Dense Depth Maps using Stereo Vision Head

Luis Almeida ⋆ *and Jorge Dias*
e-mail:{laa, jorge}@isr.uc.pt

Instituto de Sistemas e Robótica
Departamento de Engenharia Electrotécnica - Universidade de Coimbra-Polo II
3030 COIMBRA, PORTUGAL

Abstract. In this paper, we examine and present the development of a depth map acquisition system with real-time characteristics based on a binocular active vision system. In order to obtain depth information we present a method that combines stereo matching with mechanical activity, reducing the time spent to perform the correspondence between the image points on the left and right images. Controlling the cameras' vergence or the baseline distance it is possible to change continuously the fixation point in the space and, at the same time, to select points with *correspondent* image projections. Computing the distance from those points to the vision system, is possible to obtain a dense relative map of the scene. The correspondence is established based on similarity measures between image regions. This measures are performed by operators with characteristics that makes this method suitable for parallel implementation. Since the depth information is relative, the calibration of the active vision system is minimal.

1 Introduction

One of the very important tasks in computer vision is to extract depth information of the world. Stereoscopy is a technique to extract depth information from two images of a scene taken from different view points. This information can be integrated on a single entity called dense depth map.

Dense depth maps are arrays with the distances from the object to the imaging system and the computation of precise depth information is, in generally, a time consuming task. In this paper we propose an algorithm to extract a *relative* dense depth map that has not requirements of a precise knowledge of the stereo system settings (calibration precision)[2]. The information of these dense relative maps can be integrated as a depth cue on higher level processes including object recognition, robot navigation or any other task that requires a three-dimensional representation of the physical environment.

In vision research several methods had been used to extract information about the structure of the world from video pictures obtained by active vision systems - some examples are [1], [6], [8], [9], [12].

A common way of approaching this is to use two stereo cameras in a similar fashion to the human visual system (see figure 1). A given point in the scene will, in general, project into two different planes and the information about its 3D position can be inferred from the vector between the two images. Depth recovering algorithms are well documented in the literature (references with pointers to other references [11],[10], [7], [9], [14],[15],[13]), nevertheless the algorithms included in our tests were selected for their potential robust real-time operation and their moderate hard and software implementation cost.

The main difficulty encountered in practice is the so-called, *correspondence problem*, namely identifying the same point in two images. Active vision simplify this problem by converging the cameras to a target of interest so that disparity is zero at the center of field of view and minimum at peripherical zone. The task of correspondence operators used to identify corresponding image point are less complex because they do not need to perform search for high magnitudes of disparities.

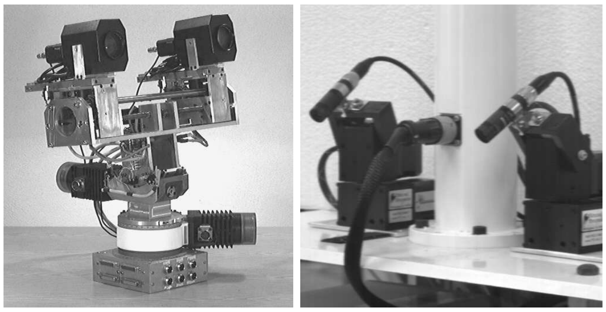

Fig. 1.: (a)-Active vision system (b)-Experimental cameras setup

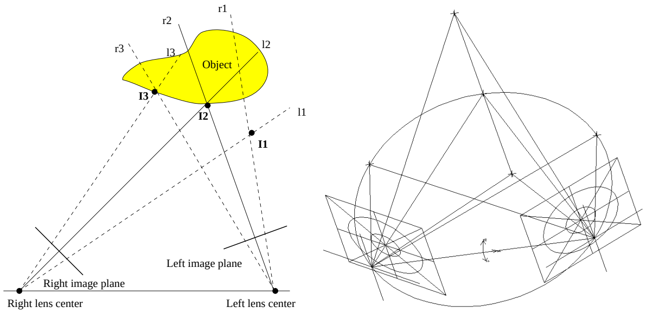

Fig. 2.: Convergent cameras schematic geometry (top view and 3D view)

