STEREO AND FOCUS TO IMPROVE DEPTH PERCEPTION

JORGE DIAS, H. de ARAÚJO, J. BATISTA, A. de ALMEIDA

Departamento de Engenharia Eletrônica, Universidade de Coimbra,
Largo Marques de Pombal, 3000 COIMBRA, PORTUGAL

Abstract: This work reports an improvement on the computer vision technique to obtain three-dimensional data by using focus and stereo vision techniques simultaneously. The main goal is to combine both techniques to speed up the depth perception algorithm, to decrease the errors on matching and to improve the precision on the depth measure. The algorithms proposed by other authors, normally do not explore the co-operation between different techniques to improve the outputs. The algorithms integrates the outputs only after processing and not along the processing. In this work the stereo vision algorithm combines the output of a visual module (depth from focus) during the processing. This approach has similarities with the multi-sensor fusion techniques, since it explores different techniques to improve the same data. A new constraint for the correspondence problem is established based on the principles of depth from focus and stereo vision geometry. The experimental site is based on a stereo pair of cameras attached to the last link of a manipulator. This type of configuration enables the control of the cameras' position and geometric relations between the vision system and the scene. This fact is fundamental to explore active vision techniques.

1. INTRODUCTION

The majority of actual and old research on computer vision uses static images as input for the algorithms. However the human perception is not static since the visual information is obtained by exploratory movements of the visual system. It is important to transfer this human behaviour to machines. A big step is made in robotics if we build a system that execute intelligent movements based on data provided by visual sensors. Studies on active vision are fundamental to obtain such machines. Exploratory and functional perceptions are two steps of the active vision that are essential for the visual system to operate [2] [19]. Exploratory perception has two main goals -- locating regions of interest and selectively processing them. Visual systems require these capabilities to deal with the computational complexity of visual tasks. This exploratory perception can be described as "smart sensing". Functional perception is another component of active vision and it is related with the top-down process of the images. This component of the active vision process is triggered by the exploratory perception. The functional perception provides also the integration of modular processes, characterising middle-level vision proposed by Marr [15]. This paper presents a technique that can be included on the techniques of the functional perception because integrates two visual modules and can be included on a more complex system doing active vision.

Vision systems are a possible choice to obtain sensorial data about the world in robotics. These systems help to understand the three-dimensional world where these systems develop and to control their movements. One important task in these vision systems is to extract depth information from two-dimensional images, which is normally followed by the three-dimensional reconstruction of the scene. This three-dimensional data is essential for later stages like object recognition, scene interpretation or path planning. To obtain good results in these later stages is essential to have good data as input and this implies to have depth measurements with known uncertainty. In vision, we obtain depth information by using techniques like -- stereo, motion, or focus. Stereo vision is one of the most studied techniques for depth estimation, but the algorithms proposed until now are very dependent of the correspondence technique used. Focus is another possible technique to depth recovery and the algorithms proposed until now normally explore the focus technique by controlling the positions of the sensor plate [9, 13, 18] or to obtain the depth of a single object. This work presents a different approach to the problem by establishing a method to obtain a depth map from focus and to combine it with a stereo vision method. A new constraint for the correspondence problem is established based on the principles of depth map from focus and stereo vision geometry. This new constraint can help in occlusion problems and to control the multiplicity of correspondences.

2. DEPTH FROM FOCUS

Introduction

The focus analysis has been used to automatically focus imaging systems or to obtain coarse depth information from the observed scene. One of the first publications about this type of problems was made by Horn [11]. Horn proposed to focus the imaging systems by using the Fourier transform and analysing the frequency content in the image. Pentland [18] proposed two methods for finding the depth-map of a scene. The first method was based on measuring the "blur" of the edges of a defocused image.

On the other hand, there are several algorithms based on depth from focus, rather than depth from defocus. Some of them use active range finding with infrared or sonar sensors to measure the distance from the camera to the object. This distance is used to changing the position of the lens and to focusing the images. Other algorithms use criterion functions to measure the sharpness of focus that would be equivalent to getting the sharpest focus and after recover the depth of the object. The performance of different focus criterion's functions as well a method to estimate the depth of an image area by actively changing the positions of the lens is evaluated and compared by Krolik [13].

