Intelligent Robots need Intelligent Vision: Visual 3D Perception

Geert De Cubber[1], Lazaros Nalpantidis[2], Georgios Ch. Sirakoulis[3] and Antonios Gasteratos[2]
[1]Royal Military Academy
Department of Mechanical Engineering (MSTA)
Av. de la Renaissance 30, 1000 Brussels
Geert.De.Cubber@rma.ac.be
[2]Democritus University of Thrace
Department of Production and Management Engineering
lanalpa@pme.duth.gr, agaster@pme.duth.gr
[3]Democritus University of Thrace
Department of Electrical and Computer Engineering
gsirak@ee.duth.gr

Keywords

Stereo Vision, Structure from Motion, Dense depth estimation, Integrated depth perception

1. Introduction

Contemporary autonomous robots are generally equipped with an abundance of sensors like for example GPS, Laser, ultrasound sensors, etc to be able to navigate in an environment. However, this stands in contrast to the ultimate biological example for these robots: us humans. Indeed, humans seem perfectly capable to navigate in a complex, dynamic environment using primarily vision as a sensing modality. This observation inspired us to investigate visually guided intelligent mobile robots.

In order to *understand* and reason about its environment, an intelligent robot needs to be aware of the three-dimensional status of this environment. The problem with vision, though, is that the perceived image is a two-dimensional projection of the 3D world. Recovering 3D-information has been in the focus of attention of the computer vision community for a few decades now, yet no all-satisfying method has been found so far. Most attention in this area has been on stereo-vision based methods, which use the displacement of objects in two (or more) images. Where stereo vision must be seen as a spatial integration of multiple viewpoints to recover depth, it is also possible to perform a temporal integration. The problem arising in this situation is known as the "Structure from Motion" (SfM) problem and deals with extracting 3-dimensional information about the environment from the motion of its projection onto a two-dimensional surface.

In this paper, we investigate the possibilities of stereo and structure from motion approaches. It is not the aim to compare both theories of depth reconstruction with the goal of designating a winner and a loser. Both methods are capable of providing sparse as well as dense 3D reconstructions and both approaches have their merits and defects. The thorough, year-long research in the field indicates that accurate depth perception requires a combination of methods rather than a sole one. In fact, cognitive research has shown that the human brain uses no less than 12 different cues to estimate depth. Therefore, we also finally introduce in a following section a methodology to integrate stereo and structure from motion.

The idea of this approach is sketched on Figure 1. The setup we consider for this approach is that of a stereo vision system installed on a mobile robot. While navigating in the environment, the different camera images can be interpreted in several ways to obtain a structural 3D view of the environment. Stereo vision employs the differences in the left and right camera images to obtain a sparse 3D reconstruction, which can be upgraded to a dense reconstruction. In a similar manner, SfM employs the movement between two successive camera images to infer sparse, and eventually also dense, depth information. Finally, all these cues are integrated to obtain a better reconstruction result, bringing together the advantages of stereo and SfM- based approaches.