1.1 Vergence geometry

Consider two cameras horizontal displaced, sharing a common tilt angle and assuming a convergent configuration (see figure 3). The vergence angle v subtended at the target P (fixation point) by the two camera optical center, E_l and E_r (nodal points), can be inscribed on a circle through these three points. From plane geometry the view rays to camera optical center from any point I (intersection point) on this circle subtend the same angle. Thus, since the optical camera rays yield zero disparity, so do the images of all points on this circle throughout the field of view. This special circle is known as the *geometric horopter* (Vieth-Muller circle of zero disparity). Object at this locus in the environment will be in correspondence (same pixel address) in two images. Two isodisparity circles with negative and positive disparity in term of angular difference (smaller / larger radius than circle E_l, E_r, P) are also marked.

Small disparities correspond to small deviation in depth from horopter. Such disparities can be measured by simple local neighbourhood operators, to build up a dense surface map of environment near horopter [8][16].

Coombs [4] characterises horopter as being the surface in three dimensional space defined by the points that stimulates exactly corresponding points (i.e., that as zero stereo disparity) in two cameras.

If $P = [X, Y, Z, 1]^T$ is an arbitrary point in 3D space defined in homogeneous co-ordinates, then the transformation of these world co-ordinates systems to a co-ordinate system aligned with camera E_i ($i\epsilon[l_{(left)}, r_{(right)}]$) will be $P_i^E = ^E T_W P_W$. $^E T_W$ is the transformation which takes world co-ordinates to camera co-ordinates E_i and $P_i^E = [x, y, z, 1]^T$ the co-ordinates on a camera referential. Finally a pinhole camera E_i maps P_i^E to the point $(u_i, v_i) = (f_i(x_i/z_i), f_i(y_i/z_i))$ (f_i is the focal length).

The region of space that has the same vertical and horizontal positions in left and right camera is then given by $u_l = u_r$ and $v_l = v_r$. Regardless of the complexity of the geometry relating the two camera, possible differences in focal length, and even possible misalignments, each of these contraints simplifies if we consider only vergence and a common tilt movements

The fixation point (X, Z) co-ordinates, for a $\mathbf{C}YCLOP$ $\mathbf{C}$ referential are expressed by equation (1):

$$\begin{cases} \tan\theta_l = \frac{X+b/2}{Z} \\ \tan\theta_r = \frac{X-b/2}{Z} \end{cases} = \begin{cases} \tan\theta_l = \frac{b/2+X}{Z} \\ -\tan\theta_r = \frac{b/2-X}{Z} \end{cases} = \begin{cases} X = \frac{b}{2}\frac{\tan\theta_l+\tan\theta_r}{\tan\theta_l-\tan\theta_r} \\ Z = \frac{b}{\tan\theta_l-\tan\theta_r} \end{cases} \quad (1)$$

* Departamento de Engenharia Electrotécnica - Instituto Politécnico de Tomar, 2300 TOMAR, PORTUGAL

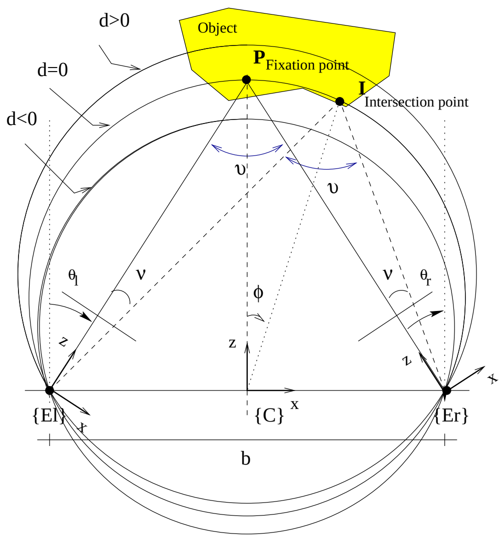

Fig. 3.: Geometric horopter

Thus $(X, Z) = (\frac{b}{2}\frac{\tan\theta_l + \tan\theta_r}{\tan\theta_l - \tan\theta_r}, \frac{b}{\tan\theta_l - \tan\theta_r})$ where Z is the perpendicular distance from the fixation point to the baseline.

