The International Journal of
Robotics Research
32(2) 206–217

DOI: 10.1177/0278364912469420
ijr.sagepub.com

Plenoptic cameras in real-time robotics

Fengchun Dong[1], Sio-Hoi Ieng[2], Xavier Savatier[1], Ralph Etienne-Cummings[3] and Ryad Benosman[2]

Abstract

Real-time vision-based navigation is a difficult task largely due to the limited optical properties of single cameras that are usually mounted on robots. Multiple camera systems such as polydioptric sensors provide more efficient and precise solutions for autonomous navigation. They are particularly suitable for motion estimation because they allow one to formulate a linear optimization. These sensors capture the visual information in a more complete form called the plenoptic function that encodes the spatial and temporal light radiance of the scene. The polydioptric sensors are rarely used in robotics because they are usually thought to increase the amount of data produced and require more computational power. This paper shows that these cameras provide more accurate estimation results in mobile robotics navigation if designed properly. It also shows that a plenoptic vision sensor with a resolution ranging from 3×3 *to* 40×30 *pixels camera, provides higher accuracy than a mono-SLAM running on a* 320×240 *pixels camera. The paper also gives a complete scheme to design usable real-time plenoptic cameras for mobile robotics applications by establishing the link between velocity, resolution and motion estimation accuracy. Finally, experiments on a mobile robot are shown allowing for a comparison between optimal plenoptic visual sensors and single high-resolution cameras. The estimation with the plenoptic sensor is more accurate than a monocular high-definition camera with a processing time 100 times lower.*

Keywords

plenoptic function, egomotion, non central vision, scalespace, vision based navigation

1. Introduction

Efficient vision-based navigation in mobile autonomous robotics is usually a difficult task to achieve due to the real-time requirements and limited computational power available. Current approaches rely on the use of one or two perspective cameras simultaneously. Adding more cameras is generally incompatible with real-time robotics as more data would need to be processed (Rivas Lopez and Tyrsa, 2008). Standard perspective cameras capture only one ray from each point in space. Motion is generally estimated from the periodic measurements of light as the camera moves in the scene. This estimation is generally nonlinear and difficult to solve because it is ill-posed (Neumann et al., 2002). Monocular vision systems also have limitations due to their narrow field of view that introduces ambiguities in distinguishing small rotations from translations or in identifying the exact scale factor.

These problems can be solved by increasing the number of cameras so that larger portions of the scene are covered. However, this option does not appear optimal for robotics as it introduces an increase of the amount of data to acquire, transfer and process, thus requiring more computational power. Multiple camera systems acquire multiple rays from the same scene point allowing the estimation of motion to be linear. A more complete light model introduced by Adelson and Bergen (1991) is also used for a more accurate motion estimations. This light model is expressed by the plenoptic function which is a six-dimensional (6D) mapping (if light wavelength is neglected) used to represent time-varying light rays. An imaging device captures only a small subset of the plenoptic function. If a plenoptic camera captures the plenoptic function at few locations in space, it can then be seen as a set of classic perspective cameras.

[1]Instrumentation, IT and Systems Department IRSEEM, Rouen, France
[2]UPMC Univ Paris 06, UMR 7210, UMR_S968, The Vision Institut, Paris, France
[3]Computational Sensory-Motor System Lab, The John Hopkins University, Baltimore, MD, USA

Corresponding author:
Siohoi Ieng, UPMC Univ Paris 06, UMR 7210, UMR_S968, The Vision Institut, 4 place Jussieu, 75005 Paris, France.
Email: sio-hoi.ieng@upmc.fr

These sensors are generally referred to as plenoptic cameras (Adelson and Wang, 1992).

This principle originates from a photographic technique called "integral photography" described by Okoshi (1976). A plenoptic camera is composed of several cameras set close to each other, it can then simultaneously capture a large number of images. Recent technological advances have allowed the production of several plenoptic related cameras, namely, Stanford's multi-camera array (Wilburn, 2004), the Panoptic camera (Raboud, 2009) and micro lens array cameras (Ng et al., 2005; Raskar et al., 2008; Georgiev and Lumsdaine, 2010). Handheld plenoptic cameras are also available in the market from companies such as Raytrix or Lytro. Other small vision-sensor chips have also been developed for applications in which thickness is an issue (Brady and Morrison, 2000; Tanida and Yamada, 2002; Duparré et al., 2004; Völkel and Duparré, 2008).

Evolution has selected multiple camera systems when it comes to processing complex visual information with the lowest power consumption. Insects with limited cognitive abilities rely on multiple visual sensors that prove to be particularly efficient for performances that current robotics struggle to attain (Franceschini, 2008). There is a wide variety of compound eyes, in various shapes and in size, allowing insects to perform complex manoeuvres such as reactive obstacle avoidance (Land and Collett, 1974), fast prey detection (Olberg et al., 2000), target localization and tracking (Forster, 1979; Jackson and R., 2002), and accurate vision-based odometry (Srinivasan et al., 2000) at a very low power cost. Compound eyes also allow a wide coverage of scenes and can provide, in some cases, an omnidirectional field of view. Insect brains seem capable of processing the amount of information mainly because of two factors. First, the compound eyes have a more optimal coverage of the scene and, second, they allow for measurement of light from several viewpoints and directions. The high-resolution needs are consequently reduced as the number of necessary pixels decreases. For example, while the number of photosensitive elements, the ommatidia (or unit eyes), approximates 8000 in bees, it stands around 30,000 in dragonflies (Tinbergen, 1980).

Works on motion estimation with compound eyes first focused on elementary motion detectors (EMDs) and then transposed to the plenoptic cameras. EMDs are based on models of insect motion detection (Hassenstein and Reichardt, 1956; Borst and Egelhaaf, 1993; VanSanten and Sperling, 1984), with an important property being that motion detection is sensitive to image contrast (Borst and Egelhaaf, 1993). As such, the amplitude of motion detected will be greater where higher contrast exists given the same underlying motion. The EMDs are therefore incapable of providing precise visual motion estimates unless an automatic gain control in the ommatidia is implemented. The EMDs are usually used in reflexive obstacle avoidance and stabilization in flying robots where only discrete motion adjustments are required. Some recent examples include (Iida and Lambrinos, 2000; Ruffier and Franceschini, 2003; Liu and Usseglio-Viretta, 2001; Wei Chung et al., 2003). However, given their lack of precision, EMDs are not suitable for general navigation, particularly tasks requiring fine motion control.

The applications of plenoptic cameras in real robotics tasks are still scarce. This is probably due to the fact that currently available plenoptic sensors are built for computational imaging tasks. These sensors aim at studying the principle of integral imaging that allows for image refocusing, realistic image rendering, building high dynamic range images and achieving 3D imaging (Adelson and Wang, 1992; Gortler and Grzeszczuk, 1996; Levoy and Hanrahan, 1996; Camahort and Fussell, 1999).

A recent work close to that presented in this paper has been published by Dansereau et al. (2011) where the plenoptic function is used to achieve real-time navigation. The approach is however fundamentally different from the presented work. Dansereau et al. (2011) applied the plenoptic function to a set of cameras with fixed parameters. They introduce three distinct closed-form solutions to extract the motions parameters from the plenoptic function. Our effort focuses on studying how to reduce data to allow real-time processing without altering the accuracy of the visual odometer by optimizing the sensor's parameters. We inquire into the problem of defining an optimal plenoptic camera for a robotics task. We compare the use of a single high-resolution camera versus a lower-resolution plenoptic sensor. At the core of the question is the problem of visual navigation and motion estimation: what plenoptic camera should be built for a specific robotics navigation task according to a given range of speed of the robot? What is the lowest possible resolution of a plenoptic camera needed to subsequently facilitate the motion estimation problem in the best possible way and perform as well or even better than a single high-resolution camera?

We will show that there is an optimal set of parameters allowing a tradeoff between resolution, field of view and accuracy. The optimal plenoptic camera reduces drastically the unnecessary data with minimal accuracy loss. As will be shown, high-resolution sensors are often less efficient in most navigation tasks, performing worse than a carefully designed plenoptic camera with a much lower resolution.