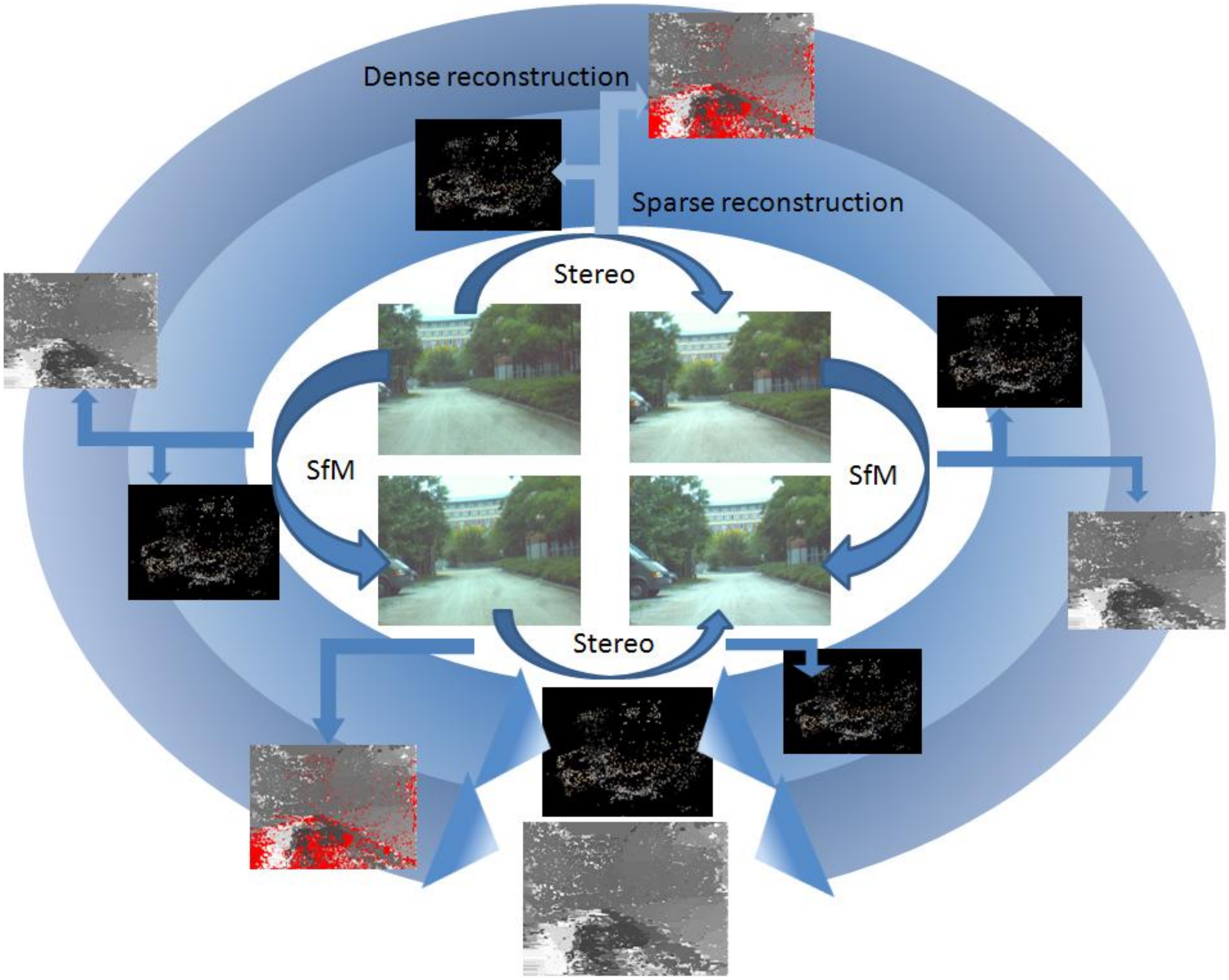

Figure 1: Overview of the sparse and dense 3D reconstruction techniques discussed in this paper: Stereo vision extracts 3D information from left-to-right image correspondences, whereas Structure from Mtion (SfM) uses inter-frame correspondences. Integration of these results can lead to optimized 3D reconstructions.

The remaining of this paper is organized as follows: sections 2 and 3 respectively explain the stereo-vision and SfM-based 3D reconstruction approaches in more detail. Section 4 explains how the results of both methods can be fused to from an optimized model.

2. Stereo Vision

Stereo vision deals with the issue of stereo correspondence and depth estimation. It is of great importance in the field of machine vision, virtual reality, robot navigation and environment reconstruction as well as in many other aspects including production, security, defence, exploration and entertainment. Calculating the distance of various points or any other primitive in a scene relative to the position of a camera is one of the important tasks of a computer vision system. The most common method for extracting depth information from intensity images is by means of a pair of synchronized camera-signals, acquired by a stereo rig [3]. The point to point matching between the two images from the stereo setup derives the disparity map. The difference on the horizontal coordinates of these points is the disparity. The disparity map consists of all disparity values of the image. Having extracted the disparity map, problems such as 3D reconstruction, positioning, mobile robot navigation, obstacle avoidance, etc., can be dealt with in a more efficient way.

To determine the correspondence of two points lying in the two images, it is necessary to measure the similarity of the points. The point to be matched without any ambiguity should be distinctly different from its surrounding pixels. Several algorithms have been proposed in order to address this problem. However, every algorithm makes use of a matching cost function so as to establish correspondence between two pixels [4]. The most common ones are absolute intensity differences (AD), the squared intensity differences (SD) and the normalized cross correlation (NCC). The selection of the appropriate disparity value for each pixel is performed afterwards. In many cases this is an iterative process. An additional disparity refinement step is frequently used.

