Reliable Planar Object Pose Estimation in Light Fields From Best Subaperture Camera Pairs

Nathan Crombez, Guillaume Caron, Takuya Funatomi, Yasuhiro Mukaigawa

Reliable planar object pose estimation in light-fields from best sub-aperture camera pairs

Nathan Crombez[1], Guillaume Caron[2], Takuya Funatomi[3] and Yasuhiro Mukaigawa[3]

***Abstract*—A light-field camera can obtain richer information about a scene than a usual camera. This property offers a lot of potential for robot vision. In this paper, we present a method for pose estimation of a planar object with a light-field camera.**

The light-field camera can be regarded as a set of sub-aperture cameras. Although any combination of them can theoretically be used for the pose estimation, the accuracy depends on the combination. We show that the estimated pose error can be reduced by selecting the best pair of sub-aperture cameras. We have evaluated the accuracy of our approach with real experiments using a light-field camera in front of planar targets held by an industrial manipulator for ground truth comparison.

***Index Terms*—Visual Tracking, Computer Vision for Other Robotic Applications.**

I. Introduction

A. Motivation

The pose estimation of an object is a fundamental task for a variety of purposes in the robotics field such as 3-D tracking [1], visual servoing [2], and motion estimation [3]. When the target is a known planar object such as a checkerboard, a single camera can be used for this purpose, and a lot of geometric algorithms have been proposed. The Zhang's algorithm [4] is a standard algorithm of intrinsic parameters estimation, based on corner feature points, and it also provides the poses of the checkerboard for each input views. A binocular stereo pair might also be considered for increasing the accuracy [5] and relaxing all the knowledge about the object dimensions, except it is planar.

However, Hartley and Zisserman [5] highlight that some feature points locations with respect to the pair of images leads to poor estimates of the plane parameters and pose, usually combined as a homography. More precisely, "image points [...] close to collinear with the epipole [leads to] poorly conditioned estimate" [5, Sec. 13.2.1, p. 330]. The latter issue is jointly raised by the points configuration and the cameras configuration, *i.e.* their relative pose. The problem of points configuration for the robust pose estimation of a planar object has very recently been studied in the literature in order to consider configurations that ensure a precise estimation [6].

In this work, we propose a complementary approach in which we aim at selecting the best pair of cameras to ensure a precise estimation of the planar object pose. The selection of a pair of cameras among many is done in this work thanks to light-field (LF) imaging, contrary to the classical use of a stereo rig.

B. Related works

Recently, a LF camera can be used to obtain richer information of a scene, such as 3-D reconstruction [7], [8], instead of usual 2-D cameras. The information obtained by the LF camera is inherently equivalent to a set of images captured from different viewpoints, so-called sub-aperture images (SAI). SAIs, made from the transformation of the image taken by the LF camera or made from several actual camera arrays are full of interest in robotics [9], [10]. This is due to the high redundancy and strong constraints brought by the configuration of sub-aperture cameras (SAC) in a LF camera, different from general multi-view settings.

The constraints inherent to SAIs have been considered in general Structure-From-Motion [11] with the point feature. Zhang *et al.* [12] recently extended the research to the line feature. The latter work also introduces the modeling of planar constraint on feature point manifolds. However, that theoretical modeling is not considered in any motion or structure estimation.

C. Contributions

In this paper, we propose the detailed modeling of relationships between feature points belonging to a planar area of the scene, observed by one and two LF cameras. Then, linear and non-linear estimation algorithms of planar object pose are proposed, exploiting the minimum number of SAIs. To reach precise planar object pose estimation, we propose a new algorithm to select the best pair of SAIs. This selection avoids poor conditioning in the estimation and saves processing time, with respect to a greedy estimation that would consider every available SAI.

The contributions of the article are summed up below:

- The expression of the homography between corresponding rays of a LF.
- The exploitation of the LF camera geometry to simplify the homography content between two SAIs of a LF, used for plane parameters estimation.

Manuscript received: February, 24th, 2018; Revised May, 20th, 2018; Accepted June, 18th, 2018.

This paper was recommended for publication by Editor François Chaumette upon evaluation of the Associate Editor and Reviewers' comments.

[1]Nathan Crombez is with the Le2i laboratory, ERL VIBOT CNRS 6000, Université de Bourgogne Franche-Comté, 71200 Le Creusot, France nathan.crombez@u-bourgogne.fr

[2]Guillaume Caron is with the MIS laboratory, EA 4290, Université de Picardie Jules Verne, 80000 Amiens, France guillaume.caron@u-picardie.fr

[3]Takuya Funatomi and Yasuhiro Mukaigawa are with the OMI laboratory, Nara Institue of Science and Technology, Nara 630-0192, Japan funatomi@is.naist.jp, mukaigawa@is.naist.jp

Digital Object Identifier (DOI): see top of this page.

- The proposition of a method to select which pair of SAIs is the best for plane estimation.
- The exploitation of the best pairs of SAIs in two different LFs to estimate the planar object motion.
- The efficiency of the approach with a low processing time while being precise.

Combining the latter contributions leads to the reliable estimation of planar object poses over time in LF sequences.

Theoretical and methodological contributions are deeply evaluated on real LF images with respect to the robotic arm-based acquired ground truth. Comparisons are also done with respect to the extrinsic parameters estimation of a LF camera calibration toolbox [13], when considering a checkerboard as planar object, and with respect to the standard binocular stereo pair (*e.g.*: two horizontally aligned side-by-side cameras), when considering a textured planar object.

II. Light-field camera model

This section recalls how the image data acquired by a LF camera is transformed to the 4-D structure that parameterizes the acquisition. Throughout this paper, we use $\mathbf{x} \in \mathbb{R}^{n_d}$, a vector of Euclidean coordinates in the n_d-dimensions (n_d-D) space and $\tilde{\mathbf{x}} \in \mathbb{P}^{n_d}$, its homogeneous equivalent. Transformations from Euclidean to homogeneous and vice versa are omitted since they are very common.

In this paper, we consider the model of a LF camera introduced by Dansereau *et al.* [13]. The latter model is also considered in several computer vision applications [11], [9], [14]. This LF camera model describes how light rays emitted by a 3-D point $\mathbf{P} = [X, Y, Z]^\top$are observed by the camera. The local two-plane parameterization (L2PP) describes a light ray as its intersections with two parallel planes, *i.e.* the focal plane Π and the image plane. The latter is considered the closest to the 3D point and the focal plane the farthest.

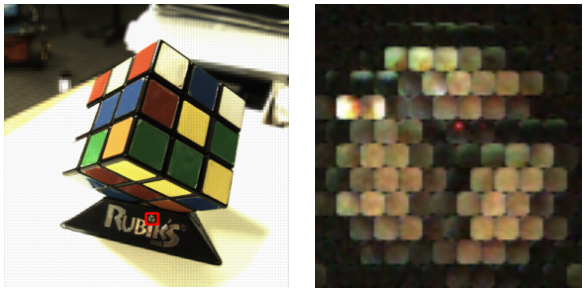
The LF camera coordinates frame $\mathcal{F}_0$ is attached to the focal plane. Then, $\mathbf{o} = [m, w]^\top$ are the coordinates, expressed in the global frame $\mathcal{F}_0$, of the intersection between the ray and the focal plane. The intersection of the light ray with the image plane gives two more coordinates, $\mathbf{p} = [x, y]^\top$, but expressed relatively to $\mathbf{o}$. Stacking both coordinates leads to the ray:

$$\tilde{\boldsymbol{\phi}} = [m, w, x, y, 1]^\top, \tag{1}$$

which fully describes both spatial ($\mathbf{o}$) and angular ($\mathbf{p}$) information about the incident light ray.

However, physically, the LF camera has a unique photosensitive matrix. So, the four parameters of (1) are acquired in the same raw real image plane. Figure 1 shows a subpart of such a raw image on which one can see that it contains a set of small lenslet images.