2. Motion and structure from the plenoptic function

2.1. Motion from the plenoptic function

The plenoptic function is a more complete physics model of light that describes an image from a bundle of sampled rays where each ray is characterized by the position and orientation from which it is seen, at a given time t. Such a light model defines the structure of the visual space, denoted $L(\mathbf{x}, \mathbf{r}, t)$, with:

- L as the radiance;
- $\mathbf{x} = (x, y, z)^{\mathrm{T}} \in \mathbb{R}^3$ as the viewing position;
- $\mathbf{r} = (\theta, \varphi)^{\mathrm{T}} \in [0,\ 2\pi]^2$ as the direction from which L is seen;
- $t \in \mathbb{R}^+$ as the time.

The plenoptic-based motion estimation is applied with the formalism established by Neumann et al. (2002, 2004). The vision sensor is moving through the scene and each of its cameras at $\mathbf{x}$ samples the plenoptic function from $\mathbf{r}$, at t. The egomotion is rigid, i.e. given by a rotation matrix R and a translation vector $\mathbf{T}$.

Estimating the motion parameters from the plenoptic function assumes the smoothness of the signal, which allows a Taylor's first-order expansion. A second hypothesis is the constant scene illumination overtime. If this is fulfilled, the plenoptic function satisfies the photo-consistency constraint which is similar to that used in optical flow computation (Neumann, 2004):

$$-L_t = (\nabla_{\boldsymbol{x}} L)^{\mathrm{T}} \frac{d\boldsymbol{x}}{dt} + (\nabla_{\boldsymbol{r}} L)^{\mathrm{T}} \frac{d\boldsymbol{r}}{dt}, \tag{1}$$

where $\nabla_x L$ and $\nabla_r L$ are the spatial gradients of L and L_t its partial time derivative $\frac{\partial L}{\partial t}$. If $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)^{\mathrm{T}}$ and $\mathbf{q} = (q_x, q_y, q_z)^{\mathrm{T}}$ are, respectively, the angular velocity and the instantaneous translation, then Equation (1) becomes

$$-L_t = (\nabla_x L)^{\mathrm{T}} \boldsymbol{q} + (x \times \nabla_x L + r \times \nabla_r L)^{\mathrm{T}} \boldsymbol{\omega}, \tag{2}$$

where $\times$ is the cross product. This equality is referred to as "differential plenoptic motion constraint" which is a linear, scene-independent constraint in the motion parameters and the plenoptic partial derivatives. The mixture of Cartesian and spherical coordinates ($\mathbf{x}$ and $\mathbf{r}$) in Equation (2) is inconvenient to manipulate. This equation is replaced by a more flexible but equivalent form derived from the two-plane parametrization (Levoy and Hanrahan, 1996). Equation (2) can then be rewritten by computing the Jacobian between the original and the two-plane parametrizations, and then simplified into the matrix form:

$$-L_t = \begin{pmatrix} L_x & L_y & L_u & L_v \end{pmatrix} M \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{\omega} \end{pmatrix}, \tag{3}$$

with

$$M = \left(\begin{array}{ccc|ccc} 1 & 0 & -\frac{u}{f} & -\frac{uy}{f} & \frac{ux}{f} + Z_\pi & -y \\ 0 & 1 & -\frac{v}{f} & -(\frac{vy}{f} + Z_\pi) & \frac{vx}{f} & x \\ 0 & 0 & 0 & -\frac{uv}{f} & \frac{u^2}{f} + f & -v \\ 0 & 0 & 0 & -(\frac{v^2}{f} + f) & \frac{vu}{f} & u \end{array}\right),$$

the transformation matrix for the translation and rotation.

The meaning of each variable in (x, y, u, v, f, Z_π) is detailed in Neumann (2004). An over-determined linear system can be built with the contribution of each camera and solved for the rigid motion parameters with standard least-squares techniques.

The mechanism explaining why a plenoptic camera is more suitable for navigation task is shown in Figure 1. When moving through the scene, a conventional camera can truly satisfy the photo-consistency expressed by Equation (1) only for small motions while the plenoptic camera can handle much larger displacements. For such motions, a single ray can actually be captured by another camera unit in the plenoptic device, while for a single camera, this property no longer holds.

2.2. Multi-scale plenoptic function

Cameras' motions across the scene span from slow, rectilinear translations to complex and non-coplanar trajectories. The resulting velocity field is a complex mixture of motions which hardly satisfies the constant illumination hypothesis, especially for large displacements. Correct motion parameters would sometimes be impossible to estimate. One classic solution is to apply a multi-scale representation of the signal and to perform a multi-layered motion estimation, starting from the coarsest level that underlines larger motions and erases smaller ones. The results are propagated to the next finer level to contribute to a refined motion estimation. Classical application of the multi-scale technique consists of building a Gaussian pyramid on the image space.

In our case, a pyramid structure for a vector field l that samples the plenoptic function data is built and defined as follows: $\mathbf{l} = (x, y, u, v, L, L_x, L_y, L_u, L_v, L_t)^{\mathrm{T}}$, with

$$\begin{array}{ccc} l : \mathbb{R}^5 & \rightarrow & \mathbb{R}^{10} \\ (\mathbf{x}, \mathbf{r}, t) & \mapsto & \mathbf{l} = (x, y, u, v, L, L_x, L_y, L_u, L_v, L_t)^{\mathrm{T}}, \end{array} \tag{4}$$

where $L_x, L_y, \ldots$ are the partial derivatives of the plenoptic function with respect to each variable.

The multi-scale strategy is related to the scale space theory (Lindeberg, 1994) which is a formal theory that maps any vector field to a one-parameter family of smoothed vector fields. In the case of l, the multi-scale space represents a family of subsampled signals $l(\mathbf{x}, \mathbf{r}, t, k)$ smoothed by Gaussians of variance k, with k being also the scale parameter.

The motion estimation is performed by iterative propagations from the lowest resolution (highest level) to higher resolutions (lower level) with the possibility to stops and returns the results if the motion estimation satisfies some prefixed conditions. The pyramids are assumed n-level high, with the highest level storing the lowest-resolution data. The iterative motion estimation operates according to Algorithm 1. The threshold mentioned in line 8 is determined experimentally using a set of trajectories of the robot, it is taken as the mean value giving the lowest estimation error. This value is equal to 5×10^{-2}.

2.3. Optimal scale factors

The scale space theory is widely used in signal and image processing. Lindeberg (1994) gives a detailed formalization and analysis of the scale space representation for 2D

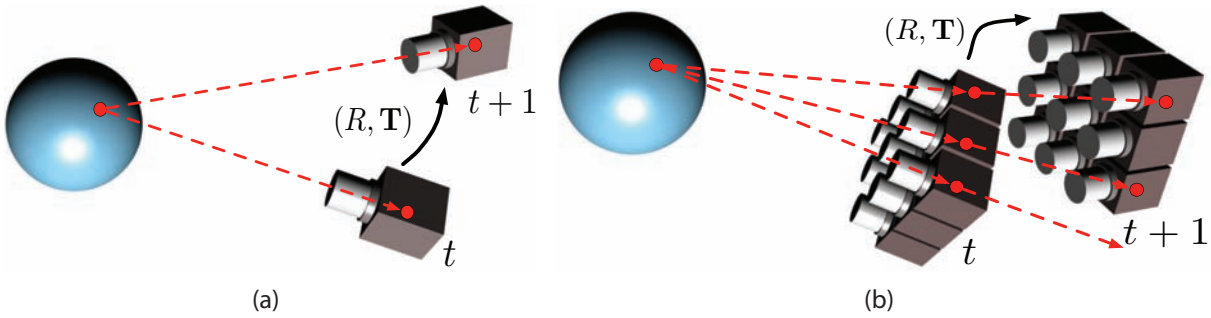

Fig. 1. (a) A monocular camera and (b) a plenoptic system observing the same scene while undergoing the identical rigid motion. In (a), the photo-consistency is true for small motions or if the scene reflects light isotropically since the camera is observing two different rays (dashed arrows) from its two positions. In (b), the photo-consistency is true because a same ray can be captured by two different cameras (e.g. the top dashed arrow captured by the sensor at t and $t+1$).