Stereo correspondence algorithms can be grouped into those producing sparse output and those giving a dense result. Feature based methods stem from human vision studies and are based on matching segments or edges between two images, thus resulting in a sparse output [12]. This disadvantage, dreadful for many purposes, is counterbalanced by the accuracy and speed obtained. However, contemporary applications demand more and more dense output. In order to categorize and evaluate stereo algorithms producing dense output, a context has been proposed [9]. According to this, dense matching algorithms are classified in local and global ones. Local methods (area-based) swap accuracy with speed. The disparity computation at a given point depends only on intensity values within a finite support window [14]. Global methods (energy-based) are time consuming but very accurate. Their goal is to minimize a global cost function, which combines data and smoothness terms, taking into account the whole image [2]. Of course, there are many other methods that are not strictly included in either of these two broad classes.

Algorithms resulting in sparse disparity maps are very useful when fast depth estimation is required and at the same time detail, in the whole picture, is not so important. This type of algorithms tends to focus on the main features of the images, i.e. edges, leaving occluded and poorly textured areas unmatched.

Disparity and hence reconstructed depth is calculated only for those features. Consequently high processing speeds, accurate and reliable results but with limited density are achieved.

Methods that produce dense disparity maps gain popularity as the computational power grows. Moreover, contemporary applications are benefited by, and consequently demand dense depth information. This kind of algorithms assigns disparity values for almost every pixel of the input images. On the other hand they are more computational demanding compared to the sparse ones. Autonomous robot navigation requires high execution speed and dense information in order the robot to be able to navigate itself in real-time. The most appealing choice is a local, Sum of Absolute Differences (SAD) based method. Such a method is rapidly executed and its results can be further improved in a post-processing stage.

It is obvious that both sparse and dense output producing algorithms have their own advantages and limitations. Methods leading to sparse reconstruction are generally fast and accurate. On the other hand dense reconstruction, despite being more computation and time demanding, is more suitable for contemporary applications, such as robot navigation.

2. Structure from Motion

Stereo vision relates two images which are displaced slightly in place to recover depth information from the difference in projection on the 2 camera image fields. When only one camera is present, it is also possible to perform the same relation between two camera images from the same imaging device which are displaced slightly in time. The problem arising in this situation is known as the ”Structure from Motion” (SfM) problem and deals with extracting 3-dimensional information about the environment from the motion of its projection onto a two-dimensional surface. The extra difficulty as opposed to stereo vision is that in the case of SfM the spatial relationship between the two camera viewpoints is a priori unknown. Indeed, in stereo vision, the stereo rig is generally a fixed setup of 2 cameras, whereas in SfM the camera undergoes an unknown motion pattern between 2 shots. This means that the camera motion first needs to be estimated precisely before 3D reconstruction is possible.

In general, there are two approaches to SfM. The first, feature based method is closely related to stereo vision. It uses corresponding features in multiple images of the same scene, taken from different viewpoints. The basis for feature-based approaches lies in the early work of Longuet-Higgins [7], describing how to use the epipolar geometry for the estimation of relative motion. These techniques have matured a lot over the past two decades, but – of course - they only deliver sparse 3D information. The second approach for SfM uses the optical flow field as an input instead of feature correspondences. Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image. Optical flow can arise from relative motion of objects and the viewer. Consequently, optical flow can give important information about the spatial arrangement of the objects viewed and the rate of change of this arrangement. The applicability of the optical flow field for SfM calculation originates from the epipolar constraint equation which relates the optical flow, or the apparent image motion field, to the relative camera motion (translation and rotation) and 3D structure in a non-linear fashion.

Sparse feature-based approaches are based upon the work of Longuet-Higgins [7], introducing the 8-points algorithm. It features a way of estimating the relative camera motion, using the essential matrix, which constrains feature points in two images. The first problem with these feature based techniques is of course the retrieval of correspondences, a problem which cannot be reliably solved in image areas with low texture. From these correspondences, estimates for the motion vectors can be calculated, which are then used to recover the depth. An advantage of feature based techniques is that it is relatively easy to integrate results over time, using bundle adjustment [11] or Kalman filtering [6]. Bundle adjustment is a maximum likelihood estimator that consist in minimizing the re-projection error. It requires a first estimate of the structure and then adjusts the bundle of rays between each camera and the set of 3D points.