Hence, a transformation is needed to get the 4-D $\tilde{\boldsymbol{\phi}}$ from pixels of the raw acquired image. To do so, two main stages are required. First, omitting image processing of the raw image to segment lenslets [13], the initial common 2-D raw pixels structure is transformed to a 4-D structure (Fig. 2a). The latter organizes raw image pixel coordinates in order to build subsets containing one pixel, of coordinates $\mathbf{i} = [i, j]^\top$, of each lenslet of central coordinates $\mathbf{k} = [k, l]^\top$, with respect to which $\mathbf{i}$ is



(a) (b)

Fig. 1: LF camera acquisition: a complete raw LF image (a) and a zoomed subpart of this LF (b).

expressed. Stacking $\mathbf{i}$ and $\mathbf{k}$ leads to the decoded LF image "pixel" coordinates:

$$\tilde{\mathbf{r}} = [i, j, k, l, 1]^\top. \tag{2}$$

Then, the 4-D structure is easily re-organized from $\tilde{\mathbf{r}}$ coordinates in order to build SAIs, as shown in Figure 2b.

As, in the rest of the article, we focus on considering $\tilde{\boldsymbol{\phi}}$'s and not $\tilde{\mathbf{r}}$'s, $\tilde{\mathbf{r}}$ is transformed to $\tilde{\boldsymbol{\phi}}$ thanks to the LF camera intrinsic parameters matrix $\boldsymbol{\mathcal{K}}$ such that:

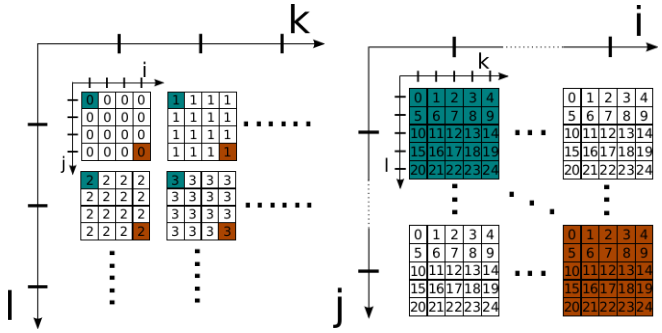
$$\begin{bmatrix} m \\ w \\ x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \mathcal{K}_{1,1} & 0 & \mathcal{K}_{1,3} & 0 & \mathcal{K}_{1,5} \\ 0 & \mathcal{K}_{2,2} & 0 & \mathcal{K}_{2,4} & \mathcal{K}_{2,5} \\ \mathcal{K}_{3,1} & 0 & \mathcal{K}_{3,3} & 0 & \mathcal{K}_{3,5} \\ 0 & \mathcal{K}_{4,2} & 0 & \mathcal{K}_{4,4} & \mathcal{K}_{4,5} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \\ l \\ 1 \end{bmatrix}. \tag{3}$$

The twelve non-zero elements of the intrinsic matrix $\boldsymbol{\mathcal{K}}$ are estimated thanks to the LF camera calibration toolbox[1] implementing [13].

Finally, since a LF camera captures several light rays emitted by a unique 3-D point $\mathbf{P}$, there are as much $\tilde{\boldsymbol{\phi}}$'s corresponding to $\mathbf{P}$ as there are SACs seeing $\mathbf{P}$. Hence, we indicate the SAC which, virtually, acquires the ray as superscript left to $\tilde{\boldsymbol{\phi}}$, *i.e.* ${}^{c_a}\tilde{\boldsymbol{\phi}}$, for SAC c_a of frame $\mathcal{F}_{c_a}$, ${}^{c_b}\tilde{\boldsymbol{\phi}}$, for SAC c_b of frame $\mathcal{F}_{c_b}$, and so on. More precisely, we have:

$${}^{c_a}\boldsymbol{\phi} = [{}^0m^{c_a}, {}^0w^{c_a}, {}^{c_a}x, {}^{c_a}y, 1]^\top, \tag{4}$$

[1] http://dgd.vision/Tools/LFToolbox



(a) (b)

Fig. 2: (a) Decoded data structure (4-D). (b) SAIs (each bloc of figures from 0 to 24 is a SAI) generated from the decoded 4-D data structure. This figure is a rewriting of Fig. 6 in [8].

considering the coordinates of SAC optical centers are expressed in the global frame $\mathcal{F}_0$ (*e.g.* ${}^0\mathbf{o}^{c_a} = [{}^0m^{c_a}, {}^0w^{c_a}]$, regarding camera c_a). In (4), ${}^{c_a}\mathbf{p} = [{}^{c_a}x, {}^{c_a}y]^\top$, is nothing but a precision, since coordinates x and y are expressed relatively to ${}^0\mathbf{o}^{c_a}$ (1).

***Remark** 1 (Coplanarity of SACs):* By construction from the L2PP parameterization, optical centers of SACs belong to the same plane Π (the focal plane). That is why these optical centers coordinates are expressed in 2-D in $\mathcal{F}_0$.

III. Multiple view geometric relationships in a light-field under planar constraint

A LF camera captures several projections of the same 3-D point in one shot. Thus, we propose to consider an observed planar object and to study how its parameters are involved in SAIs relationships. First, we describe how feature correspondences in LF structures are expressed. Then, we introduce for the first time, to our knowledge, the homography matrix related to light rays.

A. Light ray correspondences

In an image captured with a regular camera, a single 3-D point ${}^{c_a}\mathbf{P} = [{}^{c_a}X, {}^{c_a}Y, {}^{c_a}Z]^\top$, expressed in the c_a camera frame $\mathcal{F}_{c_a}$, is projected onto a unique 2D point ${}^{c_a}\mathbf{p} = [{}^{c_a}x, {}^{c_a}y]^\top$ in the sensor frame (${}^{c_a}\mathbf{k} = [{}^{c_a}k, {}^{c_a}l]^\top$ in the 4-D structure, see (2), adding superscript c_a to show to which SAC it belongs to). The usual input to many computer vision algorithms, for object pose estimation, is a set of 2D points correspondences ${}^{c_a}\mathbf{k}$ and ${}^{c_b}\mathbf{k}$ (from a second camera c_b) between two (or more) images.

In a calibrated LF camera, correspondences between two SACs c_a and c_b images are rays, parameterized as 4-D (Sec. II), below written in the L2PP modeling and as homogeneous coordinates:

$$ {}^{c_a}\tilde{\boldsymbol{\phi}} \;\leftrightarrow\; {}^{c_b}\tilde{\boldsymbol{\phi}} \; . \tag{5}$$

As in classical multiple view computer vision, many correspondences are usually available, *i.e.* ${}^{c_a}\tilde{\boldsymbol{\phi}}_1 \leftrightarrow {}^{c_b}\tilde{\boldsymbol{\phi}}_1$, ${}^{c_a}\tilde{\boldsymbol{\phi}}_2 \leftrightarrow {}^{c_b}\tilde{\boldsymbol{\phi}}_2$, and so on. Indices are not used in every equation of the paper to limit heavy mathematical writings.

B. Rays homography

Corresponding $\tilde{\boldsymbol{\phi}}$'s are issued from the same 3-D point $\mathbf{P}$. When $\mathbf{P}$ is lying on a planar object, its rays captured by the LF camera follows a homography projective transformation. Considering the common multiple view computer vision, we have, for two camera frames $\mathcal{F}_{c_a}$ and $\mathcal{F}_{c_b}$, the projective homography matrix ${}^{c_b}\mathbf{H}_{c_a(3\times3)}$ transforming ${}^{c_a}\mathbf{P}$ to ${}^{c_b}\mathbf{P}$, up to a scale factor, is:

$$ {}^{c_b}\mathbf{P} \propto {}^{c_b}\mathbf{H}_{c_a(3\times3)} \; {}^{c_a}\mathbf{P} \; . \tag{6}$$