Algorithm 1 Multi-scale motion estimation.

1: **Inputs:** Motion parameters estimated at level $k+1$, the Gaussian and plenoptic pyramids.
2: **Outputs:** Motion parameters $(\mathbf{q}, \boldsymbol{\omega})^{\mathrm{T}}$.
3: **for** each level k **do**
4: Build Equation (3) with current level data and solve for the motion parameters: $(\tilde{\mathbf{q}}, \tilde{\boldsymbol{\omega}})^{\mathrm{T}}$.
5: Compute the plenoptic function and its derivatives with current level data, then update odd rows and columns of the plenoptic pyramid. Even rows and columns are updated with level $k+1$ of the plenoptic pyramid.
6: For each element at $(x, y)^{\mathrm{T}}$ of the pyramid, update it with a weighted mean value of the plenoptic vector over a 3×3 neighborhood (including the data of current level k and the data of previous level $k+1$).
7: Build a new Equation (3) with the updated pyramid data and solve for motion parameters: $(\mathbf{q}, \boldsymbol{\omega})^{\mathrm{T}}$.
8: **if** $e = \|(\mathbf{q}, \boldsymbol{\omega})^{\mathrm{T}} - (\tilde{\mathbf{q}}, \tilde{\boldsymbol{\omega}})^{\mathrm{T}}\| <$threshold **then**
9: exit and return the motion parameters
10: **end if**
11: **end for**

image signals. In this section, we extend the multi-scale representation to the plenoptic function, i.e. instead of building a pyramidal set of 2D images, we build a pyramid of 10-dimensional functions.

The multi-scale approach is motivated by the ability to deal with a wide range of natural scenes. The sensor motion estimation from the image is largely conditioned by the size and depth of image structures. Conversely, the sensor motion's velocity and direction have a direct impact on the resulting image signal. To reduce noise that limits the estimation quality, one should find the best scale of the signal representation that captures, as accurately as possible, the sensor's egomotion. Finding a unique global optimal scale assumes that motion components (i.e. instantaneous translation $\mathbf{q}$ and instantaneous rotation $\boldsymbol{\omega}$) are affected in a similar way. However, according to Equation (2), we can see that $\mathbf{q}$ depends only on the viewing positions while $\boldsymbol{\omega}$ depends on both the viewing positions and rays direction. This underlines that translations are mainly captured by $\nabla_x L$ while the instantaneous rotation are sensitive to more local changes.

According to this observation, we stipulate that both terms are not influenced by the scale factor in a similar way and each component should be optimized with its own scale factor. The best scale is given by the one producing the largest signature operators response. For example, in edges detection, the trace or the determinant of the Hessian matrix should be maximized while in texture characterization, the Hessian matrix is replaced by the second moment matrix. Figure 2 shows the response of the second moment matrix μ computed at the center of each image: $\mu = \begin{pmatrix} L_x^2 & L_x L_y \\ L_y L_x & L_y^2 \end{pmatrix}$. More precisely, $\det \mu$ is computed for each scale and the one maximizing its value gives the optimal scale for texture characterization. This shape signature is actually one of those used by Lindeberg (1994) to set an automatic scale detection. To underline that the scale has different influence on the $\nabla_x L, \nabla_y L$ (i.e. the inter-camera derivatives) and the $\nabla_u L, \nabla_v L$ (i.e. intra-camera derivatives), we computed the determinant of μ for each set of gradients (normalized by the maximum value). As we can see, in one case the maximum response is reached at scale (160×120) for the intra-camera derivatives and scale (10×8) for the inter-camera derivatives. In the second case these values are (40×30) and (10×8), respectively. Hence, the scale optimizing the estimation of $\mathbf{q}$ and that optimizing $\boldsymbol{\omega}$ are different.

This observation is a major issue as it requires Algorithm 1 to be redesigned to optimize separately the estimation of $\mathbf{q}$ and $\boldsymbol{\omega}$. One way to do it is to apply Algorithm 1 to find the optimal scale that stabilizes $\mathbf{q}$, then fixes $\mathbf{q}$ with the estimated value and re-apply Algorithm 1 to optimize $\boldsymbol{\omega}$.

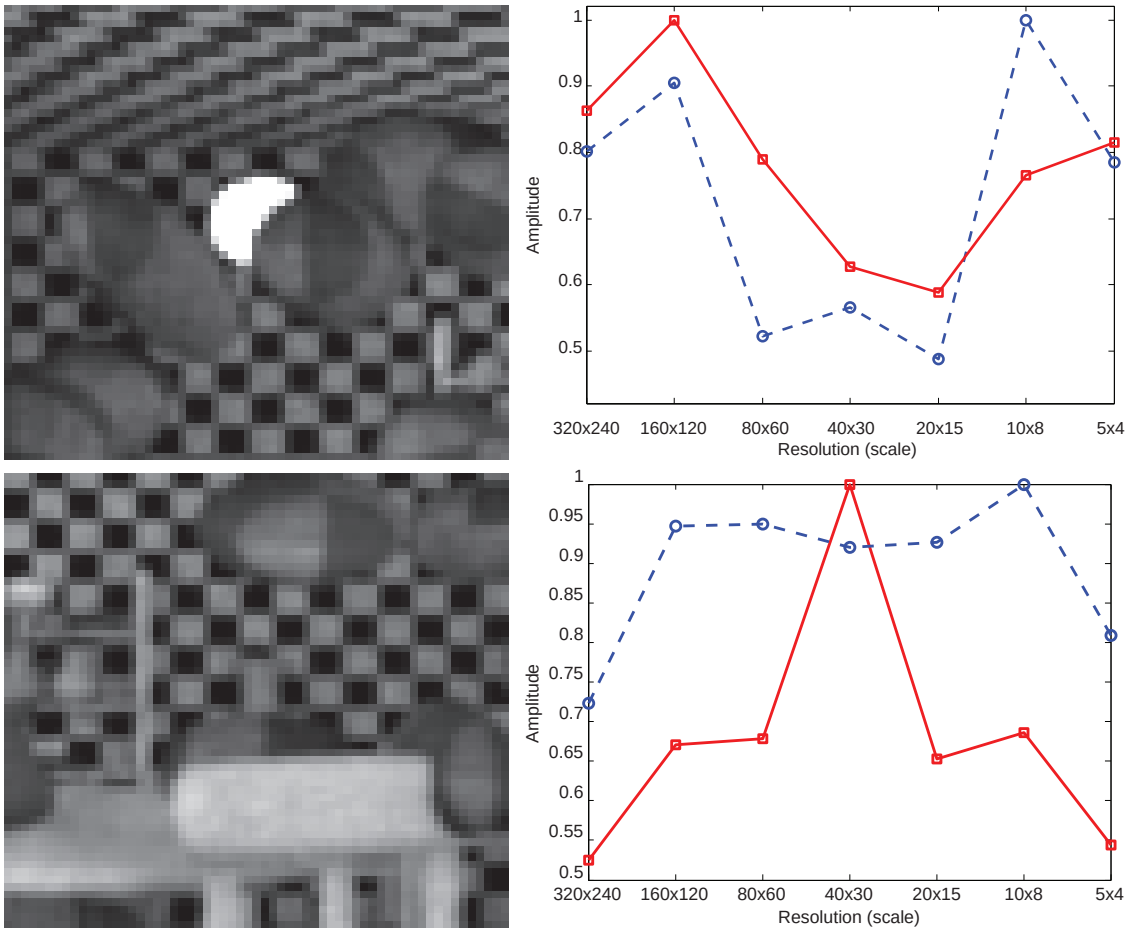

Fig. 2. Mean value of the determinant of the second moment matrix μ computed at the center of each image. The determinant is used as scale-space signature. The plain curves are the responses obtained for L_u, L_v (intra-camera gradient) while the dashed curves are those for L_x, L_y (inter-camera gradient). In each case, the maximums are not the same as we conjectured: scales 160×120 and 10×8 for the top and scales 40×30 and 10×8 for the bottom.