2 Description of the algorithm

The active vision systems used in our experiences (figure 1) enables the control of the the cameras' vergence or the distance between the cameras. Such features are used during the algorithm execution.

The set of *view rays* with origin on left and right cameras intersect in space and if we select the intersections from *view rays* that have the same order in left and right cameras and verify the order constraint we will obtain *intersection points* (see figures 2 and 3), $I1 = l1 \cap r1, I2 = l2 \cap r2, I3 = l3 \cap r3, ...$). These points have zero disparity and we can filter those on the object's surface through similarity operators. The match probability between the correspondent image points is high. Controlling the cameras' vergence or the baseline distance we can sweep the space with this set of *view rays* and, at the same time, select the *intersection points* that have the maximum correlation. If we record the position of these points (that are related with the vergence angle or the baseline distance) we will obtain a dense depth map where the depth distances are recorded as positions where the *intersection point* intersect the object' surface. The method combines a software and a mechanical search to perform the correspondence between the image points on the left and right images-see figure 4. Verging the cameras from near parallel to several degrees of vergence, sweeps the visual rays and the intersection points through the scene.

The intersection points that are in the the objects' surface present high value in a similarity measurement. During the algorithm execution this output is continuously analyzed, for each corresponding image point pair. The peaks of this one-dimensional signal, correspond to likely object depths and when associated with the vergence angles, a three-dimensional map of depth is possible to obtain (see figure 4). Each similarity operator is implemented with operators that work independently and always on the same image point pair. This fact makes the method fast and ideal for hardware parallel implementation, generating simultaneously the similarity measurements for all image pairs.

The key problem on depth map building by stereo vision can be identified as finding the correct correspondence of image projections, i.e. homologous image points that represent a single point in the physical scene. There are not standard solution for the so-called *correspondence problem*, but the majority of approaches used can be roughly classified into two classes: correlation-based and feature-based methods.

The solution on this article belongs to the *correlation-based techniques*, which continuously comparing areas in the left and right images. The algorithm uses a simplification of the epipolar constraint because it assumes that for a small vergence movement, any 3D point will be projected always in the same image row. This assumption is also considered for small baseline movements. It also uses the continuity constraint and assumes that the intensities at two corresponding pixels are approximately the same.

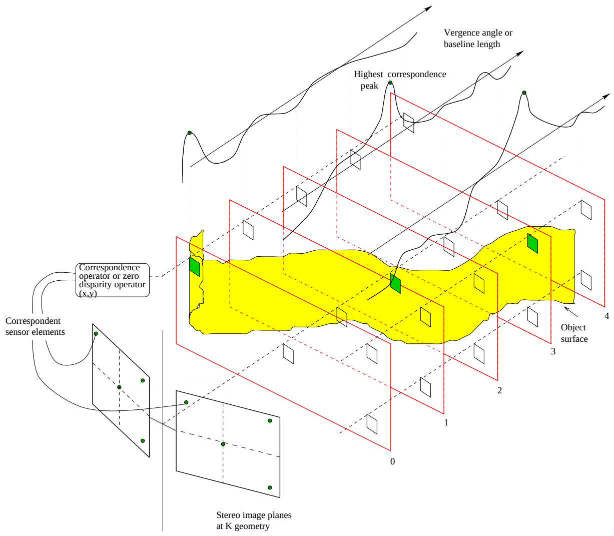

Fig. 4.: Space being sweep and the similarity outputs

The important constraint in our approach consists on the application of similarity operators just to image points that have same coordinates on both images. Using this constraint, that acts like a zero-disparity filter for convergent cameras, we ensure that the observed 3D point is on the horopter, i.e. it belongs to a curve in space that has zero disparity (or at least the disparity is minimum), which is desirable because points or features on such condition can be easily picked from the scene. This approach enables the separation of the object of interest from surroundings and simplify the calculus.