Then, as a 3D point ${}^{c_a}\mathbf{P}$ is projected in the sensor frame of camera c_a following the perspective projection:

$$ {}^{c_a}\tilde{\mathbf{p}} = \lambda^{c_a} \; {}^{c_a}\mathbf{P} \iff \begin{bmatrix} {}^{c_a}x \\ {}^{c_a}y \\ 1 \end{bmatrix} = \lambda^{c_a} \begin{bmatrix} {}^{c_a}X \\ {}^{c_a}Y \\ {}^{c_a}Z \end{bmatrix} \; , \tag{7}$$

where $\lambda^{c_a} = \frac{1}{{}^{c_a}Z}$, and similarly done for ${}^{c_b}\mathbf{P}$, corresponding $\tilde{\mathbf{p}}$'s are related by the same ${}^{c_b}\mathbf{H}_{c_a}$:

$$ {}^{c_b}\tilde{\mathbf{p}} \propto {}^{c_b}\mathbf{H}_{c_a} \; {}^{c_a}\tilde{\mathbf{p}} $$
$$ \begin{bmatrix} {}^{c_b}x \\ {}^{c_b}y \\ 1 \end{bmatrix} \propto \begin{bmatrix} H_1 & H_2 & H_3 \\ H_4 & H_5 & H_6 \\ H_7 & H_8 & 1 \end{bmatrix} \begin{bmatrix} {}^{c_a}x \\ {}^{c_a}y \\ 1 \end{bmatrix} \; . \tag{8}$$

The analytical expression of ${}^{c_b}\mathbf{H}_{c_a}$ is known to involve the rotation matrix ${}^{c_b}\mathbf{R}_{c_a}$, belonging to the SO(3) group of the Lie algebra, the translation vector ${}^{c_b}\mathbf{t}_{c_a(3\times1)} = [{}^{c_b}t^X_{c_a}, {}^{c_b}t^Y_{c_a}, {}^{c_b}t^Z_{c_a}]^\top$, belonging to $\mathbb{R}^3$ between the frame $\mathcal{F}_{c_a}$ and the frame $\mathcal{F}_{c_b}$. It also involves the parameters of the planar surface on which the 3-D point belongs to [5]:

$$ {}^{c_b}\mathbf{H}_{c_a} = {}^{c_b}\mathbf{R}_{c_a} - \frac{1}{{}^{c_a}d} {}^{c_b}\mathbf{t}_{c_a} \; {}^{c_a}\mathbf{n}^\top \; , \tag{9}$$

where ${}^{c_a}\mathbf{n} = [{}^{c_a}n^X, {}^{c_a}n^Y, {}^{c_a}n^Z]^\top$ is the normal vector of the planar surface expressed in $\mathcal{F}_{c_a}$ and ${}^{c_a}d$ is the orthogonal distance between the planar surface and the origin of the camera c_a. Taking (8) and (9) together leads to the general analytical expression of each $H_{\{1,\dots,8\}}$.

Whereas homographies above are, as often considered, 2-D, $\tilde{\boldsymbol{\phi}}$'s are 4-D. Thus, as ${}^{c_a}\tilde{\boldsymbol{\phi}}$ and ${}^{c_b}\tilde{\boldsymbol{\phi}}$ are rays captured by SACs, crossing at $\mathbf{P}$, we extend (8) to rays as:

$$ {}^{c_b}\tilde{\boldsymbol{\phi}} \propto {}^{c_b}\boldsymbol{\mathcal{H}}_{c_a} \; {}^{c_a}\tilde{\boldsymbol{\phi}} $$
$$ \begin{bmatrix} {}^0m^{c_b} \\ {}^0w^{c_b} \\ {}^{c_b}x \\ {}^{c_b}y \\ 1 \end{bmatrix} \propto \begin{bmatrix} \mathcal{H}_1 & \mathcal{H}_2 & \mathcal{H}_3 & \mathcal{H}_4 & \mathcal{H}_5 \\ \mathcal{H}_6 & \mathcal{H}_7 & \mathcal{H}_8 & \mathcal{H}_9 & \mathcal{H}_{10} \\ \mathcal{H}_{11} & \mathcal{H}_{12} & \mathcal{H}_{13} & \mathcal{H}_{14} & \mathcal{H}_{15} \\ \mathcal{H}_{16} & \mathcal{H}_{17} & \mathcal{H}_{18} & \mathcal{H}_{19} & \mathcal{H}_{20} \\ \mathcal{H}_{21} & \mathcal{H}_{22} & \mathcal{H}_{23} & \mathcal{H}_{24} & \mathcal{H}_{25} \end{bmatrix} \begin{bmatrix} {}^0m^{c_a} \\ {}^0w^{c_a} \\ {}^{c_a}x \\ {}^{c_a}y \\ 1 \end{bmatrix} \; . \tag{10}$$

We can identify in (10) the presence of the (8) which gives the following nine elements of ${}^{c_b}\boldsymbol{\mathcal{H}}_{c_a}$:

$$ \begin{aligned} \mathcal{H}_{13} &= H_1 & \mathcal{H}_{14} &= H_2 & \mathcal{H}_{15} &= H_3 \\ \mathcal{H}_{18} &= H_4 & \mathcal{H}_{19} &= H_5 & \mathcal{H}_{20} &= H_6 \\ \mathcal{H}_{23} &= H_7 & \mathcal{H}_{24} &= H_8 & \mathcal{H}_{25} &= 1 \end{aligned} $$

The latter elements of ${}^{c_b}\boldsymbol{\mathcal{H}}_{c_a}$ only relate the third to fifth coordinates of $\tilde{\boldsymbol{\phi}}$'s. Their first and second coordinates, ${}^0\mathbf{o}^{c_a}$ and ${}^0\mathbf{o}^{c_b}$, correspond to optical centers coordinates of SACs in the LF camera frame $\mathcal{F}_0$. Since, optical centers of SACs belong to the same plane (Remark 1), and assuming optical axes of SACs are parallel (assumption discussed in Remark 2, below), the transformation between two SACs is a translation of length ${}^{c_b}t^X_{c_a}$ and ${}^{c_b}t^Y_{c_a}$ along the two axes of the focal plane.

That translations between $\mathcal{F}_{c_a}$ and $\mathcal{F}_{c_b}$ can be written w.r.t the global frame $\mathcal{F}_0$:

$$ {}^{c_b}t^X_{c_a} = {}^{c_b}t^X_0 + {}^0t^X_{c_a} = {}^0t^X_{c_a} + {}^0t^X_{c_b} = {}^0m^{c_a} - {}^0m^{c_b} \tag{11}$$

$$ {}^{c_b}t^Y_{c_a} = {}^{c_b}t^Y_0 + {}^0t^Y_{c_a} = {}^0t^Y_{c_a} + {}^0t^Y_{c_b} = {}^0w^{c_a} - {}^0w^{c_b} \; . \tag{12}$$

Equations (11) and (12) provide alone the useful missing relationships to set the remaining sixteen elements of ${}^{c_b}\boldsymbol{\mathcal{H}}_{c_a\,(5\times5)}$:

$$ {}^{c_b}\boldsymbol{\mathcal{H}}_{c_a\,(5\times5)} = \begin{bmatrix} 1 & 0 & 0 & 0 & {}^{c_b}t^X_{c_a} \\ 0 & 1 & 0 & 0 & {}^{c_b}t^Y_{c_a} \\ 0 & 0 & H_1 & H_2 & H_3 \\ 0 & 0 & H_4 & H_5 & H_6 \\ 0 & 0 & H_7 & H_8 & 1 \end{bmatrix}. \tag{13} $$

Furthermore, assumptions considered to establish (11) and (12) lead to the expressions of ${}^{c_b}\mathbf{R}_{c_a}$ and ${}^{c_b}\mathbf{t}_{c_a}$:

$$ {}^{c_b}\mathbf{R}_{c_a}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, {}^{c_b}\mathbf{t}_{c_a}=\begin{bmatrix} {}^{c_b}t^X_{c_a} \\ {}^{c_b}t^Y_{c_a} \\ 0 \end{bmatrix}=\begin{bmatrix} {}^0m^{c_a} - {}^0m^{c_b} \\ {}^0w^{c_a} - {}^0w^{c_b} \\ 0 \end{bmatrix}, \tag{14} $$

thus simplifying ${}^{c_b}\mathbf{H}_{c_a}$ (9) as:

$$ \begin{bmatrix} 1 - \frac{{}^{c_b}t^X_{c_a}\,{}^{c_a}n^X}{{}^{c_a}d} & -\frac{{}^{c_b}t^X_{c_a}\,{}^{c_a}n^Y}{{}^{c_a}d} & -\frac{{}^{c_b}t^X_{c_a}\,{}^{c_a}n^Z}{{}^{c_a}d} \\ -\frac{{}^{c_b}t^Y_{c_a}\,{}^{c_a}n^X}{{}^{c_a}d} & 1 - \frac{{}^{c_b}t^Y_{c_a}\,{}^{c_a}n^Y}{{}^{c_a}d} & -\frac{{}^{c_b}t^Y_{c_a}\,{}^{c_a}n^Z}{{}^{c_a}d} \\ 0 & 0 & 1 \end{bmatrix}. \tag{15} $$