3. Experiments

3.1. Setup

The vision sensor used for the experiments is built out of a set of nine perspective cameras arranged in a 3×3 array, reproducing the functioning of a compound eye (see Figure 3). These cameras are configured to work synchronously with the same frame rate. A rigid mount was also designed to ensure the stability of the array structure spacing cameras by 5 cm both horizontally and vertically. The system is then calibrated with the multi-camera calibration method described by Svoboda et al. (2005), giving the relative pose of each camera. Their synchronization is achieved by sending a triggering signal to start the acquisition from the unique computer which also collects all of the incoming data streams. The image maximal resolution is set to 320×240 pixels as a tradeoff between signal resolution and available bandwidth allowed by the connection to the computer. Finally, the entire system is embedded on a mobile Pioneer platform to record data while the latter executes trajectories at various speeds up to 1.8 m s^{-1}. Higher translational speeds which are beyond the platform's ability are approximated by changing the cameras' frame rate. The estimation results at these speeds should be considered as the upper bounds of the odometer's accuracy.

In a fast acquisition process, the acquired images will suffer from motion blur as the platform's velocity increases. Results should however be comparable as for speeds exceeding 2 m s^{-1} higher frame rates are usually used.

A Microsoft Kinect sensor is used to ensure ground truth and assess the motion estimation performances. It is calibrated using Burrus (2011) so that each pixel in the light intensity image is mapped to the depth image. Hence, the trajectory of the robot moving in the scene can be estimated directly. Tracked motions are also constrained within the Kinect's range of accuracy, from 0.5 to 3 m in front of the device, for which the estimation error and its variance are less than 5% (see Khoshelham, 2011; Panaite et al., 2011).

3.1.1. Results Five types of motions are executed by the mobile robot with the plenoptic system embedded on it, and

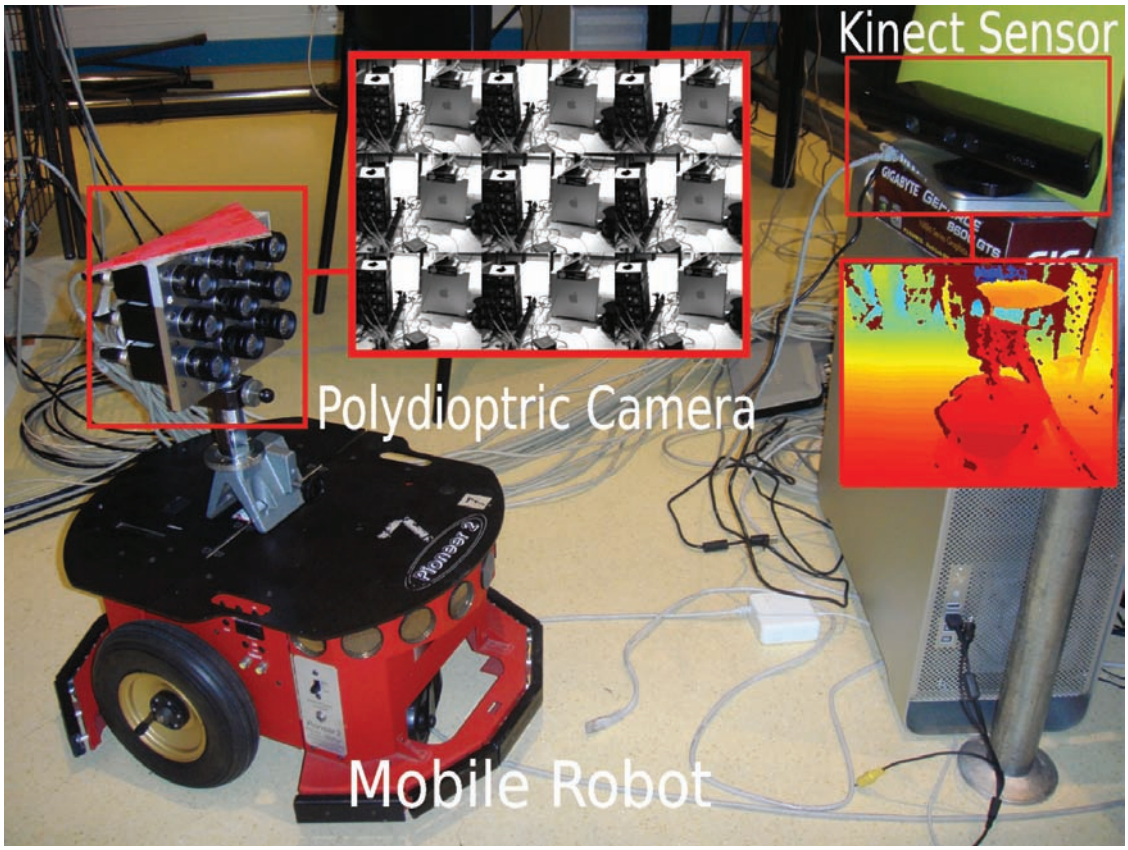

Fig. 3. Experimental setup: a 3×3 camera array on a mobile robot monitored by a Kinect sensor. The camera array captures nine video sequences at 20 fps while the Kinect sensor provides the ground truth of its motion.

for each motion a total of six trials are performed. Figure 4 shows a set of circular trajectories corresponding to six trials. Each of them are estimated with the plenoptic visual odometry for a total of 100 samples, only a fraction of them are plotted as markers for readability purpose.

Figure 5 shows the histogram of the translational errors for all of the mentioned motions. These errors are represented as absolute errors with a mean value equals to 0.07 m. All trajectories have an average length of 10 m. We have also applied the monocular algorithm designed by Civera et al. (2010), which has been largely tested and used for robotic navigation, on the same set of data. The estimated motion parameters obtained using either technique are compared with the same ground truth and the normalized errors are shown in Figure 6. Finally and because it is usually claimed as being the gold standard in structure from motion (SFM) problems, the bundle adjustment is known to provide reliable and stable solutions. We then also apply to the data the monocular simultaneous localization and mapping (SLAM) algorithm presented by Strasdat et al. (2010). As one can expect, the error curve is more stable, the algorithm even produces the best performances at the beginning of the motion and performs slightly worse than the plenoptic method as the distance increase.

The relative errors are normalized by the traveled distance. The plenoptic algorithm is performing better than the monocular algorithm as the error is always lower. The bundle adjustment algorithm proves to work efficiently by giving the lowest mean error at short distances. As distance increases, the bundle adjustment algorithm error is slightly above the plenoptic estimation's error. The mean error is around 2.85% for the plenoptic estimation while Civera et al.'s monocular SLAM provides a mean error of 6%. Strasdat et al.'s algorithm is giving a lower mean error of 2.5%. The decreasing behavior of the curves underlines the stability of the estimations techniques: the errors are decreasing as the lengths of the trajectories increase, meaning that the errors are bounded all of the time.

The monocular algorithm performances drop as the motion speed increases. Above the speed of 1 m s^{-1}, the monocular algorithm often fails to estimate the motion. One main reason for this failure is the difficulty in setting the initial guess of the Kalman filter used in the algorithm. Without prior knowledge of the motion, the algorithm diverges at high speed if the initial guess is chosen loosely. Another reason for the poor performances at high speed is due to motion blur which increases with the velocity and causes feature tracking to fail when only one camera is used.