Since we know which points match, the measurement of the disparity is trivial and to know the real distance just requires the knowledge of the cameras' geometry.

During the execution, we perform the correlation over all image points and just store the best matches. The accuracy of these measures is very important because all the algorithm depends of the certainty of these matches. Similarity measures are computed by comparing a fixed window in the left image with a corresponding window in the right image while the vergence or baseline movements are performed. For each corresponding pair of pixel a curve of correlation scores is generated and the highest (or lowest depending of the similarity operator) give us the best matching point.

We tested some similarity operators such as the sum of absolute differences (SAD) (2), normalized cross-correlation operator (NCC) (3) and the zero mean normalized cross-correlation operator (ZNCC) (4)[3][5] . L and R stands for the left and right gray level images, with a wxw search window and $\overline{L}$ and $\overline{R}$ are the average gray level for the left and right image, respectively. Although the first operators have a low computational cost the results with real images are not good enough. The best operator seems to be the zero normalized cross-correlation operator but the computational cost is considerable. Experimentally this operator presents the best results because it is most invariant to affine transformations of the images which may result from slightly different cameras' settings. In our case the gray level distribution between images is slightly different and the equalization performed by this operator is essential.

$$SAD(x,y) = \sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} |L(x+i,y+j) - R(x+i,y+j)| \tag{2}$$

$$NCC(x,y) = \frac{\sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} L(x+i,y+j)R(x+i,y+j)}{\sqrt{\sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} L^2(x+i,y+j) \sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} R^2(x+i,y+j)}} \tag{3}$$

$$ZNCC(x,y) = \frac{\sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} (L(x+i,y+j) - \overline{L(x+i,y+j)})(R(x+i,y+j) - \overline{R(x+i,y+j)})}{\sqrt{\sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} (L(x+i,y+j) - \overline{L(x+i,y+j)})^2 \sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} (R(x+i,y+j) - \overline{R(x+i,y+j)})^2}} \tag{4}$$

2.1 Validating matches and the hierarchical algorithm

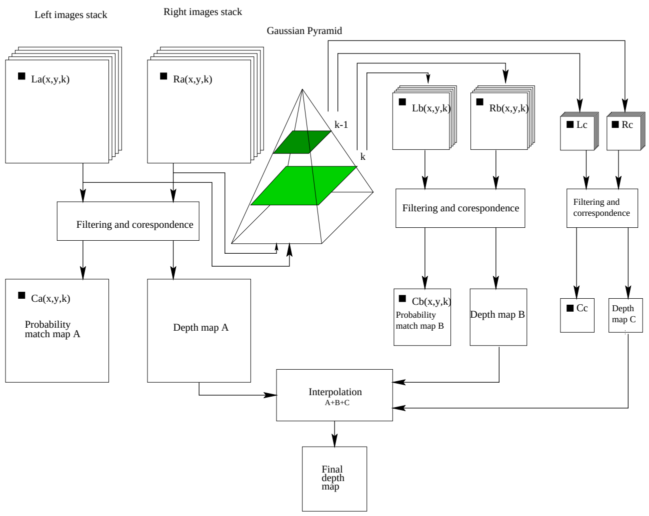

Fig. 5.: Algorithm overview scheme

Validating matches presents a serious problem and in order to certify the corresponding points we are using some specific constraints:

- The temporal continuity of the correlation scores curve, i.e. for each correlation curve associated to a corresponding pair of pixel's we analyze if the maximum(or minimum) values are consistent with the earlier and posterior values. Abrupt peaks can be noise.
- An uncertainty measure, based on the spatial matches distribution, i.e. we expect that the point with the highest match is surrounded by high match probability points.
- Multi-resolution coherence. Finally to increase the reliability we perform matching at several levels of resolution (computed by sub-sampling original images or gaussian smoothed images) with equal size windows (see figure 5).

