as the level curves for this data in Fig. 5(c), where the light areas correspond to confidence above a threshold and dark areas correspond to confidence below that same level.

The resulting reconstruction presents three major error areas which could be mitigated if the patches were not estimated independently.

To assess the estimation performance we calculate the mean absolute error (MAE), on the estimations without the centroid. This yields a structural MAE (SMAE) that better reflects the estimated structure. The reconstruction results are promising, the MAE value was 0.50 units (around 7 cm) reflecting an estimation offset. Regarding the SMAE, we obtained a value of 0.17 units (around 2 cm), which is justified by the lack of coherence between patches. The assumption of constant depth within the patches also drives the error up, since abrupt changes are not captured.

1) Real Data: We use the light field acquired by us, corresponding to pose 1, to test the proposed method. The obtained confidence map and corresponding level curves are depicted in Fig. 11 (a), with light areas corresponding to high confidence and dark areas to low confidence. The reconstruction results, depicted in Fig. 11 (b), possess a correct depth range and most patches possess values around to 0.7 meters, which is around the correct estimated depth. However, inconsistencies are found in some patches (see Fig. 11 (b)).

The first reconstruction, depicted in Fig. 11 (c), possesses several blank areas. These areas were complemented with our estimates, as shown in Fig. 11 (d). There, the issues with our estimation are noticeable. Despite that, the remaining estimation is coherent with the first reconstruction, blending in well. This compelling result demonstrates that we can leverage strengths from both methods.

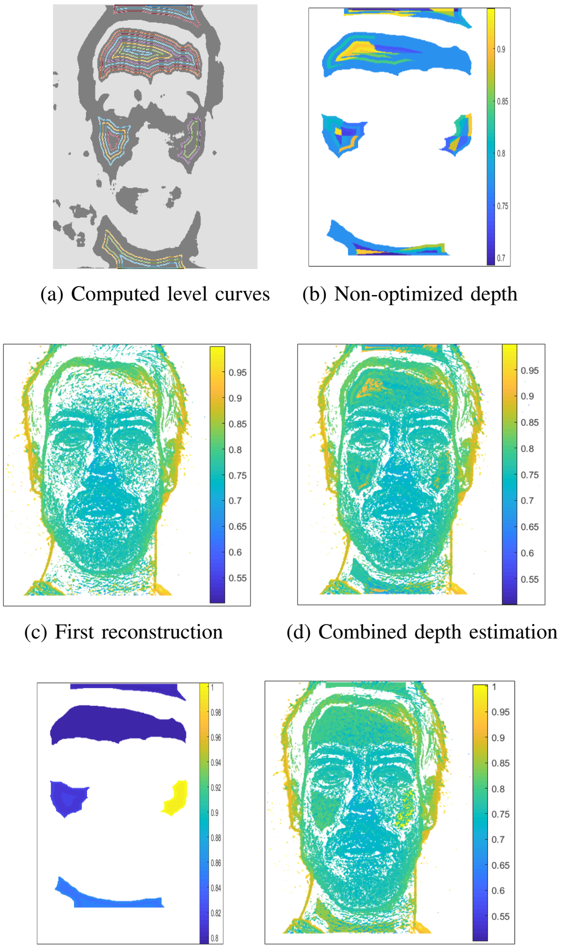
The performance of the proposed optimization was also tested, on the same data, with the best results attained for $\omega_1 = 1000$ and $\omega_2 = 1$ (see Fig. 11 (e)). The first reconstruction was complemented with this estimation. As shown in Fig. 11 (f), a correct depth range and overall coherence with the first reconstruction values was attained. Specifically, the forehead presents estimations highly consistent with the closest high confidence areas (on the hair/forehead frontier). Furthermore, the difference between forehead and neck is around 5 cm, which is consistent with real distances.

Comparing both approaches we conclude that the optimization step is valuable as it enforces consistency between patch estimations. Moreover, the complemented estimation presents a convincing result, as low confidence areas present estimates coherent with the closest high confidence areas of the first reconstruction.

VI. Conclusion and Future Work

The work described in this report comprised the study of both the plenoptic camera and the light fields acquired by it, targeting the usage of light field imagery to perform 3D reconstruction of human faces. Traditional edge-based methods struggle in low gradient areas, as demonstrated by our experiments.