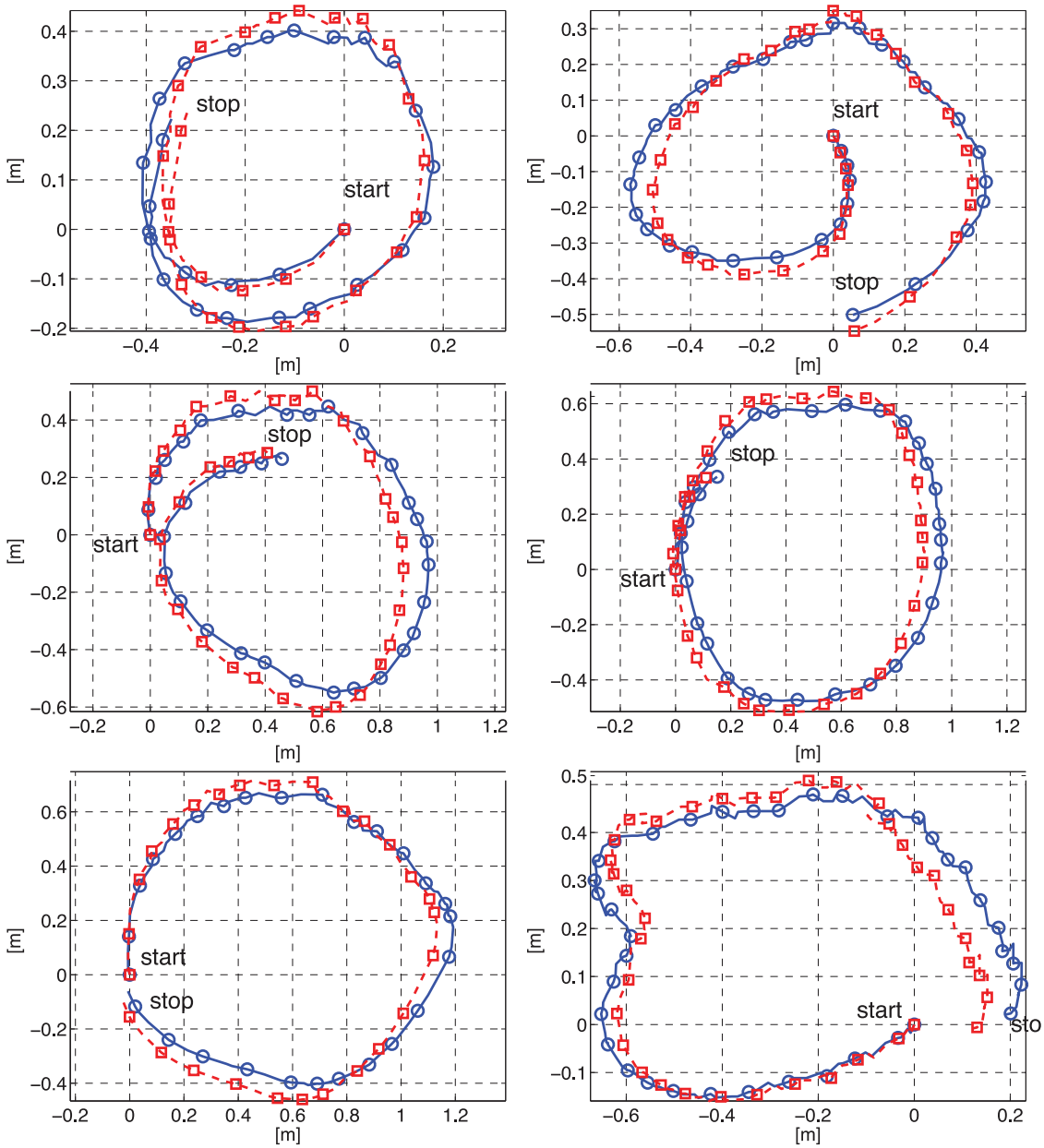

Fig. 4. Six circular trajectories built from the Kinect (circles) and the plenoptic estimation (squares).

A camera array used in conjunction with the plenoptic function is a featureless technique, so it is more robust to large displacements as explained in Section 2.1.

3.2. Optimal setup

Motion estimation accuracy varies according to the motion speed, sensor resolution, and field of view. Although there is no simple way to express analytically the estimation error with respect to those variables, optimal parameters must be identified in order to produce the most accurate motion estimation. As discussed in previous sections, the multi-scale approach aims to recover the motions accurately while avoiding having to process images at unnecessarily high-resolution levels. However, this is still a computationally expensive process that prevents real-time applications. By identifying optimal parameters for certain motions, it is possible to reduce the search domain and more important to decrease the processing time.

3.2.1. Impact of resolution The influence of the resolution is examined in this section. The estimation error f is expressed as a function of the scale factor and the velocity. The optimal parameters are determined by minimizing f with a standard gradient descent:

$$f : \begin{array}{ccc} \mathbb{R}^2 & \rightarrow & \mathbb{R} \\ \mathbf{X} & \mapsto & f(\mathbf{X}) \end{array}, \tag{5}$$

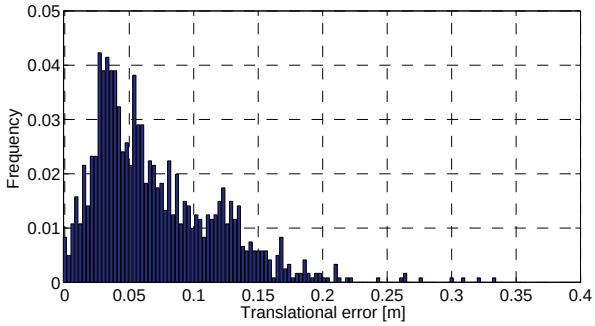

Fig. 5. Histogram of the translational absolute errors estimated by the plenoptic algorithm, for the six trials of the five motions. The average length of the trajectories is 10 m and the mean estimation error is equal to 0.07 m.

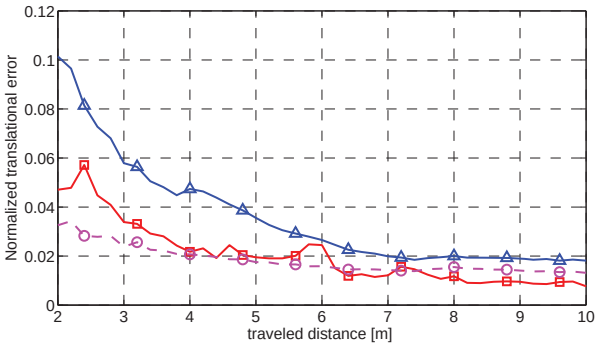

Fig. 6. Normalized translational errors with respect to the traveled distance. The triangles show the estimation results from monocular algorithm, the squares show the plenoptic ones and finally the circles represent the results from the SLAM algorithm incorporating the bundle adjustment. The errors are computed for five types of trajectories, each repeated six times.

where

$$f(\mathbf{X}) = \frac{||(\mathbf{q}(\mathbf{X}), \boldsymbol{\omega}(\mathbf{X}))^{\mathrm{T}} - (\tilde{\mathbf{q}}(\mathbf{X}), \tilde{\boldsymbol{\omega}}(\mathbf{X}))^{\mathrm{T}}||^2}{||(\mathbf{q}(\mathbf{X}), \boldsymbol{\omega}(\mathbf{X}))^{\mathrm{T}}||^2},$$

and $\mathbf{X} = (k, v)^{\mathrm{T}}$ is the vector of the scale factor and the velocity. The scale factor is the same as that introduced as the scale space parameter. Figure 7 shows the error function established for velocity ranging from 0.2 to 5 m s^{-1} and resolution from (320×240) to (5×4) pixels. A local minimum of $f(\mathbf{X})$ is found at $\mathbf{X} = (4.46, 1.32\ \mathrm{m} \cdot \mathrm{s}^{-1})$, i.e. at a resolution of 29×22. This minimum is also confirmed by the experimental results. The algorithm 1 selects frequently the level 20×15 when the speed $\in [0.9, 2]$ m s^{-1}.

The motion estimation accuracy is usually expected to be an increasing function of the sensor resolution. However, experiments show surprising results where lower resolutions yield superior results than higher resolutions. This counter-intuitive result can be explained by the fact that given the robot's speed and the scene content, the egomotion estimations are often biased by local minor motions or lighting changes. Reducing the resolution filters out these perturbations. This process has a limit as shown in Figure 7. A resolution that is too low tends to provide highly imprecise estimations. Hence, the estimation error has the shape depicted in Figure 7, with a local minimum corresponding to the optimal parameters. We may then wonder whether there exists a methodology to predict and exploit this relationship, rather than using the brute-force search presented? Unfortunately, this behavior is largely dependent on the scene contents for which we usually do not have control. It is then appears difficult to predict an a priori shape to the estimation error. However, for large classes of scene a "mean" optimal parameter set could probably be estimated.