For dense 3D reconstruction from motion, the first question to ask is what kind of input data one could use. For motion, what is available to people is the optical flow, which is the best we can hope to recover starting from the intensity images alone. Optical flow is defined [5] as apparent motion of brightness patterns observed when a camera is moving relative to the objects being imaged. It can be represented with a two-dimensional velocity vector associated with each point on the image plane. It contains important information about cues for region and boundary segmentation, shape recovery, and so on. A dense optical flow field provides an excellent starting point for 3D reconstruction. In the literature, there exist numerous computational approaches for estimating optical flow, The optical flow is related to the structure and motion parameters through the rigid motion equation. This allows us to write the depth d of each image pixel $\boldsymbol{x}$ as a function of camera translation $\boldsymbol{t}$ and rotation $\boldsymbol{\varpi}$ and optical flow $\boldsymbol{u}$: $d=d(\boldsymbol{x},\boldsymbol{t},\boldsymbol{\varpi},\boldsymbol{u})$. We can now exploit this information to do the 3D reconstruction, as knowing the optical flow and the motion parameters, the depth field can calculated directly. This process is called backprojection. The problem is that backprojection alone is very sensitive to errors. In practice, to reconstruct a dense depth field, it is necessary to maximize the information which can be retrieved out of the given data. Dense 3D reconstruction algorithms can be roughly subdivided into two categories depending on how they deal with this problem. On one hand, there are volume based approaches [10]. Examples are voxel coloring, photo hulls, and level sets. They use a large amount of images integrated into a single scheme. These approaches use a discretized volume and restrict possible depth values (3D points) to a predefined accuracy. A second series of methods are pixel-based approaches [1], which do not need 3D discretization and compute depth (disparity) with higher precision for every pixel. These approaches are often based upon the iterative solution of a partial differential equation (PDE) minimizing an energy functional. Pollefeys presented in [8] an approach to solve the dense reconstruction problem by combining state-of-the-art algorithms for uncalibrated projective reconstruction, self- calibration and dense correspondence matching.

Dense SfM is a fairly young evolution, as it is only recent that the computing power required for these approaches is available to a wider audience. As a result, these approaches are still a field of much research and haven't matured completely. This means that the results are still sometimes unreliable and that computing times are generally large. Sparse SfM approaches on the other hand are nowadays capable of providing qualitative results very fast. However, it must be noted that they only provide sparse data and that most experiments are performed under controlled lab circumstances. More research is needed to improve their performance on complex large outdoor scenes.

3. Integrated Depth Perception

Generally speaking, to estimate 3D motion and structure from multi-view image sequences, it is desirable to fuse stereo and motion constraints to some extent [13]. However, combining motion/stereo constraints from multi-view image sequences requires extra caution. This is because some points in the reference image may be invisible (occluded) in another view. If the algorithm is not aware of this and still combines the motion and stereo constraints from the occluded view, the results could be very wrong. Stereo and SfM essentially return the same data: a sparse point cloud or a dense surface representation. In the sparse case, the problem is that the SfM point cloud is scaled, whereas stereo can measure absolute depths. Therefore we need to re-scale the SfM and stereo point clouds to the same dimensions. In the dense case, the problem is similar, only the input data is different. Indeed whereas in the sparse case point clouds are considered, dense 3D reconstructions are often not represented as 3D point clouds but as maps of surface normals, curvature-based descriptors, 2D depth images (depth maps), meshes, … In order not to complicate the analysis, we will here consider only 3D point clouds, such that the problems of sparse and dense 3D model integration can be treated together.