Therefore ${}^{c_b}\boldsymbol{\mathcal{H}}_{c_a}$ becomes:

$$ \begin{bmatrix} 1 & 0 & 0 & 0 & {}^{c_b}t^X_{c_a} \\ 0 & 1 & 0 & 0 & {}^{c_b}t^Y_{c_a} \\ 0 & 0 & 1 - \frac{{}^{c_b}t^X_{c_a}\,{}^{c_a}n^X}{{}^{c_a}d} & -\frac{{}^{c_b}t^X_{c_a}\,{}^{c_a}n^Y}{{}^{c_a}d} & -\frac{{}^{c_b}t^X_{c_a}\,{}^{c_a}n^Z}{{}^{c_a}d} \\ 0 & 0 & -\frac{{}^{c_b}t^Y_{c_a}\,{}^{c_a}n^X}{{}^{c_a}d} & 1 - \frac{{}^{c_b}t^Y_{c_a}\,{}^{c_a}n^Y}{{}^{c_a}d} & -\frac{{}^{c_b}t^Y_{c_a}\,{}^{c_a}n^Z}{{}^{c_a}d} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{16} $$

***Remark** 2 (Parallelism of SAC optical axes):* From (11), SAC optical axes are considered to be parallel. Although it allows reaching (16) simple as it is, the assumption reduces the rest of the modeling and planar object pose estimation (Sec. IV and V) to the class of LF camera devices following a design having the same optical properties as the Lytro Photo LF camera device [15]. In the latter, lenses of the considered micro-lenses array, to make the camera a LF one, are coplanar and have their optical axes parallel. The latter design fact leads to the parallelism of SAC optical axes. Several other LF acquisition designs, including multi-camera ones [16], can also benefit from the method we propose.

IV. Plane parameters estimation

A. Linear estimation

Considering the analytical expression of ${}^{c_b}\boldsymbol{\mathcal{H}}_{c_a}$ in (16) and after some manipulations and rearrangements of (10), we can collect the plane parameters ${}^{c_a}\mathbf{n}$ and ${}^{c_a}d$ on one side:

$$ \begin{bmatrix} {}^{c_b}x_1 - {}^{c_a}x_1 \\ {}^{c_b}y_1 - {}^{c_a}y_1 \\ \vdots \\ {}^{c_b}x_* - {}^{c_a}x_* \\ {}^{c_b}y_* - {}^{c_a}y_* \end{bmatrix} = \begin{bmatrix} {}^{c_b}t^X_{c_a}\,{}^{c_a}x_1 & {}^{c_b}t^X_{c_a}\,{}^{c_a}y_1 & {}^{c_b}t^X_{c_a} \\ {}^{c_b}t^Y_{c_a}\,{}^{c_a}x_1 & {}^{c_b}t^Y_{c_a}\,{}^{c_a}y_1 & {}^{c_b}t^Y_{c_a} \\ \vdots & \vdots & \vdots \\ {}^{c_b}t^X_{c_a}\,{}^{c_a}x_* & {}^{c_b}t^X_{c_a}\,{}^{c_a}y_* & {}^{c_b}t^X_{c_a} \\ {}^{c_b}t^Y_{c_a}\,{}^{c_a}x_* & {}^{c_b}t^Y_{c_a}\,{}^{c_a}y_* & {}^{c_b}t^Y_{c_a} \end{bmatrix} \begin{bmatrix} \frac{{}^{c_a}n^X}{{}^{c_a}d} \\ \frac{{}^{c_a}n^Y}{{}^{c_a}d} \\ \frac{{}^{c_a}n^Z}{{}^{c_a}d} \end{bmatrix} \tag{17} $$

where ${}^{c_b}t^X_{c_a} = ({}^0m^{c_a} - {}^0m^{c_b})$, ${}^{c_b}t^Y_{c_a} = ({}^0w^{c_a} - {}^0w^{c_b})$ as seen in (11) and (12). Therefore, the plane parameters can be computed directly from a set of at least 3 light rays correspondences ($* >= 3$) following:

$$ {}^{c_a}\boldsymbol{\eta}= \begin{bmatrix} {}^{c_b}t^X_{c_a}\,{}^{c_a}x_1 & {}^{c_b}t^X_{c_a}\,{}^{c_a}y_1 & {}^{c_b}t^X_{c_a} \\ {}^{c_b}t^Y_{c_a}\,{}^{c_a}x_1 & {}^{c_b}t^Y_{c_a}\,{}^{c_a}y_1 & {}^{c_b}t^Y_{c_a} \\ \vdots & \vdots & \vdots \\ {}^{c_b}t^X_{c_a}\,{}^{c_a}x_* & {}^{c_b}t^X_{c_a}\,{}^{c_a}y_* & {}^{c_b}t^X_{c_a} \\ {}^{c_b}t^Y_{c_a}\,{}^{c_a}x_* & {}^{c_b}t^Y_{c_a}\,{}^{c_a}y_* & {}^{c_b}t^Y_{c_a} \end{bmatrix}^+ \begin{bmatrix} {}^{c_b}x_1 - {}^{c_a}x_1 \\ {}^{c_b}y_1 - {}^{c_a}y_1 \\ \vdots \\ {}^{c_b}x_* - {}^{c_a}x_* \\ {}^{c_b}y_* - {}^{c_a}y_* \end{bmatrix} \tag{18} $$

where superscript $^+$ denotes the pseudo-inverse and ${}^{c_a}\boldsymbol{\eta} = [\frac{{}^{c_a}n^X}{{}^{c_a}d}\frac{{}^{c_a}n^Y}{{}^{c_a}d}\frac{{}^{c_a}n^Z}{{}^{c_a}d}]^\top$. Then the parameters of the planar object w.r.t. the camera c_a can be obtained from ${}^{c_a}\boldsymbol{\eta}$ following:

$$ {}^{c_a}d = \frac{1}{\|{}^{c_a}\boldsymbol{\eta}\|} \tag{19} $$

$$ {}^{c_a}\boldsymbol{n} = {}^{c_a}d\;{}^{c_a}\boldsymbol{\eta}. \tag{20} $$

B. Non-linear optimization

To estimate ${}^{c_a}\boldsymbol{\eta}$ more robustly and to allow the method to be extended to relative pose estimation between two LFs (Sec. V), we formulate its estimation as a non-linear optimization of which the initial guess is provided by the linear estimation (Sec. IV-A).

As ${}^{c_b}\mathbf{H}_{c_a}$ in (15) depends on ${}^{c_a}\boldsymbol{\eta}$, we rewrite it as ${}^{c_b}\mathbf{H}_{c_a}({}^{c_a}\boldsymbol{\eta})$. Considering Q corresponding points, the optimization problem to solve is:

$$ {}^{c_a}\hat{\boldsymbol{\eta}} = \min_{{}^{c_a}\boldsymbol{\eta}} \sum_{q\,=\,0}^{Q} ||{}^{c_b}\tilde{\mathbf{p}}^*_q \ominus {}^{c_b}\tilde{\mathbf{p}}_q||, \tag{21} $$

where the $\ominus$ operator denotes the difference after resetting the homogenous coordinate at 1 and with:

$$ {}^{c_b}\tilde{\mathbf{p}}^*_q = {}^{c_b}\mathbf{H}_{c_a}({}^{c_a}\boldsymbol{\eta})\,{}^{c_a}\tilde{\mathbf{p}}_q, \tag{22} $$

${}^{c.}\tilde{\mathbf{p}}_q$ being the measured point coordinates in the image of SAC $c.$ and ${}^{c.}\tilde{\mathbf{p}}^*_q$ a computed one.