The matching computation proceeds independently at all levels of resolution and in the end the results generated at low resolution are used to validate the results at higher levels. The method assumes that the results generated at low resolution are more reliable, if less precise, than those generated at high resolutions. For each image pair iteration, and at each level (a,b,c), we obtain a match probability map (i.e. a zero-disparity map) and a depth map. These depth maps are continuously updated during the vergence or baseline movements (each movement correspond to a few pixel shifts in the images rows). In the end, the low level resolution depth maps are expanded to the original size and together with the original one we compute a weighted average depth map (or simply reject those points where the peaks do not match). By the algorithm description it is possible to see that several processes can be executed in parallel and to speeding up the depth map acquisition process.

2.2 Experiments with Baseline Control

The software matching processes (correlations) are the time consuming tasks. The processing time is proportional to the size window search. However similarity operators SAD (equation 2) and NCC (equation 3) can be implemented to avoid redundant multiplication using recursions over the indices. With these operators it is possible to make the processing time independent of the window size but, as mentioned, the result's precision are not as good as the ZNCC operator. We are using search windows with 21x21 pixels. Smaller search windows can speedup the system, but the results are not so reliable (i.e. we miss the object pattern). Because the method is iterative the result precision is also dependent on the number of image pair acquisitions and on the range of each verging or baseline increments.

Figure 7.a shows a relative depth map of a big rotated box that, has, on top some other small boxes at different distance ranges. The big box is in front of the robot head around 215 cm from the baseline. The map is represented in shades of gray, dark meaning close, white meaning far. Figure 7.b represents the score curve correlation of a pixel with 100,128 coordinates. Figure 6 is one of the stereo image pair acquired during the baseline movements. The illustrated experience was achieved with a symmetric fixation geometry and changing the baseline distance with decrements of 1.5 mm (1.0823 cm in depth). We started with a initial baseline distance of 29.01cm.

Fig. 6.: Stereo image pair

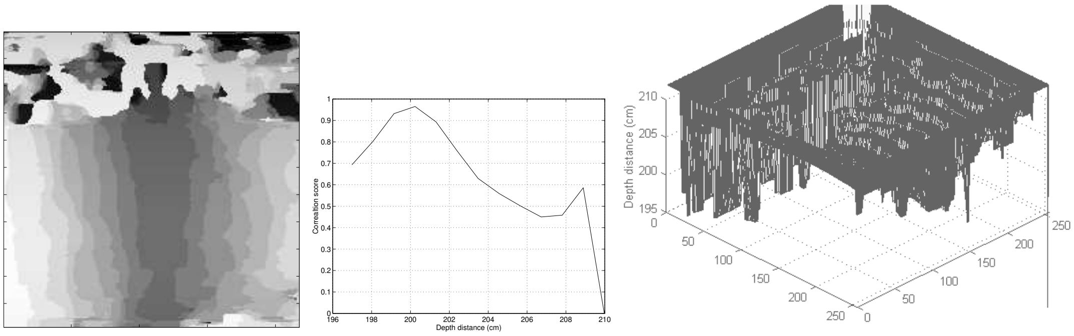

Fig. 7.: (a)-Depth map of a box. (b)-Score curve (c)-3D Depth map

2.3 DSP implementation

The active vision system consists of a 5 d.o.f. (see figure 1a) robotics platform with independent vergence, baseline, pan and tilt axes. This robot head is controlled by one host computer (PC 486/66 MHz) and a dual C40 Image Processing and frame-grabber PC board.

The image acquisition and processing is performed by a dual C40 image processing board from *Transtech Parallel Systems*. One of the two DSP C40 is associated with a frame-grabber (TDM435) and is designated as master while the other is called slave. The host computer communicates with the master C40 through a FIFO memory channel allowing transfer rates of 2MBytes/sec. The transference between the two C40's is performed through DMA controlled bi-directional comports allowing transfer rates of 40MBytes/sec even while they are processing (see figure 8). The frame-grabber can do acquisitions at a video rate frequency of 25Hz and it has multiple video entries, but just one at each time can be captured.