3.2.2. Impact of the field of view The field of view is one of the critical parameters for motion estimation problems. It is known that visual sensors with narrow field cannot distinguish translations from rotations (the aperture problem (Hildreth, 1984)). Increasing the field of view can however significantly decrease these ambiguities. In this experiment we start by studying its influence by iteratively cropping the original images from their maximal resolution of 320×240 pixels to artificially decrease resolution. Each camera is mounted with identical lenses providing a 35° field of view both vertically and horizontally. Figure 8 shows the motion estimation results for arbitrary translations with different values of the field of view expressed as ratios of the maximal value (i.e. 35°).

From the error curves, one can see that the motion estimation is the most accurate when the sensor is working with its maximum field of view. The lower the field, the faster the estimation performance decreases with estimation errors higher than 10%. For too small values of the field of view the algorithm simply fails because of the lack of sufficient overlapping areas. It also induces a subsequent reduction in the visual cues required for the photo consistency.

The nine used cameras have fixed focal length lenses, so it was not possible to study the impact of higher values of the field of view. In order to provide upper bound results, we generated data from artificial scenes filled with textured spheres randomly distributed. In this second experiment the field of view is studied for a range of 10–150°, with cameras' spacing ranging from 5 to 30 cm. The remaining parameters are set to the same values as those of the developed plenoptic sensor. The virtual camera velocity is set to 1 m s^{-1}. The multi-resolution algorithm is applied to estimate motion providing mean errors shown in Figure 9.

A mean error of 12% is achieved for a field of view of 35° and camera spacing of 5 cm. This is also consistent with the results presented on real data in Section 3.1.1. The error curve reported by Neumann et al. (2002) provide similar conclusions, the error decreases as the field of view becomes larger. It is also less sensitive to the camera spacing at large values of the field of view. The multi-resolution method reduces the amount of errors by 10%. This can be explained by the optimization process that selects the most

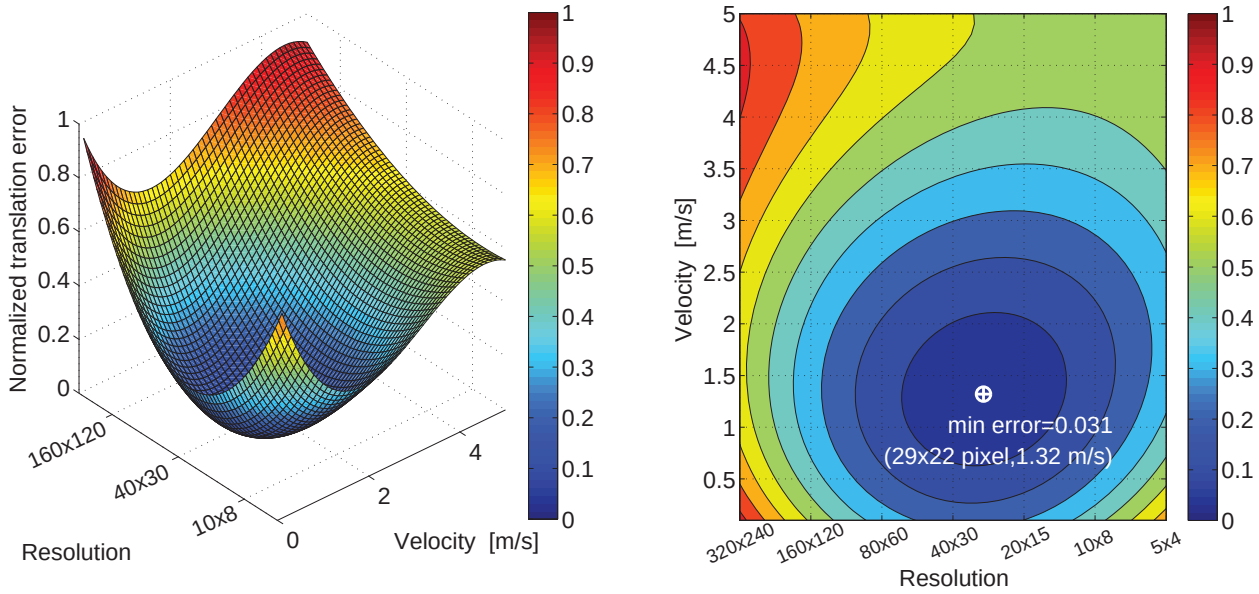

Fig. 7. Estimation errors with respect to the signal resolutions and motion velocities. The plenoptic sensor is particularly adapted for speeds from 1 to 2 m s^{-1} for resolutions between 40 × 30 and 20 × 15 pixels.

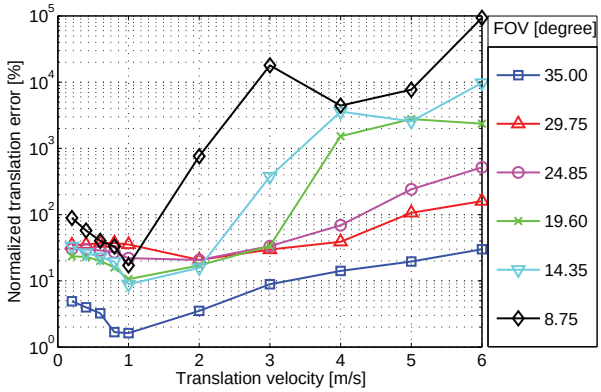

Fig. 8. Normalized translation errors plotted with a logarithmic scale to make the errors at slow motions (< 1 m s^{-1}) visible. Each curve has a fixed field of view ratio value.

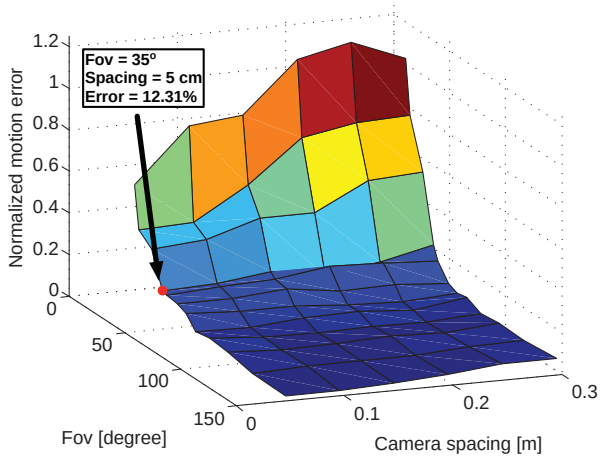

Fig. 9. Mean estimation error computed using data from artificial scenes filled with textured spheres randomly distributed. The error is plotted as function of the field of view and the camera spacing. The red dot shows the estimation mean error when the field of view and the camera spacing are set to the same values as those in the real plenoptic sensor.

suitable resolution level given the speed of 1 m s^{-1}. According to the error curve, one can expect with this approach to go below 5% of error with a 65° field of view.

We intentionally discarded the study of spatial arrangement of cameras in order to optimize the field of view, this goes beyond the scope of this paper, as the practical experiments and comparisons would then have to use omnidirectional sensors for comparisons, which is a complete other topic.

3.2.3. Processing time The plenoptic motion estimation algorithm has been tested on a Core 2 Duo, running at 2.40 GHz, the algorithm is implemented in Matlab without specific optimization. The motion estimation is performed independently for each resolution level, starting from 5 × 4 to 320×240 pixels. Regardless of the estimations' accuracy, the mean processing times at each step (i.e. pair of images) are measured for all of the images of each sequence and shown in Table 1. The mean processing time is also measured for both the monocular technique that also relies on a Matlab implementation and the multi-scale plenoptic algorithm which is supposed to provide the optimal results with the shortest processing time. The results are shown in the two last rows of Table 1.