A gradient descent-like algorithm (Gauss-Newton in this paper) solves (21). It requires to express the Jacobian matrix of image points with respect to plane parameters, *i.e.*:

$$ \mathbf{J}_{\mathcal{F}_1,\mathcal{F}_2} = \frac{\partial\,{}^{\mathcal{F}_2}\mathbf{p}^*}{\partial\,{}^{\mathcal{F}_1}\boldsymbol{\eta}}, \tag{23} $$

where the index q is omitted for compactness. $\mathcal{F}_1$ and $\mathcal{F}_2$ denote generic coordinate systems of the homography relationship:

$$ {}^{\mathcal{F}_2}\tilde{\mathbf{p}}^* \propto {}^{\mathcal{F}_2}\mathbf{H}_{\mathcal{F}_1}\,{}^{\mathcal{F}_1}\tilde{\mathbf{p}} \tag{24} $$

similarly to (8).

Thus, the detailed generic expression of $\mathbf{J}_{\mathcal{F}_1,\mathcal{F}_2}$ is (with $\mathcal{F}_1$ omitted, again for compactness):

$$ \mathbf{J}_{\mathcal{F}_1,\mathcal{F}_2}=\frac{1}{\left((\mathbf{R}_3 - \boldsymbol{\eta}^\top t^Z)\tilde{\mathbf{p}}\right)^2}\begin{bmatrix} \left((\mathbf{R}_1 t^Z - \mathbf{R}_3 t^X)\tilde{\mathbf{p}}\right)\tilde{\mathbf{p}}^\top \\ \left((\mathbf{R}_2 t^Z - \mathbf{R}_3 t^Y)\tilde{\mathbf{p}}\right)\tilde{\mathbf{p}}^\top \end{bmatrix}, \tag{25} $$

with $\mathbf{R}_{\cdot}$, the $\cdot$-th row of the rotation matrix between $\mathcal{F}_1$ and $\mathcal{F}_2$, and t^X, t^Y and t^Z the translation vector components, between $\mathcal{F}_1$ and $\mathcal{F}_2$ as well.

Considering only two SACs c_a and c_b of the same LF, $\mathbf{J}_{\mathcal{F}_1,\mathcal{F}_2}$ is simplified as $\mathbf{J}_{c_a,\ c_b}$:

$$\mathbf{J}_{c_a,\ c_b} = -\begin{bmatrix} {}^{c_b}t^X_{c_a} & {}^{c_a}\tilde{\mathbf{p}}^\top \\ {}^{c_b}t^Y_{c_a} & {}^{c_a}\tilde{\mathbf{p}}^\top \end{bmatrix}, \tag{26}$$

due to the same SACs orientation and the coplanarity of their optical centers (14).

As much $\mathbf{J}_{c_a,\ c_b}$ are computed as there are point pairs between SAIs of c_a and c_b. These Jacobians are stacked as matrix $\mathbf{J}$, considered in the plane parameters increment ${}^{c_a}\dot{\boldsymbol{\eta}}$ computation of an optimization iteration:

$${}^{c_a}\dot{\boldsymbol{\eta}} = -\mathbf{J}^+\mathbf{e}, \tag{27}$$

where $\mathbf{e}$ is the error vector stacking every ${}^{c_b}\mathbf{p}^*_q - {}^{c_b}\mathbf{p}_q$, for each point pair q. Then, ${}^{c_a}\dot{\boldsymbol{\eta}}$ updates ${}^{c_a}\boldsymbol{\eta}$ by addition, allowing to compute a new ${}^{c_b}\mathbf{H}_{c_a}$ and new ${}^{c_b}\tilde{\mathbf{p}}^*_q$ values, for all q, and this process loops until convergence.

C. The best SACs pair to estimate plane parameters

As introduced in Section I, classical stereo-vision would have, by definition, only two views at once of the same scene, or planar object, to be considered in the estimation of plane parameters. The LF camera, by providing many views of the scene at once, offers a new possibility: choosing which pair of SACs to consider in two views-based computer vision algorithms. Two criteria are considered below to make that choice.

First, as well known, the widest baseline between SACs, however allowing feature points matching, should be considered to solve (18) and (21) in order to reduce uncertainties due to slight imprecisions of $\tilde{\phi}$ values allowing to get reliable estimates of the plane parameters.

Furthermore, one should also avoid solving (18) and (21) if $\tilde{\phi}$'s are collinear, or close to be, with the epipoles of both SAIs [5]. The key idea is, then, to choose a pair of SACs for which the set of $\tilde{\phi}$ pairs, to be considered in (18) and (21), is far from being collinear with SACs epipoles.

Combining both baseline and non-epipolar collinearity constraints leads to following best SACs pair selection strategy to get reliable plane parameters estimations:

- if the center of gravity of the SAIs $\tilde{\phi}$ pairs is in the top-left corner of the LF camera field of view, then choose top-right and bottom-left SACs to solve (18) and (21) (similar way for the three other field of view corners)
- if the center of gravity of the SAIs $\tilde{\phi}$ pairs is in the top-center of the LF camera field of view, then choose bottom-right and bottom-left SACs to solve (18) and (21) (similar way for the bottom-center, left-middle and right-middle areas of the LF camera field of view)

The latter selection scheme is experimentally validated in Section VI-B1 and estimations compared to the greedy approach of stacking every available SAC.

V. Planar object pose estimation from two LFs

$\mathbf{LF}_1$ and $\mathbf{LF}_2$ are two light-fields. A SAC c_a of $\mathbf{LF}_1$ is written c_{a_1} and we write similarly for $\mathbf{LF}_2$ as well as c_b.

The goal is to exploit the planar geometry of the observed object to compute the frame change between two of its poses or, equivalently, ${}^{c_{a_2}}\mathbf{M}_{c_{a_1}}$, the rigid transformation matrix between $\mathcal{F}_{c_{a_1}}$ and $\mathcal{F}_{c_{a_2}}$. The constant frame change ${}^{c_a}\mathbf{M}_0$ from SAC frame $\mathcal{F}_{c_a}$ to the $\mathbf{LF}$ frame $\mathcal{F}_0$ easily leads to the frame change from $\mathcal{F}_{\mathbf{LF}_1}$ to $\mathcal{F}_{\mathbf{LF}_2}$.

We consider c_{a_1} as the reference frame, without loss of generality since ${}^{c_a}\mathbf{M}_0$ is constant, so the plane parameters are estimated as ${}^{c_{a_1}}\boldsymbol{\eta}$. Thus, $\mathbf{LF}_1$ only contributes to the computation of ${}^{c_{a_1}}\boldsymbol{\eta}$ whereas $\mathbf{LF}_2$ contributes to both ${}^{c_{a_1}}\boldsymbol{\eta}$ and ${}^{c_{a_2}}\mathbf{M}_{c_{a_1}}$. This is written by extending (21) and considering $\mathbf{m}$, a vector representation of ${}^{c_{a_2}}\mathbf{M}_{c_{a_1}}$, to:

$$\begin{bmatrix} {}^{c_{a_1}}\hat{\boldsymbol{\eta}} \\ \hat{\mathbf{m}} \end{bmatrix} = \min_{{}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m}} \sum_q {}^1C_q({}^{c_{a_1}}\boldsymbol{\eta}) + {}^2C_q({}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m}), \tag{28}$$

where

$${}^1C_q({}^{c_{a_1}}\boldsymbol{\eta}) = ||{}^{c_{b_1}}\mathbf{H}_{c_{a_1}}({}^1\boldsymbol{\eta})\,{}^{c_{a_1}}\tilde{\mathbf{p}}^*_q \ominus {}^{c_{b_1}}\tilde{\mathbf{p}}_q||, \tag{29}$$

and

$$\begin{aligned} {}^2C_q({}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m}) = &\ ||{}^{c_{c_2}}\mathbf{H}_{c_{a_1}}({}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m})\,{}^{c_{a_1}}\tilde{\mathbf{p}}^*_q \ominus {}^{c_{e_2}}\tilde{\mathbf{p}}_q|| \\ + &\ ||{}^{c_{d_2}}\mathbf{H}_{c_{a_1}}({}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m})\,{}^{c_{a_1}}\tilde{\mathbf{p}}^*_q \ominus {}^{c_{f_2}}\tilde{\mathbf{p}}_q|| \end{aligned}. \tag{30}$$