CV-5.7.1
Focused and defocused images

Defocused images are processed versions of images well focused. Normally real imaging systems are composed of several lenses but conceptually we can combine all the imaging elements into a single thick lens with a focal length $f$ as shown in the figure 1. To analyse the image formation process we can use two different approaches: using the classical geometric optics or diffraction theory. The geometric optics uses ray-tracing process to explain the image formation and is identical to a first-order approximation model since the geometrical distortion of the image is neglected. The classical physical optics and the diffraction theory can be used to analyse the images. The cameras used do not have sufficient spatial resolution to make diffraction effects significantly and we considered only geometric effects to analyse the process of image formation.

![Diagram of image formation](image)

Figure 1. Formation of focused and defocused images using thick lens. The defocused image is obtained by changing the place of the image sensor.

If the lenses are thick and, the front and back focal planes of them are planes normal to the optic axis situated at equal distances $f$, the following Gauss lens law holds:

$$\frac{1}{d_{in}} + \frac{1}{d_{out}} = \frac{1}{f}$$

The values $d_{in}$ and $d_{out}$ represent the distance to the image plane and the distance to the object imaged.

To explain the effects of defocusing an image we can trace the rays that light takes trough the optical system. Using this principle, for a point at infinity $d_{out}=\infty$ we obtain by equation (1) $d_{in}=f$. In this case the defocused image plane is placed at $f$ and the image of the three-dimensional point $P$ is also a point $Q$. If the image plane is displaced by an amount $\delta w_{out}$ then the image formed on it will be a circle with diameter $2r_c$ since if the aperture of the lens is also a circle. Using simple geometry and by similar triangles we find the following expression.

$$\frac{r_c}{d_{in}} = \frac{r_c}{\delta} = \frac{\delta r_c}{d_{in}}$$

The values $r_c$ and $r_c$ are the radius of the lens and the circle respectively. The physical optics explains the circle formation by the distribution of the energy received by the optical system.

Focus measure

To automatically measure the focus of a region in the image we need a criterion to measure the sharpness in that region. The notion of defocus has the inherent sensing of loss of information compared to focused images. This is equivalent to loss the quality of the image. Analysing this phenomenon using physical optics, the defocusing appears due to the loss of high-frequency components of light energy arriving to the optical system. If we ignore lens aberrations, a source of distortion of the image is due to diffraction caused by the wave nature of the light for coherent or for incoherent monochromatic illumination. Defocusing the optical system attenuates the high-frequency components arriving to the image plate. To measure the focus quality we can quantify the high-frequency content of the image. Several focus measure operators as base for different criterion of "sharpness" have been proposed by vision researchers. Krolik [13] in his thesis, compared some of them only using information in the image.

The Fourier transform is the first candidate for a criterion's function, but due to its high computational complexity and the use of special-purpose hardware to improve the calculation speed, it is not normally considered. Since the defocusing affects the edge characteristics it is natural to use an edge detector for the criterion function to measure the quality of focus. Based in this methodology we can estimate the gradient $V(l(x, y))$ at each image point and sum all the absolute values. The final value is a measure of the quality of focus. Acquiring a set of images at different focus positions and measuring the quality of focus for each one, we obtain a curve of values. The following step is to position the global maximum of this curve. That position is the index to an image well focused (by this criterion).

Other focus criteria can be used. For example we can measure the focus quality by filtering the images with an high-pass filter like the Laplacian

$$\nabla^2 I = \frac{d^2 I}{dx^2} + \frac{d^2 I}{dy^2}$$

The factor $I(x, y)$ represents the intensity of the point $(x, y)$. The focus quality can be computed by summing the absolute value of the Laplacian at each image point with magnitudes greater than a threshold value. The values obtained establish a curve for the different positions of focus. The criterion function, as above, searches by a global maximum in the curve.

Comparing the grey level histograms of two images, one well focused and another defocused, we observe that the variances of the two histograms are different. The histogram of the focused image presents a high variance and for the defocused image the histogram presents a low variance. The high variance of the image histogram is associated with sharp quality of the images while low variances are associated with blurring. The blurring reduces the amount of grey-level fluctuation and the variance can be a criterion to measure the defocus. Using this fact we can adopt the standard definition of variance as a focus measure

$$\sigma^2 = \frac{1}{N^2} \sum_{x=1}^{N} \sum_{y=1}^{N} (I(x, y) - \mu)^2$$

The parameter $\mu$ on the expression represents the mean of the grey-level distribution. In this case the criterion to search the focused image is given by the maximisation of $\sigma^2$.