A common approach towards the problem of matching 2 point clouds is the Iterative Closest Point (ICP) algorithm presented in [15]. This method optimizes a set of free-form curves represented by a set of chained points to come to an optimal 3D representation. The data of the space curves are available in the form of a set of chained 3-D points from the stereo or SfM algorithm. The key idea underlying the ICP approach is the following. Given that the motion between two successive frames is small, a curve in the first frame is close to the corresponding curve in the second frame. By matching points on the curves in the first frame to their closest points on the curves in the second, we can find a motion that brings the curves in the two frames closer (i.e., the distance between the two curves becomes smaller). Iteratively applying this procedure, the algorithm yields successively better motion estimates. The algorithm to obtain a unified sparse point cloud can thus be summarized as follows [15]:

- Input 2 3D frames of space curves where each curve is a set of chained 3D points obtained from a 3D reconstruction algorithm
- Find the closest points by satisfying distance and orientation constraints
- Update the matching through statistic analysis of distances
- Compute the motion between the 2 frames from the updated matches
- Apply the estimated motion to all points
- Iterate until convergence

4. Conclusions

In this paper we have discussed 2 paradigms for 3D reconstruction: stereo vision and structure from motion. When considering sparse reconstruction, stereo vision is now able to produce excellent results at high framerates as long as the environment is well-textured. Due to the fact that structure from motion algorithms always need an extra processing step to estimate the camera movement, their results are more error prone and require more processing time, although this is still falls into reasonable delays in the sparse case. SfM approaches offer the evident benefit that only 1 simple camera is necessary, whereas a stereo rig is still specialized equipment.

When dense reconstruction is concerned, multiple techniques have been developed to upgrade the sparse stereo reconstruction to a dense one at reasonable framerates. These techniques work quite well in man-made environments where the form of objects can be estimated or where piecewise planarity constraints can be imposed. In natural conditions, more research is still needed. Dense SfM approaches have the major disadvantage of disposing of a very long processing pipeline, making them extremely slow. They also face the same problems as dense stereo reconstruction techniques when presented large outdoor scenes.

The question is not which one of these techniques is the best one. It can be easily shown that humans use both methods to perceive depth. In order for robots to robustly see in 3 dimensions, we also presented an approach to combine the advantages of both approaches in an integrated framework.

5. Acknowledement

The research work presented in this paper has been funded by the View-Finder FP6 IST 045541 Project.

6. Bibliography

[1] L. Alvarez, R. Deriche, J. Weickert, and J. Sanchez. Dense disparity map estimation respecting image discontinuities: A PDE and scale-space based approach. In IAPR International Workshop on Machine Vision Applications, 2000.

[2] M. Bleyer and M. Gelautz. A layered stereo algorithm using image segmentation and global visibility constraints. International Conference on Image Processing, 2004.

[3] O.Faugeras. Three-dimensional computer vision: a geometric viewpoint. MIT Press, Cambridge, Ma, 1993

[4] H. Hirschmuller and D. Scharstein. Evaluation of Cost Functions for Stereo Matching. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

[5] Horn. Robot Vision. MIT Press, Cambridge, Ma, 1979.

[6] H. Jin, P. Favaro and S. Soatto. A semi-direct approach to structure from motion. The Visual Computer, 19(6):377–394, October 2003.

[7] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. Nature, 293(5828):133135, September 1981.

[8] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Automated reconstruction of 3d scenes from sequences of images. ISPRS Journal Of Photogrammetry And Remote Sensing, 55(4):251–267, 2000

[9] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47 (1-3):7-42, 2002.

[10] Christoph Strecha, Tinne Tuytelaars, and Luc Van Gool. Dense matching of multiple wide-baseline views. In ICCV 03: Proceedings of the Ninth IEEE International Conference on Computer Vision, page 1194, Washington, DC, USA, 2003. IEEE Computer Society.

[11] B. Triggs, P.F. Mclauchlan, R.I. Hartley and A.W. Fitzgibbon. Bundle adjustment – a modern synthesis. Lecture Notes in Computer Science, 1883:298 – 372, 2000.

[12] O. Veksler. Dense Features for Semi-Dense Stereo Correspondence. International Journal of Computer Vision 47 (1-3):247-260, 2002

[13] A.M. Waxman and J H Duncan. Binocular image flows: steps toward stereo-motion fusion. IEEE Tran. Pattern Anal. Mach. Intell., 8(6):715–729, 1986

[14] K.J. Yoon and I.S. Kweon. Adaptive support-weight approach for correspondence search. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4):650-656, 2006.

[15] Z. Zhang. Iterative Point Matching for Registration of free-form curves. INRIA Research Report 1658, March 1992