A method was proposed to complement this methods' estimations where they have low confidence. Using the reconstruction confidence extracted from the structure tensor enabled targeting low confidence areas for reconstruction. The method yielded promising results, with consistent depth estimations within each region. Throughout most of this work real data was inaccessible, therefore further experimentation is required



(a) Computed level curves (b) Non-optimized depth

(c) First reconstruction (d) Combined depth estimation

(e) Optimized depth (f) Combined depth estimation

Fig. 11. Depth estimation using level sets of the reconstruction confidence on real data: (a) confidence map and level curves; (b) non-optimized estimation; (c) first reconstruction; (d) first reconstruction complemented with (b); (e) optimized estimation; (f) first reconstruction complemented with optimized estimation. [Credit for the light field acquisition: Miguel Rodrigues.]

both in regards of real data as well as a more detailed study of the optimization parameter tuning.

In future work the different integration techniques should be explored, such as inclusion of edge points in the frontier as a constraint, enabling techniques like Poisson blending [16] to be used. Furthermore, regarding plenoptic setups an effort to investigate the potential of smartphones for light field imagery acquisition [12] would allow the general public to benefit from its capabilities.

References

[1] Anna Alperovich, Ole Johannsen, Michael Strecke, and Bastian Goldluecke. Light field intrinsics with a deep encoder-decoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9145–9154, 2018.

[2] Josef Bigun. Optimal orientation detection of linear symmetry, 1987.

[3] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1027–1034, 2013.

[4] Mehmet Dikmen. 3d face reconstruction using stereo vision. *Master's Thesis, Supervisor: Ugur Halıcı, Middle East Technical University, Ankara*, 2006.

[5] Mingtao Feng, Syed Zulqarnain Gilani, Yaonan Wang, and Ajmal Mian. 3d face reconstruction from light field images: A model-free approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518, 2018.

[6] Rodrigo Ferreira and Nuno Goncalves. Fast and accurate micro lenses depth maps for multi-focus light field cameras. In *German Conference on Pattern Recognition*, pages 309–319. Springer, 2016.

[7] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*. Springer, 2016.

[8] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, et al. A taxonomy and evaluation of dense light field depth estimation algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 82–99, 2017.

[9] Numair Khan, Qian Zhang, Lucas Kasser, Henry Stone, Min H Kim, and James Tompkin. View-consistent 4d light field superpixel segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7811–7819, 2019.

[10] Simao Graça Marto, Nuno Barroso Monteiro, Joao Pedro Barreto, and José António Gaspar. Structure from plenoptic imaging. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 338–343. IEEE, 2017.

[11] Simão Marto. Structure reconstruction using plenoptic camera. Master's thesis, Instituto Superior Técnico, University of Lisbon, 11 2017.

[12] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[13] Nuno Barroso Monteiro, Joao P Barreto, and José António Gaspar. Standard plenoptic cameras mapping to camera arrays and calibration based on dlt. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[14] Nuno Barroso Monteiro, Joao Pedro Barreto, and José Gaspar. Dense lightfield disparity estimation using total variation regularization. In *International Conference on Image Analysis and Recognition*, pages 462–469. Springer, 2016.

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

[16] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[17] NG Ren. Digital light field photography. *Ph. D. thesis Stanford University*, 2006.

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[19] Alireza Sepas-Moghaddam, Valeria Chiesa, Paulo Lobato Correia, Fernando Pereira, and Jean-Luc Dugelay. The ist-eurecom light field face database. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2017.

[20] Alireza Sepas-Moghaddam, Fernando Pereira, and Paulo Lobato Correia. Light field-based face presentation attack detection: reviewing, benchmarking and one step further. *IEEE Transactions on Information Forensics and Security*, 13(7):1696–1709, 2018.

[21] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[22] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2012.

[23] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020.

[24] Jingyi Yu, Leonard McMillan, and Steven Gortler. Surface camera (scam) light field rendering. *International Journal of Image and Graphics*, 4(04):605–625, 2004.

[25] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016.

[26] Hao Zhu, Qi Zhang, and Qing Wang. 4d light field superpixel and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6384–6392, 2017.