The figure 2 shows the results of the application of some of the operators explained above for an image's sequence.

Since in the normal scenes are impossible to have all object focused, it is necessary to select evaluation windows containing features or a region of interest to evaluate the focus. Conversely, if the window contains the projection of two or more object points lying at different distances, then the criterion function will have more than one peak as show on figure 3.

CV-5.7.2
Depth calculation

Using the equation (1) for a situation of an image perfectly focused we have

\[ d_{\text{out}} = \frac{f d_{\text{in}}}{d_{\text{in}} - f} \]  

(5).

Using a lens without zoom, the focal length \( f \) is constant and the depth \( d_{\text{out}} \) from the camera to the object can be calculated.

3. DEPTH FROM STEREO VISION

Introduction

Stereo vision is an important method to obtain depth information. Once the corresponding points or primitives are identified from a pair of stereo images, the depth of the points or other primitives can be calculated by triangulation.

In our approach we use straight lines as primitives. The algorithm proposed has two steps: first -- generation of correspondence hypothesis and second -- the hypothesis verification and validation. Both the steps used the preliminary depth information obtained by focus. That information minimizes the set of possible correspondent primitives and helps to avoid the effect of occlusion.

Primitives

The primitives used are straight lines obtained by least-squares estimation using the points on the contours. The contours are obtained from images filtered by a recursive edge detector proposed by Derich [6]. The filtering operation describes the gradient magnitude and direction for each pixel of the image. Edge points are detected by threshold and by using hysteresis for non maximal suppression. A pixel is marked as an edge if the gradient magnitude at the pixel is larger than the gradient of neighbours. Marked edge points are then linked into contours and traced until an abrupt change in the local edge orientation is encountered. The line segments corresponding to these contours are then calculated by using least-squares.

Binocular stereo constraints

The recovery of scene topography from a stereo pair of images has typically preceding by a calibration phase. In this phase, the relative orientation of the two images is determined. This can be determined by a calibration where we calculate the values of the transformation between the two images and here represented by the matrices \( \text{CALIB}_{\text{LEFT}} \) and \( \text{CALIB}_{\text{RIGHT}} \). The first and second step of the algorithm are to do the correspondence between primitives in the two images. For that we create a set of constraints to generate hypothesis of correspondence.

In our algorithm the correspondences uses the midpoint point of the lines in the left and right images. The first step of the algorithm implements epipolar, disparity interval, geometric similarity and ordering constraints to generate hypothesis of correspondence. On the second step of the algorithm multiple hypotheses are solved. The final phase is the calculation of depth using triangulation. Stereo vision does not give the depth for all the points on the scene and to generate the intermediate data we can use interpolation [4].

Epipolar constraint: Once we have the transformation between the two images we can establish constraints for the position of the corresponding image points. These geometrical constraints are lines along the images. The position of corresponding image points must to lie along these lines.
A stereo vision system can be geometrically represented by two reference centres on the centres C1 and C2 representing the centres of the lenses. The images are represented by two planes \( \Gamma \) and \( \Gamma' \). The points on the scene projects on the two images using the perspective model. Given a point \( p' \) on the image \( \Gamma' \) all possible physical points \( P \) that may have produced \( p' \) are on the line supported by \( C, p' \). As a direct consequence the possible correspondences \( P' \) of \( p \) in the image \( \Gamma \) are located along a line known as epipolar. This epipolar line starts at point \( E' \), called epipolar point, which is the intersection of the line going through \( C_1 \), \( C_2 \) and the image plane \( \Gamma' \). This constraint reduced the dimension of our search space from one dimension to two dimensions.

**Geometric similarity:** The potential matches between lines must have similar proprieties such as length, orientation, gradient magnitude and phase. The orientation is a very important constraint to reject false matches.

**Ordering constraint:** Stereo projection usually preserves the order of primitives extracted from the two images along matching epipolar lines. The reason for this is that it is geometrically impossible for points arising from the same opaque surface to be differently ordered in the two images. Different orderings are possible only for scenes composed of surfaces of small extent such as isolated blobs or thin wires. Order is used in our algorithm to constrain the selection of correct matches.