These results show that by selecting the optimized resolution (here 20 × 15), accurate motion estimations can be achieved. Compared with the multi-scale technique, the

Table 1. Mean processing times for the single resolution plenoptic motion estimation algorithm at each resolution level compared with the multi-scale and monocular algorithms.

Algorithm	Resolution (pixels)	Mean processing time (s)
Mono-scale plenoptic	320×240	7.596
	160×120	1.890
	80×60	0.468
	40×30	0.118
	20×15	0.032
	10×8	0.011
	5×4	0.005
Monocular	320×240	1.320
Multi-scale plenoptic	n/a	3.154

Table 2. Optimal setups according to motion speeds. Three domains of velocity define the optimal estimation parameters which allows for accurate estimation and short processing times.

	Optimal sensor setup		
Motion speed ($m.s^{-1}$)	< 0.7	$\in [0.7, 2]$	$\in [2, 2.6]$
Center resolution (*pixel*)	40×30	20×15	40×30
Field of view	Maximum	Maximum	Maximum
Processing time (s)	0.12	0.03	0.12
Mean error	0.13	0.11	0.13

processing time is also shorter by a factor 100. Beyond the benefit of a drastic reduction in processing time, we also gain accuracy in the motion estimation. In our experiments, the plenoptic algorithm works remarkably well at the resolutions of 40×30 and 20×15 pixels, with a mean processing time of 0.03 s, while the monocular algorithm works at 1.32 s and the multi-scale at 3.15 s.

3.2.4. Proposed setup The impact of image resolution and field of view were analyzed in previous sections. These tests help to determine the optimal setup to produce correct estimation at an optimized computational time. The minimization of the error function in Equation (5) allows us to find the adequate parameters $\mathbf{X} = \text{argmin}\, f(\mathbf{X})$ where $\mathbf{X}$ is the vector of the sensor parameters (scale level and field of view) and the motion velocity. The motion velocity is not a parameter of the sensor, however it is of great importance to the vision system. The velocity range is in principle known from the specification of the robot's architecture. This brings relevant information to properly initialize the motion estimation process.

Table 2 can be seen as a chart to help design adapted vision sensors for a specific robotic task. It summarizes the setup producing the lowest mean estimation error according to the motion speed. The presented work dealt with sets of similar cameras (i.e. similar intrinsic parameters, regular spacing) as the paper's initial goal was the study of the computation properties of insects' compound eyes that generally rely on similar visual units. More general configuration would be interesting to study: we believe they are linked to higher "cognitive" tasks such as predation, reproduction, etc.

The error function is minimized over the variable $\mathbf{X}$ with a gradient descent to find the local minimum for each speed interval: < 0.7 m s^{-1}, $[0.7, 2]$ m s^{-1} and $[2, 2.6]$ m s^{-1}. For a standard wheeled robot (e.g. Pioneer, with a maximum speed of 1.8 m s^{-1}), the optimal choice will be a 40×30 pixel plenoptic camera, it produces the best balance between processing speed and accuracy. In the case of fast robots such as flying unmanned aerial vehicles (UAVs) 20×15 pixel plenoptic camera with a wide field of view can significantly reduce the processing time while maintaining an error estimation of less than 12%. For higher-speed motions, ($v > 2.6$ m s^{-1}), there seems to be no real optimal set of parameters as the sensor estimation errors are always higher than 20%.

4. Discussion

This paper shows a surprising result: that a set of low-resolution cameras can be more efficient for navigation than a high-resolution visual sensor. This can be easily explained by the fact that multiple viewpoint acquisition maximizes the signal-to-noise ratio, accurate conclusions can then be drawn. If the individual measurements of a plenoptic camera have a lot of variability it becomes difficult to determine the useful information needed to perform the task. Measuring a single item or event more than once maximizes the accuracy of results. More measurements of a single event lead indeed to greater confidence in calculating an accurate average estimation.

Another important conclusion of the paper relates to the low amount of data to process. We often think of camera-like eyes when dealing with perception, but nature contains several types of eyes, with an incredible variety of designs as shown in the work of Land and Nilsson (2002). Compound eyes directly related to this work are found in the insect world. Their visual acuity is around 100 times less than that of the human eye. The compound eyes are however known to be excellent at detecting motion (Dawkins, 1996). Compound eyes resolution is particularly adapted to the low neural processing ability of insects brain. Each facet of a compound eye encloses one single visual unit (ommatidium) containing around 7–11 sensory cells, this corresponds well to the limited processing power of the insect brain.

From a technical point of view more adapted cameras could be used. Figure 1 shows a design of the camera system where each camera is a completely closed unit, composed of a unique lens and image sensor. A more adapted camera could be built to achieve a similar 3×3 plenoptic camera system using a single image sensor and an array of micro lenses or pin-holes. Current CMOS technology

allows video image sensors with as many as 10 million pixels over a chip size of 2.5 cm $\times$ 1.4 cm (Thorpe, 2011). Such a camera can easily be organized as a plenoptic camera system when coupled with the appropriate optics. Typically for these types of large format image sensors, the entire image can be readout from the image sensor, or be down-sampled by the readout integrated circuit, as mandated by a feedback loop that sets the required resolution as a function of the estimated visual motion. Hence, a single chip plenoptic camera system using the standard video imaging chips can readily be built and coupled with the visual motion estimation and optimization algorithms discussed above for applications in mobile robotics.

Another major property of insect eyes is their high temporal resolution that was not inquired in this paper. As an object moves across the visual field of a compound eye, ommatidia are progressively turned on and off. Because of the resulting "flicker effect", insects respond far better to moving objects than stationary ones. It is known that honeybees will visit wind-blown flowers more readily than still ones. The compound eyes can detect fast moving objects with high precision, contour detection is sought to be the two most important characteristics of compound eyes. Non-scanned image sensors that use the address event representation (AER) protocol to asynchronously output the pixels are also ideally suited for plenoptic cameras (Culurciello et al., 2003; Lichtsteiner et al., 2006). In such cameras, the pixels' addresses, i.e. location in the array, and in some cases the pixel values, are output when the pixels deem that enough light has been integrated by their photo-detectors. These sensors compress information at the level of pixels and have a tremendous temporal resolution of several kilohertz.

5. Conclusion

In this paper we have presented an in-depth exploration of the plenoptic approach of a vision-based navigation. The limited field of view of a perspective camera was pointed out as one of the main causes of the poor performance of many widespread vision algorithms but several factors are also examined and are shown to contribute to the motion estimation quality: the sensor's resolution, the motion velocities of the sensor while moving, etc. Experiments preformed in real conditions with a mobile robot show the plenoptic function approach often outperforming the state-of-the-art vision algorithms. The coarse to fine scheme of the approach produces a bottleneck in the motion estimation process, but it is only exploited to identify the ideal parameters allowing to use the vision system at its highest accuracy with minimal data payload. With the adequate parameters for the cameras, we have shown that a plenoptic vision system is not only a more accurate sensor for motion recovery from the images, but it also has the ability to handle larger motions. Most interestingly, such performances are achieved with an extremely low processing power requirement, thus making it perfectly suitable for realtime robotics.

The plenoptic vision sensor is presented as the optimal technological solution to the problem of vision-based motion estimation, raising several promising expectations. While classical vision-based navigation techniques struggles to perform as efficiently as living organisms, despite the piling up of complex processing techniques, the plenoptic formulation offers an interesting alternative to solving the problem. This work highlights the irrelevance of ever-increasing resolution cameras to achieve accurate motion estimation and reinstate the benefit of using a plenoptic vision system in robotic applications. With properly defined optical parameters, it is possible to build a reactive and accurate vision navigation system requiring reasonable computation resources.

Funding

This work was supported by the Council of the Large Research Network in Energy, Electronics and Materials of the French region of Haute Normandie and the CISE-LNA project, an implementation of an autonomous navigation laboratory.

Acknowledgment

The authors are grateful to Kathrine Matho for proofreading the paper.

References

Adelson EH and Bergen JR (1991) The plenoptic function and the elements of early vision. In: *Computational Models of Visual Processing*. Cambridge, MA: MIT Press, pp. 3–20.