In order to implement the presented depth recovering algorithm in the *Transtech* image processing board we use the master C40 to perform the matching process between the left and right images at the original resolution (256x256) and their respectively filtering. At the same time, we transfer the original pair of images to the slave C40, that starts the sub-sampling and the matching process for low level resolutions, typically (128x128) and (64x64). Once the slave C40 have finished his task it sends the low resolution results to the master C40 which perform the validation process and computes of the final *relative depth map*. The master

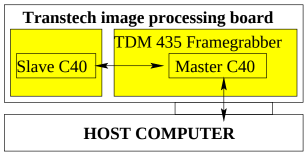

Fig. 8.: DSP hardware architecture

C40 is updated with the data geometry of the system and is responsible by the next baseline or vergence movement.

The parallel execution of the algorithm have been improved and, we still testing different CPU-C40 allocation solutions in order to improve the software performance.

2.4 Conclusions

In this article we have shown that a dense relative depth map acquisition system with real-time characteristics is possible using one active vision system. By using an active vision system we split the correspondence process between the left and right image into a software and a mechanical search. This search combination simplifies the matching phase and makes the software algorithm part suitable for parallel hardware implementation with all the advantages for a real-time system. It is robust, reliable and based on a low-cost active vision system. There is some simplifying assumptions about the environment (such as no abrupt depth changes), but they produce fairly dense results whose validity and accuracy can be quantitatively evaluated. The quality of these results is sufficient for depth cues in many 3D reconstruction applications. The independence of the algorithm from the calibration process is also an advantage because the results are relative distances between different points on the object, obtained without a precise camera calibration.

References

1. N.Ahuja and A. L. Abbott. Active stereo: integrating disparity, vergence, focus, aperture, and calibration for surface estimation, IEEE PAMI, 15(10):1007-1029, 1993.
2. J. Batista, Calibração de cameras de video, Provas de Aptidão de Capacidade Científica e Pedagógica, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, 1992.
3. P. Aschwanden, W. Guggenbulhl, Experiments Results from Comparative Study on Correlation-Type Registration Algorithms.
4. D.J.Coombs, Real-time Gaze Holding in Binocular Robot Vision,Ph. D. Dissertation, Dept. Of Computer Science, University of Rochester, June, 1992.
5. J. L. Crowley and J. Martin, Experimental Comparison of Correlation Techniques, IAS-4,International Conference on Intelligent Autonomous Systems, Karlsruhe, March 1995.
6. A. Francisco, Active structure acquisition by continuous fixation movements. Ph. D. dissertation, Dept. of Numerical Analysis and Computing Science, KTH (Royal Institute of Technology), June 1994. ISRN KTH/NA/P–94/17–SE.
7. Jorge Dias, 3D Reconstruction Using Dynamic Computer Vision,Ph. D. Dissertation, Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, November 1994.
8. M.Jenkin, J. Tsotos, and G. Dudek, The horopter and active cyclotorsion, Proc. of IEEE International Conference on Pattern Recognition, Oct. 1994, pp 707-710.
9. H. Sahabi and A. Basu, Analysis of Error in Depth Perception with Vergence and Spatially Varying Sensing, Journal of Computer vision and Image Understanding, May 1996, pp 447-461
10. D. Marr, Vision, W. Freeman, San Francisco, 1982.
11. W. Grimson, A computer implementation of a theory of human stereo vision, Philos. Trans. R. Soc. London B **292**, 1981, 217-253.
12. W. M.Theimer, H. A. Mallot, and S. Tolg, Phase Method for binocular vergence control and depth reconstruction, Proceeding of Spie Conference on intelligigent Robots and Computer Vision XI: Biological, Neural Net and 3-D Methods, Casaent, Vol 1826, pp 76-87, Boston, November, 1992.
13. E. Trucco, A. and Verri,Introductory Techniques for 3-D Computer Vision, Prentice Hall,1998.
14. D. J. Fleet, and A. D. Jepson, Stability on phase information, IEEE Trans PAMI, Vol 15, pp. 1253-1268, December, 1993.

15. C. Paredes,Fixação Visual: Uma abordagem computacional, Msc.,Faculdade de Ciência e Tecnologia da Universidade de Coimbra,Março,1998.
16. Carl F.M.Weiman,Log-polar vision system. Technical report, NASA,1994..