**Disparity:** To determine the matching point for \( p' \) we can compute the epipolar line on image \( \Gamma' \) containing the correspondent point \( p'' \). The position of \( p'' \) in the image \( \Gamma \) can be measured relative to position of \( p' \) using a single parameter, for example, the difference in abscissas of the two image points \( p' \) and \( p'' \). That difference \( \delta u' - u '' \) is normally called disparity. In practice disparity values are limited - because the points are situated in front of the image planes; - by the image dimensions; - by the dimensions of the scene when finite.

The allowable disparities can be established by the course depth obtained by focus. The depth field of the lens gives an uncertainty on the depth measure by focus. This uncertainty establishes the allowable disparity for the correspondence.

4. CO-OPERATION BETWEEN FOCUS AND STEREO

The problem of matching can be viewed as a complex optimisation in which two criteria must be satisfied simultaneously. First, the corresponding points should have similar local features (gradient, intensity, colour, ...). Secondly, the local distribution of depth estimates, or equivalently, the local distribution of disparities should be plausible with respect to the real depths of the scene in that point.

The maximum values for disparity are directly dependent of the distance between the object and the camera. If the lenses used have a focal length adapted to the interval of the distances between objects and the camera a depth map can be obtained by focus. Since we have a coarse estimation of the depth it is possible to establish an interval for the disparity. If the system uses zoom lenses we can adapt the focal length to several situations.

Another situation is the ambiguity of some correspondences. That is the case suggested in the figure below where the point \( b' \) is not viewed by the left camera. In this case a classical stereo vision algorithm generates two correspondences \((a',a'')\) and \((a,b')\). To know which one of the correspondences is correct the order constraint is not enough. By using the depth map, previously obtained by focus, we verify that the correct match is \((a',a'')\) and \(b''\) does not have correspondence.

Figure 6. Occlusion can be solved by focus

5. EXPERIMENTAL STUDIES

The experimental site uses one manipulator with six degrees of freedom and the vision system is placed on the last link of the manipulator. The reference associated to the last link of the manipulator is known as [TOOL]. Two referential called [CAM_LEFT] and [CAM_RIGHT] are placed on the left and right cameras respectively. Other two, [BASE] and [W], represent respectively the referential of the manipulator and the referential of the world. One object is normally referenced to the referential [W] and we can know the position of the referential [CAM] communicating with the manipulator controller [5].

To obtain a previous depth map we need to use focus but we can not focus all scene simultaneously since the objects are not in the infinity. To focusing the scene we must to restrict our images to small windows. Automatic window selection based on this criterion requires some form of selectivity. It is also necessary to guarantee that the feature stays within the window. Since the objects that we are interested are not within the window we must to change the position of the vision system to put the image of object into the evaluation window. To do the movement we must to have the transformation corresponding to do this movement. That transformation can be obtained if we define a line \( N \) that passes by the point that we are trying to focus. Defining the vector \( n \) as the direction of this line, the plane XOY defined by the \( K \) must to coincide with the vector \( n \). This equivalent to rotate by an angle given by \( \arccos(\hat{k}n) \) with \( kn \geq 0 \) the axis of movement given by \( \hat{k}n \).

The rotation matrix resulting of this movement is given by:

\[
R = \begin{bmatrix}
1 - \frac{n_y^2}{1 + n_z} & -n_x n_y & n_x \\
-\frac{n_x n_y}{1 + n_z} & 1 - \frac{n_x^2}{1 + n_z} & -n_y \\
n_x n_z & n_y n_z & -n_x
\end{bmatrix}
\] (6)
The figure below shows the typical images of the line segments obtained after filtering the images with the recursive filter [6]. Lists of these segments, for the left and right image, are the inputs for the stereo vision algorithm. The second image shows an image of the segments.

Figure 6 - Movement of the referential (CAM, j) to place the optical axis coincident with the vector n.

Figure 7. Typical images of segment lines. These segments correspond to the edges of the objects on the scene. These edges are detected by an edge filter and after a contour is built using these points. The contour is the input for a process that detects lines on that contour.

6. CONCLUSIONS

In this study we addressed the problem of combining two important sources of three-dimensional information: stereo vision and focus. We described a method in which depth from focus is integrated in the matching process to speed up and to improve a stereo vision algorithm. Preliminary experiences show that better results can be obtained using cooperation with focus but more experimental work must be done to compare the performance of the method with the classical methods.

REFERENCES

CV-5.7.5