In (29), ${}^{c_{b_1}}\mathbf{H}_{c_{a_1}}({}^{c_{a_1}}\boldsymbol{\eta})$ is computed using (15). In (30), we introduce SACs c_e and c_f of $\mathbf{LF}_2$ in order to get enough generality in the mathematical writings to be able to consider different pairs of SACs between LFs. Indeed, as choosing the best SACs pair leads to more reliable results (Sec. IV-C), it is also a key idea in the planar object pose estimation. Thus, ${}^{c_{e_2}}\mathbf{H}_{c_{a_1}}({}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m})$ is computed similarly to ${}^{c_{b_1}}\mathbf{H}_{c_{a_1}}({}^{c_{a_1}}\boldsymbol{\eta})$ but by, first, changing the frame of ${}^{c_{a_1}}\boldsymbol{\eta}$ to $\mathcal{F}_{c_{a_2}}$, then to $\mathcal{F}_{c_{e_2}}$, using the constant frame change ${}^{c_e}\mathbf{M}_{c_a}$, and, second, considering the varying frame change ${}^{c_{a_2}}\mathbf{M}_{c_{a_1}}$. The latter's translation and rotation (axis-angle representation) elements are merged together as $\mathbf{m} = \left[{}^{c_{a_2}}t^X_{c_{a_1}}, {}^{c_{a_2}}t^Y_{c_{a_1}}, {}^{c_{a_2}}t^Z_{c_{a_1}}, {}^{c_{a_2}}\theta^X_{c_{a_1}}, {}^{c_{a_2}}\theta^Y_{c_{a_1}}, {}^{c_{a_2}}\theta^Z_{c_{a_1}}\right]$. ${}^{c_{f_2}}\mathbf{H}_{c_{a_1}}({}^{c_{a_1}}\boldsymbol{\eta},\ \mathbf{m})$ follows similarly. Note that if both LFs have the same best SACs pair at some instant, ${}^{c_{e_2}}\mathbf{H}_{c_{a_1}}$ reduces to ${}^{c_{a_2}}\mathbf{H}_{c_{a_1}}$ and ${}^{c_{f_2}}\mathbf{H}_{c_{a_1}}$ reduces to ${}^{c_{b_2}}\mathbf{H}_{c_{a_1}}$.

To optimize simultaneously both ${}^{c_{a_1}}\boldsymbol{\eta}$ and $\mathbf{m}$, following the same methodology that leads to (27), one must compute increments:

$$\begin{bmatrix} {}^{c_{a_1}}\dot{\boldsymbol{\eta}} \\ \dot{\mathbf{m}} \end{bmatrix} = -{\mathbf{J_2}}^+\mathbf{e_2}, \tag{31}$$

used for updating ${}^{c_{a_1}}\boldsymbol{\eta}$ and $\mathbf{m}$ in a looping process until convergence. ${}^{c_{a_1}}\dot{\boldsymbol{\eta}}$ updates ${}^{c_{a_1}}\boldsymbol{\eta}$ by addition and $\dot{\mathbf{m}}$ updates $\mathbf{m}$ by composition (see the use of the exponential map of se(3) from the Lie Algebra in [17] for more details). In (31), $\mathbf{e_2}$ is the stacking of differences between computed and measured corresponding point coordinates, similarly to (27), but for three pairs of SAC images. Jacobian $\mathbf{J_2}$ first three columns contain the stacking of Jacobians $\mathbf{J}_{\mathcal{F}_1,\mathcal{F}_2}$ (see (23)), computed for SAC pairs (c_{a_1} ; c_{b_1}) (thus computed exactly as in (26), *i.e.* $\mathbf{J}_{c_{a_1},\ c_{b_1}} = \mathbf{J}_{c_a,\ c_b}$, for ${}^{c_{a_1}}\tilde{\mathbf{p}} = {}^{c_a}\tilde{\mathbf{p}}$), ($c_{a_1}$; c_{e_2}) and (c_{a_1} ; c_{f_2}), in the general case about pairs of SACs. Both latter are computed using (25), since, in these pairs, SACs belong to different LFs so $\mathbf{R}_\cdot$-s and $t_\cdot$-s may have

any values. Juxtaposing the stacking of interaction matrices $\mathbf{L}_{c_{\cdot}\mathbf{p}_q^*}$, relating ${}^{c_{\cdot}}\dot{\mathbf{p}}_q^*$ to $\dot{\mathbf{m}}$ [2], to $\mathbf{J_2}$ finishes its building:

$$\mathbf{J_2} = \begin{bmatrix} \mathbf{J}_{c_{a_1},\ c_{b_1}} & \mathbf{0} \\ \mathbf{J}_{c_{a_1},\ c_{e_2}} & \mathbf{L}_{c_{e_2}\mathbf{p}^*}\,{}^{c_e}\mathbf{V}_{c_a} \\ \mathbf{J}_{c_{a_1},\ c_{f_2}} & \mathbf{L}_{c_{f_2}\mathbf{p}^*}\,{}^{c_f}\mathbf{V}_{c_a} \end{bmatrix}, \tag{32}$$

where $\mathbf{L}_{c_{\cdot}\mathbf{p}^*}$ is, itself, the stacking of every $\mathbf{L}_{c_{\cdot}\mathbf{p}_q^*}$. ${}^{c_e}\mathbf{V}_{c_a}$, the twist transformation matrix, directly computed from ${}^{c_e}\mathbf{M}_{c_a}$, to locally transform $\mathbf{L}_{c_{e_2}\mathbf{p}^*}$ to $\mathcal{F}_{c_{a_2}}$, so that it contributes to the ${}^{c_{a_2}}\mathbf{M}_{c_{a_1}}$ frame change computation (same for ${}^{c_f}\mathbf{V}_{c_a}$).

VI. Experiments

The proposed linear and non-linear plane parameters estimations (Section IV) have been validated on simulation data before being applied on actual LF acquisitions. The estimation of a planar object pose transformation between two LFs (Section V) has similarly been evaluated. Due to lack of space, only real experiments are presented in the following.

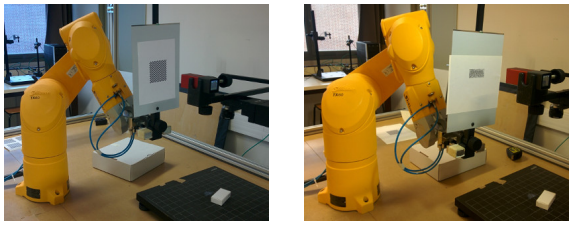
First of all, the LF sequences acquisition setup is detailed. Then, we study the influence of the selected pairs of SACs on the plane parameters estimation accuracy to validate the selection criteria proposed in this paper. Finally, the proposed planar object pose estimation is used for tracking a textured plane.

A. Data acquisition setup

The LF camera considered in our experiments is a Lytro Photo (first generation) camera. The LF camera is static and faces a planar surface moved by the end-effector of a Stäubli TX-60 6-axis robotic arm (Fig. 3).

First, a checkerboard of 19×19 squares of $3.6mm$ side is set on the planar surface of frame $\mathcal{F}_O$. Twelve $\mathbf{LF}_s$ for $s = 1...12$ are acquired, varying the pose of the checkerboard (Fig. 4). Each checkerboard pose is measured w.r.t the robot base, $\mathcal{F}_B$, by the TX-60 internal software to precisely know the 12 ${}^B\mathbf{M}_{O_s}$ (for $s = 1...12$) transformations. These 12 $\mathbf{LF}$ are set as the input of the Dansereau's MATLAB Toolbox [13] to calibrate the Lytro camera. As output, the calibration gave us, inter alia, the 5×5 intrinsics matrix (3), rectification parameters and the 12 poses of the checkerboard w.r.t. the LF camera: ${}^{\mathbf{LF}}\mathbf{M}_{O_s}$ for $s = 1...12$. We used the 12 ${}^B\mathbf{M}_{O_s}$ and the 12 ${}^{\mathbf{LF}}\mathbf{M}_{O_s}$ to estimate the fixed pose of the camera w.r.t. the robot base: ${}^B\mathbf{M}_{\mathbf{LF}}$[2]. Therefore it is now straightforward to obtain a ground truth of the pose for any planar object put on the robot end-effector w.r.t the camera to evaluate the accuracy of our method.