Adelson EH and Wang JYA (1992) Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and machine Intelligence* 14: 99–106.

Borst and Egelhaaf (1993) *Detecting Visual Motion: Theory and Models*, chapter 1. Amsterdam: Elsevier, pp. 3–27.

Brady D and Morrison RL (2000) Diffractive and micro-optics for multiplex imaging. In: *Diffractive Optics and Micro-Optics*, T. Li, ed., Vol. 41 of OSA Trends in Optics and Photonics (Optical Society of America, 2000)

Burrus N (2011) Kinect calibration. http://nicolas.burrus.name/index.php/Research/KinectCalibration.

Camahort E and Fussell D (1999) *A Geometric Study of Light Field Representations*. Technical Report TR99-35, Department of Computer Sciences, The University of Texas at Austin.

Civera J, Grasa OG, Davison AJ and Montiel JMM (2010) 1-Point RANSAC for extended Kalman filtering: Application to real-time structure from motion and visual odometry. *Journal of Field Robotics* 27: 609–631.

Culurciello E, Etienne-Cummings R and Boahen K (2003) An address event digital imager. *IEEE Journal of Solid-State Circuits* 38: 281–294.

Dansereau I, Mahon DG, Pizarro O and Williams S (2011) Plenoptic flow: Closed- form visual odometry for light field cameras. In: *Proceedings of Intelligent Robots and Systems (IROS)*.

Dawkins R (1996) *Climbing Mount Improbable*. New York: Norton.

Duparré J, Dannberg P and Schreiber P (2004) Ultra-thin camera based on artificial apposition compound eyes. In: *Proceedings of the 10th Microoptics Conference (MOC'04)*, Friedrich-Schiller-University, Germany.

Forster LM (1979) Visual mechanisms of hunting behaviour in *Trite planiceps*, a jumping spider (Aranae: Salticidae). *New Zealand Journal of Zoology* 6: 79–93.

Franceschini N (2008) Towards automatic visual guidance of aerospace vehicles: from insects to robots. *Acta Futura* 3: 12–28.

Georgiev T and Lumsdaine A (2010) Focused plenoptic camera and rendering. *Journal of Electronic Imaging* 19(2): 21106.

Gortler S and Grzeszczuk R (1996) The lumigraph. *Computer Graphics (Proceedings SIGGRAPH)*, pp. 43–54.

Hassenstein B and Reichardt W (1956) Systemtheoretische analyse der zeit, reihenfolgen, und vorzeichenauswertung bei der bewegungsperzepion des russelkafers chlorophanus. *Zeitschrift für Naturforschung* 11b: 513–524.

Hildreth E (1984) *The Measurement Of Visual Motion*. MIT Press.

Iida F and Lambrinos D (2000) Navigation in an autonomous flying robot using a biologically inspired visual odometer. In: *Sensor Fusion and Decentralized Control in Robotic System III Photonics East* (*Proceedings of SPIE*, vol. 4196), pp. 86–97.

Jackson DPH and R R (2002) Influence of cues from the anterior medial eyes of virtual prey on Portia fimbriata, an araneophagic jumping spider. *The Journal of Experimental Biology* 205: 1861–1868.

Khoshelham K (2011) Accuracy analysis of kinect depth data. In: *ISPRS Workshop Laser Scanning*.

Land MF and Collett TS (1974) Chasing behaviour of houseflies. *Journal of Comparative Physiology A* 89: 525–538.

Land MF and Nilsson DE (2002) *Animal Eyes*. Oxford: Oxford University Press.

Levoy M and Hanrahan P (1996) Light field rendering. In: *SIGGRAPH '96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM Press, pp. 31–42.

Lichtsteiner P, Posch C and Delbruck T (2006) A 128 × 128 120 dB 30 MW asynchronous vision sensor that responds to relative intensity change. In: *Digest of Technical Papers. Proceedings IEEE International Solid-State Circuits Conference ISSCC 2006*, pp. 2060–2069. DOI: 10.1109/ISSCC.2006.1696265.

Lindeberg T (1994) *Scale-Space Theory in Computer Vision*. Norwell, MA: Kluwer Academic Publishers.

Liu LC and Usseglio-Viretta A (2001) Fly-like visuomotor responses of a robot using a vlsi motion-sensitive chips. In: *Biological Cybernetics* 85: 449–457.

Neumann J (2004) *Computer Vision in the Space of Light Rays: Plenoptic Videogeometry and Polydioptric Camera Design*. PhD Thesis, University of Maryland.

Neumann J, Fermuller C and Aloimonos Y (2002) Eyes from eyes: new cameras for structure from motion. In: *Proceedings of the Workshop on Omnidirectional Vision*. Los Alamitos, CA: IEEE Computer Society.

Neumann J, Fermuller C, Aloimonos Y and Brajovic V (2004) Compound eye sensor for 3D ego motion estimation. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

Ng R, Levoy M, Bredif M, Duval G, Horowitz M and Hanrahan P (2005) *Light Field Photography with a Hand-held Plenoptic Camera*. Technical Report, Standford University.

Okoshi T (1976) *Three-dimensional Imaging Techniques*. New York: Academic Press.

Olberg RM, Worthington AH and Venator KR (2000) Prey pursuit and interception in dragonflies. *Journal of Comparative Physiology A* 186: 155–162.

Panaite J, Usciati T, Clady X and Haliyo S (2011) An experimental study of the Kinect depth sensor. In: *Robotic and Sensors Environments*.

Raboud D (2009) *The Panoptic Camera: Plenoptic Interpolation in an Omnidirectional Polydioptric Camera*. Master's thesis, EPA.

Raskar R, Agrawal A, Wilson CA and Veeraraghavan A (2008) Glare aware photography: 4D ray sampling for reducing glare effects of camera lenses. *ACM Transactions on Graphics* 27: 56:1—-56:10.

Rivas Lopez OS M and Tyrsa V (2008) *Computer Vision*. InTech.

Ruffier F and Franceschini N (2003) Octave: a bioinspired visuomotor control system for the guidance of micro-air-vehicles. In: *Proceedings of Bioengineered and Bioinspired Systems Conference*, pp. 1–12.

Srinivasan MV, Zhang SW, Altwein M and Tautz J (2000) Honeybee navigation: Nature and calibration of the "odometer". *Science* 287: 851–853.

Strasdat H, Montiel JMM and Davison AJ (2010) Scale drift-aware large scale monocular slam. In: *Robotics: Science and Systems VI*, Universidad de Zaragoza, Zaragoza, Spain, 27–30 June 2010.

Svoboda T, Martinec D and Pajdla T (2005) A convenient multicamera self-calibration for virtual environments. *PRESENCE: Teleoperators and Virtual Environments* 14: 407–422.

Tanida J and Yamada K (2002) TOMBO: thin observation module by bound optics. In: *The 15th Annual Meeting of the IEEE Lasers and Electro-Optics Society, 2002 (LEOS 2002)*, vol. 1, pp. 233–234.

Thorpe L (2011) *New 35mm CMOS Image Sensor for Digital CINE Motion Imaging*. Canon Withepaper, Canon. http://learn.usa.canon.com/app/pdfs/white_papers/EOS_C300_New_35mm_CMOS_Sensor_WP.pdf.

Tinbergen N (1980) *Animal Behavior*, 2nd edition. Time Life.

VanSanten JP and Sperling G (1984) A temporal covariance model of human motion perception. *Journal of the Optical Society of America* 1: 451–473.

Völkel R and Duparré J (2008) Technology trends of microlens imprint lithography and wafer level cameras (WLC). In: *Proceedings of the 14th Microoptics Conference (MOC'08)*, Brussels, Belgium.

Wei Chung W, Schenato L, Wood RJ and Fearing RS (2003) Biomimetic sensor suite for flight control of a micro mechanical flying insect: design and experimental results. In: *IEEE International Conference on Robotics and Automation (ICRA'03)*, pp. 1146–51.

Wilburn B (2004) *High Performance Imaging Using Arrays of Inexpensive Cameras*. PhD thesis, Stanford University.