[2]LF images and robot poses are available in the LFMIS dataset at: mis.u-picardie.fr/∼g-caron/pub/data/LFMIS_dataset.zip



(a) (b)

Fig. 3: Acquisition setup - A Lytro Photo LF camera facing a planar object that is moved by a Stäubli TX60 robotic arm.

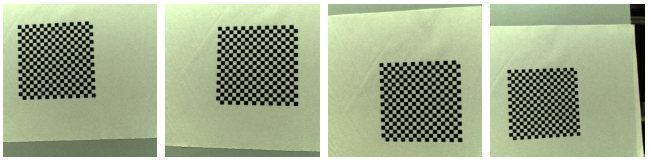


Fig. 4: Central SAI of relevant $\mathbf{LF}$ to validate our best SACs pair selection strategy.

The calibration toolbox output is also the decoded 4-D LF data structure (Section II). After re-organization, 9×9 SAIs of size 380×380 pixels are obtained from one LF. In other words, the Lytro camera can be considered as a grid of 9×9 SACs. However, due to the undistortion, the outer SAIs may be cropped or blurred. For this reason, during the following experiments, we are only considering SAIs that belong to the 5×5 centered SACs of the grid. More precisely, the top-left usable SAI is noted as $SAI_{(3\times3)}$, the top-right usable SAI is $SAI_{(7\times3)}$, the bottom-left is $SAI_{(3\times7)}$, the bottom-right is $SAI_{(7\times7)}$ and so on.

B. Best pairs of SACs study

1) Selection strategy evaluation: The purpose of this study is to experimentally validate the best SACs pair selection strategy introduced in Section IV-C.

For this experiment, we consider the 4 LFs shown in Fig. 4. We note these LFs: $\mathbf{LF}_1, \mathbf{LF}_2, \mathbf{LF}_3$ and $\mathbf{LF}_4$. As it can be seen on these 4 SAIs, the visual features (the checkerboard corners) are located in a specific part of the field of view in each LF, respectively in the top-left, top-right, bottom-right and bottom-left. These 4 LFs are then interesting candidates to validate our best SACs pair selection strategy that combines baseline and epipolar non-collinearity constraints.

We estimate the plane parameters of the 4 checkerboards following the approach described in Section IV and using different pairs of SACs: $\{SAC_{(7\times3)}, SAC_{(3\times7)}\}$ and $\{SAC_{(3\times3)}, SAC_{(7\times7)}\}$. Recall that these two pairs of SACs correspond to the two widest baselines of the usable SACs. TABLE I contains the results of this experimentation. We would like to remind that the maximum baseline between usable SACs is equal to $3mm$. The best pairs of SAC for $\mathbf{LF}_1$, $\mathbf{LF}_2$, $\mathbf{LF}_3$ and $\mathbf{LF}_4$ are highlighted in TABLE I. To quantitatively evaluate the estimation we compute the mean distance between the ground truth plane and the estimated one. More precisely, this metric corresponds to the average distance between a grid of corresponding points equally distributed on the two planes. We can see that there is a link between the pair of SACs used in the estimation of the plane

	$\mathbf{SAC}_{(7\times3)}$ $SAC_{(3\times7)}$	$\mathbf{SAC}_{(3\times7)}$ $SAC_{(7\times3)}$	$\mathbf{SAC}_{(3\times3)}$ $SAC_{(7\times7)}$	$\mathbf{SAC}_{(7\times7)}$ $SAC_{(3\times3)}$
$\mathbf{LF}_1$	3.78	3.64	6.30	6.18
$\mathbf{LF}_2$	5.32	5.36	3.21	3.17
$\mathbf{LF}_3$	2.03	2.04	4.58	4.38
$\mathbf{LF}_4$	4.07	4.38	1.50	1.42

TABLE I: Plane estimation errors [mean distance in mm] for various LFs using the four pairs of SACs of widest baselines. Bold **SAC** is the reference one.

parameters and the position of the visual features in the LF. The best pairs of SACs for $\mathbf{LF}_1$, $\mathbf{LF}_2$, $\mathbf{LF}_3$, and $\mathbf{LF}_4$ are respectively $\{SAC_{(7\times3)}, SAC_{(3\times7)}\}$, $\{SAC_{(3\times3)}, SAC_{(7\times7)}\}$, $\{SAC_{(7\times3)}, SAC_{(3\times7)}\}$, and $\{SAC_{(3\times3)}, SAC_{(7\times7)}\}$. These best SACs pairs correspond to the misalignment of the checkerboard center of gravity in the LF camera field of view w.r.t. both SACs, as awaited. Thus, these results validate the best SACs pair selection strategy proposed in Section IV-C.

2) The best SACs pair versus 25 SACs: As previously mentioned (Section IV), more than two SACs may be used in our plane parameters estimation scheme. Indeed, using the 25 available SACs seems, intuitively, interesting. Any SAC may be used as a reference, thus there are 25 combinations. We have performed the same experiment as before on the same 4 LFs (Fig. 4) but using these 25 combinations of 25 SACs. TABLE II contains for each LF, the best estimation among the 25 and the SAC that has been used as a reference to obtain this result.

First, it is interesting to note that, apart from the $\mathbf{LF}_2$, the SAC used as a reference that provides the best result is the same (or is very close, as for $\mathbf{LF}_1$) to the SAC selected following our proposed best SACs pair selection strategy (TABLE I). Considering the 25 SACs in $\mathbf{LF}_2$ not only led to a different reference SAC than for our proposed selection scheme but also to a lower estimation precision. Thus, it shows that using every available SAC does not always provide a better estimation result than using a pair, *i.e.* the best pair under the criteria proposed in this paper.

For the other considered **LF**s, thanks to the redundancy of information the estimations are more accurate than using the best pairs of SACs (TABLE I). However, the computation using 25 SACs is very time-consuming. An estimation using 25 SACs is performed in approximatively $9.0s$ when it takes $0.7s$ using a single pair (MATLAB implementation running on a laptop with an Intel i7-4900MQ CPU with 16GB RAM). So, our SAC pair selection strategy shows being an interesting tradeoff between precision and computation time.

C. Evaluation on textured planar object acquisitions

For the following experiments, we changed the planar object held by the robot (Fig. 3b) and acquired several LF sequences. The Lytro camera is not able to record a video. To overcome this issue, we capture LF sequences by taking a series of LF, slightly moving the planar object between each shots using the robotic arm. Fig. 5 shows some central SAIs of LFs extracted from two acquired sequences.

As for the first experiment, we need to build a set of light ray correspondences between two pairs of SAIs extracted from two consecutive LFs to estimate the displacement of the planar object. However, the visual features detection and matching are more difficult in the case of a textured planar scene rather than a checkerboard. In other words, we must be able to find visual features that each of the four SAIs (two SAIs in the first LF and two others in the second) have in common.

To perform this, we detect and match SURF features [18] in the four SAIs. The matching is done by setting an experimentally found hard threshold to avoid outliers while keeping enough matches (5 minimum and well balanced on the object surface) to simultaneously compute the plane parameters and its pose. The data provided by this matching has just the purpose to provide feature points with realistic noise and location uncertainties to look forward for robotics applications. True, more optimal methods of features matching or tracking could be used but the current experiment focuses on multi-LF geometry estimation, not about image processing.

The first pose of the acquired planar object is supposed to be known. Then, our method estimates its displacement, shot after shot, incrementally, and is compared to the ground truth. This experiment evaluates the accuracy of this planar object incremental pose computation depending on the pairs of SACs considered for each LF. More precisely, we compare the estimated poses under the five following SACs configurations:

- D1 - the same pair of SACs for every LF: $\{SAC_{(3\times3)}, SAC_{(7\times7)}\}$ (first diagonal baseline)
- D2 - the same pair of SACs for every LF: $\{SAC_{(7\times3)}, SAC_{(3\times7)}\}$ (second diagonal baseline)
- H - the same pair of SACs for every LF: $\{SAC_{(3\times3)}, SAC_{(7\times3)}\}$ (horizontal baseline)
- V - the same pair of SACs for every LF: $\{SAC_{(7\times3)}, SAC_{(7\times7)}\}$ (vertical baseline)
- DHV - the best pair of SACs for each LF following the strategy selection proposed in Section IV-C.

For more generality of results, we consider two LF sequences (Fig. 5) on which the latter five settings are considered for the planar object incremental pose estimation. The two sequences are composed of nine LFs each. During the acquisition of both sequences the plane has been displaced on approximatively $20cm$. In the first sequence, the plane follows pure translations. In the second sequence, translations and rotations are applied.

To take into account the estimation drift, inherent to any incremental estimation, the last planar object estimated pose of each path is compared with the ground truth (TABLE III).

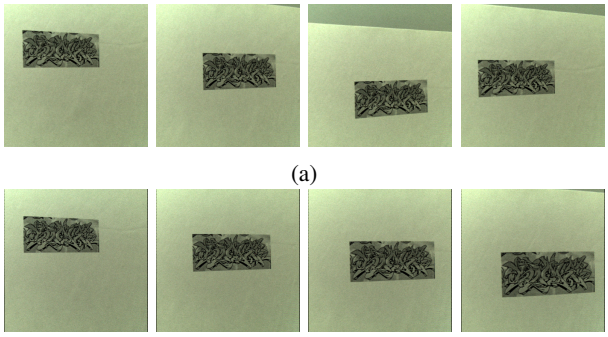
Estimation results of both experiments clearly show that the pose estimation of the planar object is highly more accurate when taking advantage of the LF camera acquisition structure, as we proposed, than considering a constant horizontal pair of cameras as in classical stereo-vision: precision improvement

	25 SACs	Reference
$\mathbf{LF}_1$	$1.84mm$	$\mathbf{SAC}_{(3\times6)}$
$\mathbf{LF}_2$	$3.35mm$	$\mathbf{SAC}_{(3\times7)}$
$\mathbf{LF}_3$	$1.41mm$	$\mathbf{SAC}_{(3\times7)}$
$\mathbf{LF}_4$	$0.62mm$	$\mathbf{SAC}_{(7\times7)}$

TABLE II: Plane estimation errors using the 25 SACs and the reference SAC. Best result among the 25 combinations.

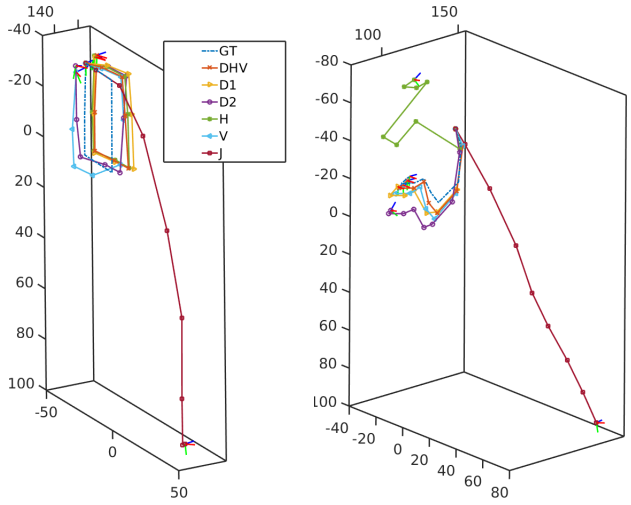
	Sequence 1	**Sequence 2**
D1	$4.48mm$; $19.05°$	$7.96mm$; $5.44°$
D2	$5.21mm$; $23.74°$	$19.68mm$; $47.20°$
H	$4.86mm$; $22.96°$	$56.43mm$; $34.51°$
V	$10.13mm$; $35.31°$	$5.41mm$; $3.68°$
DHV	$5.33mm$; $3.50°$	$2.38mm$; $0.34°$

TABLE III: Estimation errors in position and orientation at the end of the plane tracking.



(a) (b)

Fig. 5: Some central SAIs that compose two LF sequences.



(a) (b)

Fig. 6: Comparisons between ground truth (GT) and the paths estimated using the different pairs of SACs. (a), (b) the first and the second (Fig. 5) LF sequences, resp. (units: mm)

of up to an order of 14 for the first angle of DHV, w.r.t. the first angle of H, in sequence 2, for instance. The comparisons between the estimated paths and the ground truth (Fig. 6) also highlight these observations.

To compare with a LF-based approach from the state-of-the-art, Figure 6 also shows an additional estimate of the planar object path (path "J") using [11]'s Matlab scripts[3]. The latter Structure-From-Motion approach, while producing precise results in non-planar scenes, produces poor results when considering a planar one due to its essential matrix-based modeling [5, Sec. 11.9.2, p. 296]. Our approach outperforms the latter, thanks to its consideration of the fact that the observed scene is planar.

VII. Conclusion

In this paper, we have proposed a planar object pose estimation from light rays captured by a LF camera. To develop this method, we have first expressed the 5×5 homography between corresponding rays of a LF. The LF camera geometry and the 5×5 homography have been considered to implement a plane parameters estimation from one LF. A linear and a more robust estimation based on non-linear optimization have been proposed, considering the best pair of SACs. Based on all of this, we have introduced a method to estimate the rigid transformation of a planar object between two LFs. Results show that the combination of these contributions leads to the reliable planar object tracking over time from multiple LF acquisitions.

References

[1] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: A hands-on survey," IEEE Trans. on Visualization and Computer Graphics, vol. 22, no. 12, pp. 2633–2651, Dec 2016.

[2] F. Chaumette and S. Hutchinson, "Visual servo control, Part I: Basic approaches," IEEE Robotics and Automation Mag., vol. 13, no. 4, pp. 82–90, December 2006.

[3] F. Fraundorfer and D. Scaramuzza, "Visual odometry : Part ii: Matching, robustness, optimization, and applications," IEEE Robotics Automation Mag., vol. 19, no. 2, pp. 78–90, June 2012.

[4] Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. on Pattern Analysis and Machine Intell., vol. 22, pp. 1330–1334, 2000.

[5] R. I. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, 2004.

[3]http://lightfield-analysis.net/accv2016_ba_code.zip

[6] R. Acuna and V. Willert, "Robustness of control point configurations for homography and planar pose estimation," 2018. [Online]. Available: http://arxiv.org/abs/1803.03025

[7] N. B. Monteiro, S. Marto, J. P. Barreto, and J. Gaspar, "Depth range accuracy for plenoptic cameras," Computer Vision and Image Understanding, 2018.

[8] C. Hahne, A. Aggoun, V. Velisavljevic, S. Fiebig, and M. Pesch, "Baseline and triangulation geometry in a standard plenoptic camera," Int. J. Comput. Vision, vol. 126, no. 1, pp. 21–35, Jan. 2018.

[9] D. Tsai, D. G. Dansereau, T. Peynot, and P. Corke, "Image-based visual servoing with light field cameras," IEEE Robotics and Automation Lett. (RAL), vol. 2, no. 2, pp. 912–919, 2017.

[10] F. Dong, S.-H. Ieng, X. Savatier, R. Etienne-Cummings, and R. Benosman, "Plenoptic cameras in real-time robotics," The Int. Journal of Robotics Research, vol. 32, no. 2, pp. 206–217, 2013.

[11] O. Johannsen, A. Sulc, and B. Goldluecke, "On linear structure from motion for light field cameras," IEEE Int. Conf. on Computer Vision (ICCV), pp. 720–728, Dec. 2015.

[12] Y. Zhang, P. Yu, W. Yang, Y. Ma, and J. Yu, "Ray space features for plenoptic structure-from-motion," in IEEE Int. Conf. on Computer Vision (ICCV), Oct. 2017, pp. 4641–4649.

[13] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Calibration and rectification for lenselet-based plenoptic cameras," IEEE Conf. on Computer Vision and Pattern Recogn. (CVPR), pp. 1027–1034, 2013.

[14] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] T. Knight, Y.-R. Ng, and C. Pitts, "Light field data acquisition devices, and methods of using and manufacturing same," U.S. Patent US8 289 440B2, 2012.

[16] B. Wilburn, M. Smulski, H.-H. Kelin Lee, and M. Horowitz, "The light field video camera," in Media Processors, SPIE Electronic Imaging, 2002.

[17] Y. Ma, S. Soatto, J. Košecká, and S. Sastry, An invitation to 3D vision. Springer, 2004.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346 – 359, 2008.